Linguistics and Language Models:
What Can They Learn from Each Other?
Leibniz Center for Informatics
Dagstuhl, Germany

Remarks on the Distributional Foundations of Language Models

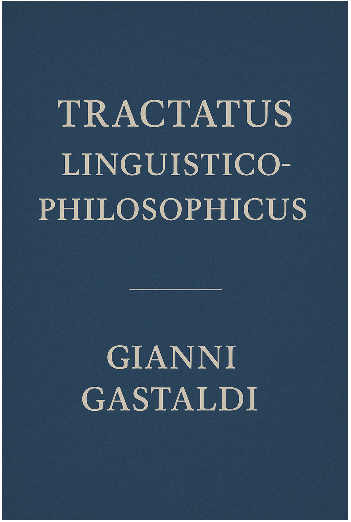
Juan Luis Gastaldi

`www.giannigastaldi.com`

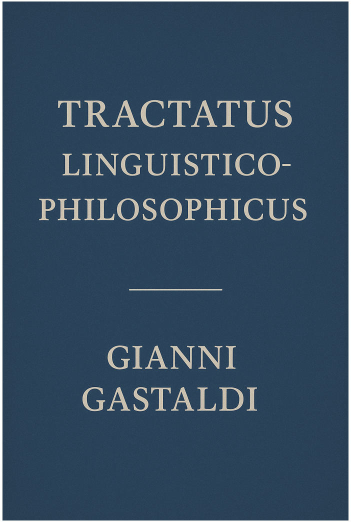
ETH zürich

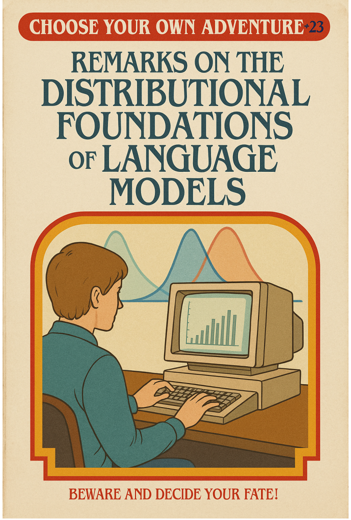
July 22, 2025

Format



Format

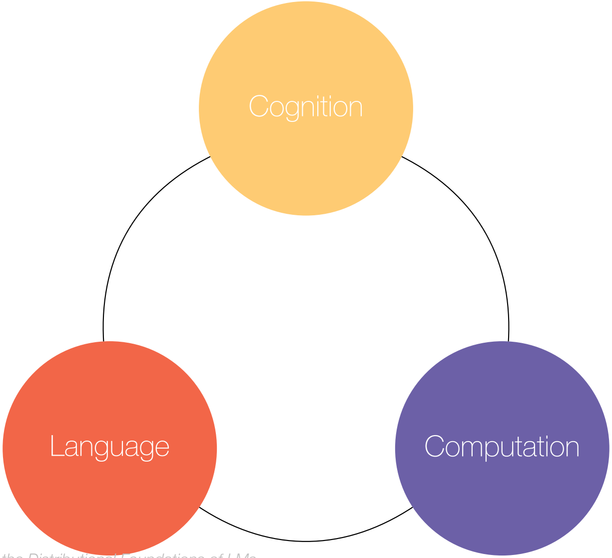

TRACTATUS
LINGUISTICO-
PHILOSOPHICUS
GIANNI
GASTALDI


CHOOSE YOUR OWN ADVENTURE
REMARKS ON THE
DISTRIBUTIONAL
FOUNDATIONS
OF LANGUAGE
MODELS
BEWARE AND DECIDE YOUR FATE!

It is sometimes said: animals do not talk because they lack the mental abilities. And this means: "They do not think, and that is why they do not talk." But — they simply do not talk.

Ludwig Wittgenstein, *Philosophical Investigations*, 1953, § 25

1.11 Chomsky's Generativist Program and the Cognitive Revolution



1.11 Chomsky's Generativist Program and the Cognitive Revolution

Cognition
vs.
Behavior
(Chomsky, 1959)

Cognition

Grammaticality
vs.
Inflection
(Chomsky, 1955)

Language

Computation

Computation
vs.
Logic
(Chomsky, 1955)

Chomsky Vs. LLMs

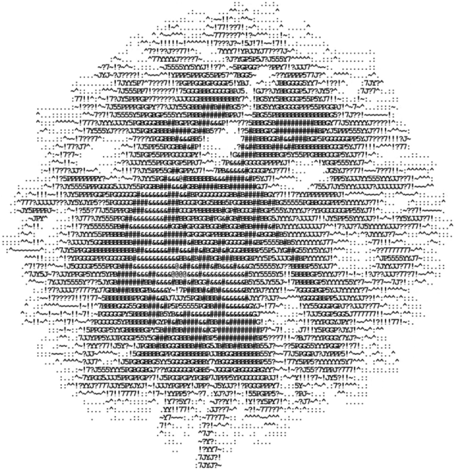
The New York Times

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)

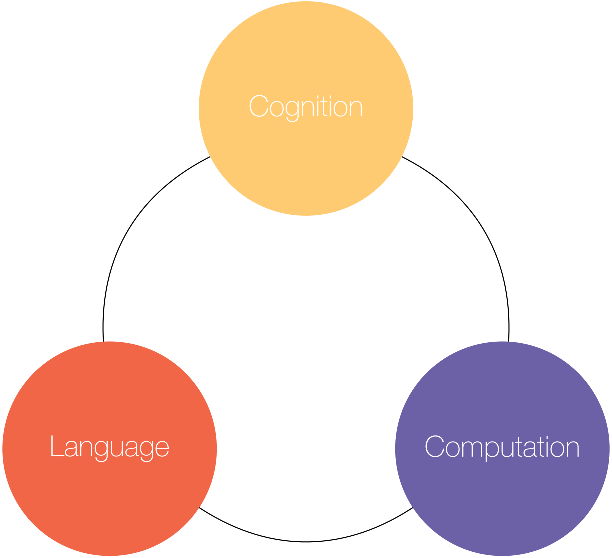
Chomsky Against Abstraction in Principle

"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."

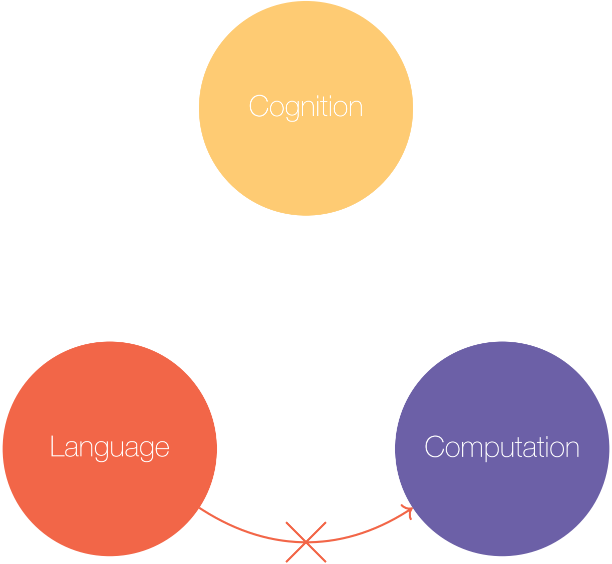
"I don't think we should cross that border, because there's no empirical claim. It is just a question of how to extend the metaphor."

(Chomsky, 1992)

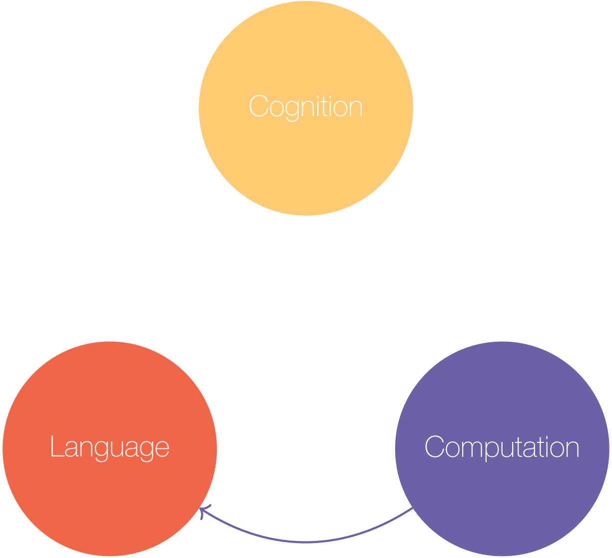
1.13 The Condition of Chomsky's Cognitive Foundations



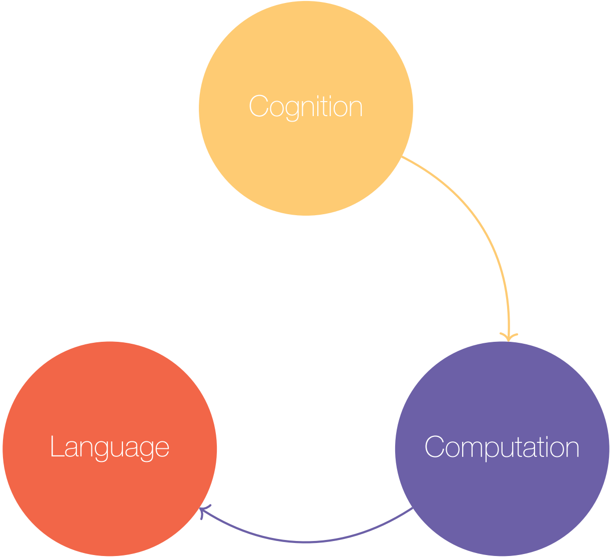
1.13 The Condition of Chomsky's Cognitive Foundations



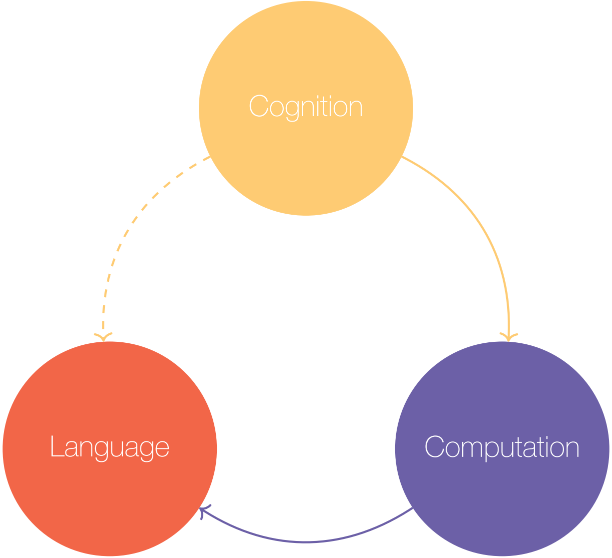
The Condition of Chomsky’s Cognitive Foundations

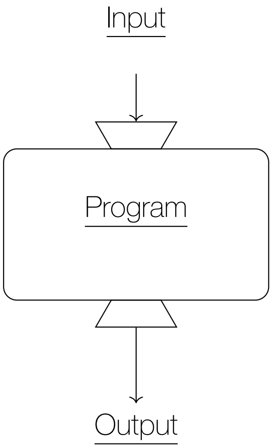


1.13 The Condition of Chomsky's Cognitive Foundations

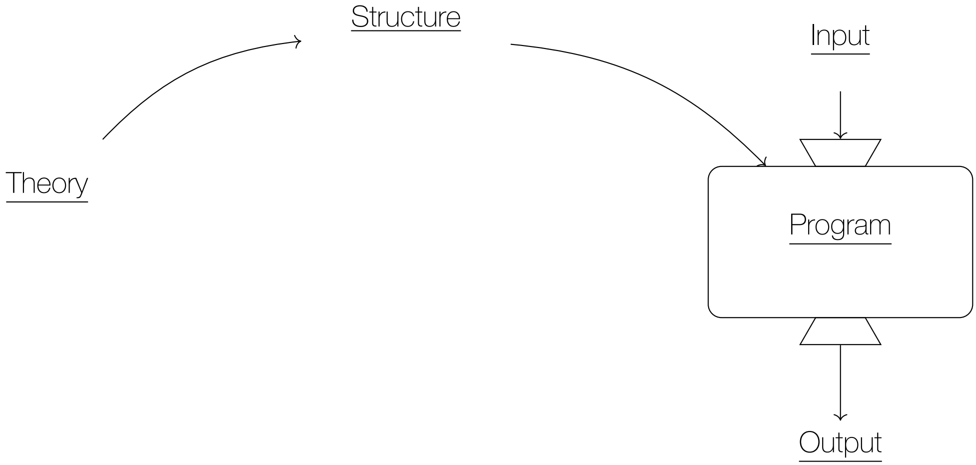


1.13 The Condition of Chomsky's Cognitive Foundations

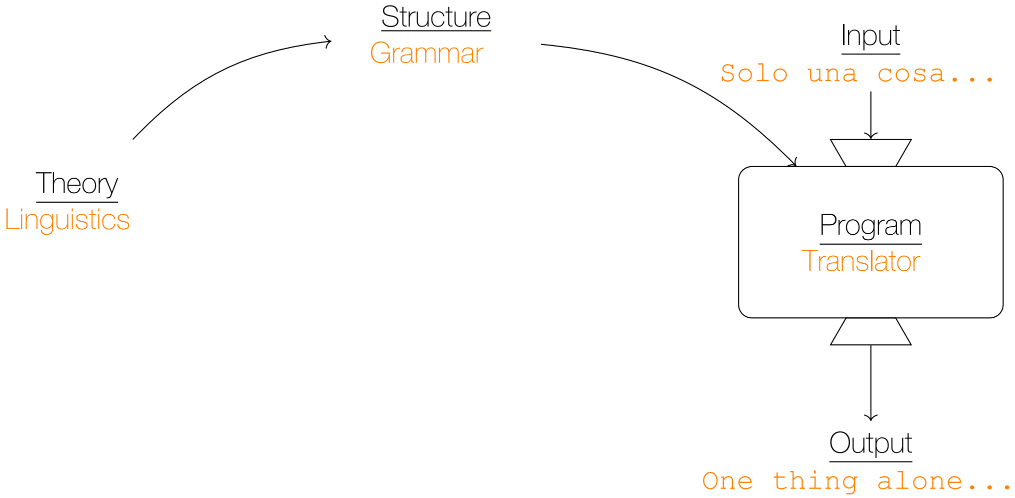



Input
Program
Output

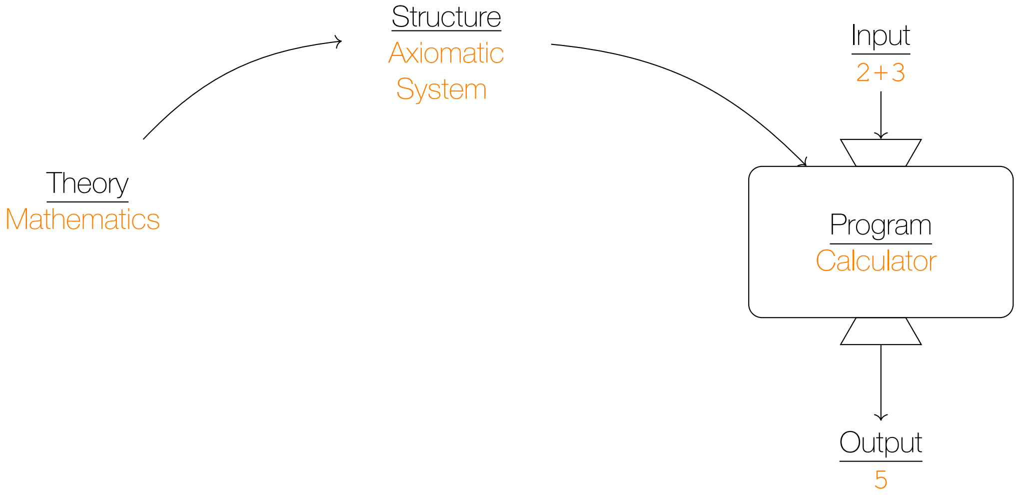
The Trap



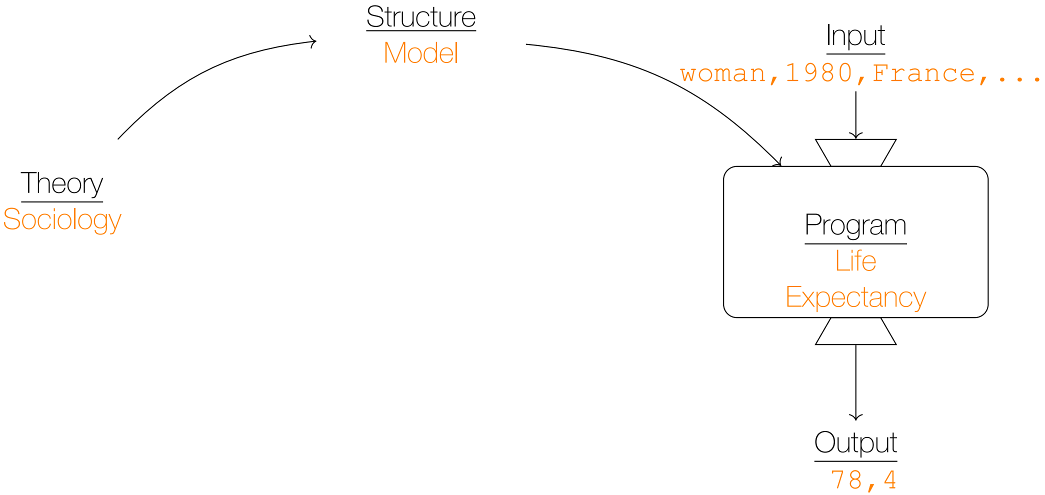
The Trap


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

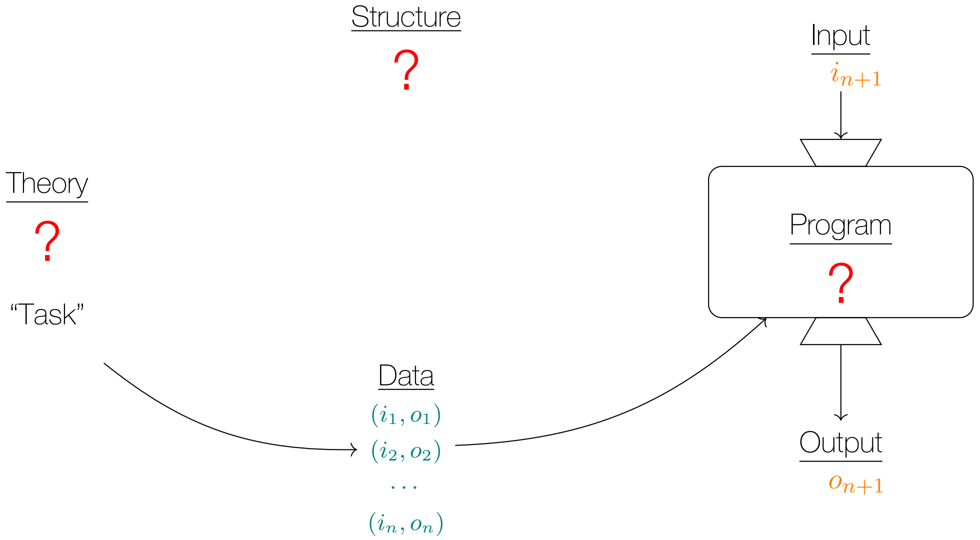

Structure
Model
Input
woman,1980,France,...
Theory
Sociology
Program
Life
Expectancy
Output
78,4

1.1 The Cognitive Import of Computational Language Models Is Not Unconditional

1.11 The contemporary connection between computational LMs and cognition was set up by Chomsky

1.12 Yet, he denies any theoretical legitimacy to LLMs

1.13 The connection set up by Chomsky has very precise epistemological conditions


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

1.2 Stochastic Parrots Vs. AGI

📉🦜

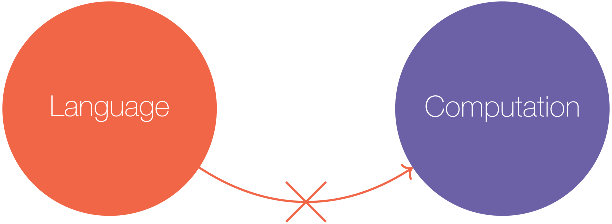
LLMs are not like us,
therefore they do not and can not have any relation to natural language.

🤖🧠

LLMs have a relation to natural language,
therefore they are like us.

The Chomskyan Condition Is Not Necessary

- 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)



The Chomskyan Condition Is Not Necessary

- Inadequacy of distributional models (Chomsky, 1953)

  Inconclusive

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

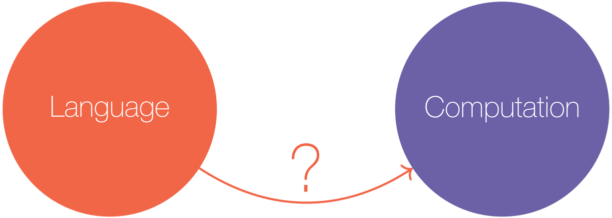
  Empirically challenged

- Limited expressive power of FSAs (Chomsky, 1956)

  The relevance is unclear

- Poverty of stimulus (Chomsky, 1959)

  Assumes what is to be proved



Formal Content

(Gastaldi and Pellissier, 2021)

Form ~~vs.~~ and ~~Meaning~~ Content

Kant, Hegel, Frege, Russian formalists, Saussure, Hjelmslev, etc.

Formal Content: The dimension of content which finds its source in the internal relations holding between the expressions of a language.

Formal Content

(Gastaldi and Pellissier, 2021)

Form ~~vs.~~ and ~~Meaning~~ Content

Kant, Hegel, Frege, Russian formalists, Saussure, Hjelmslev, etc.

Formal Content: The dimension of content which finds its source in the internal relations holding between the expressions of a language.

- *Characteristic Content*: The content resulting from the inclusion of a unit in a class of other units by which it accepts to be substituted in given contexts
- *Syntactic Content*: The content a unit receives as a result of the multiple dependencies it can maintain with respect to other units in its context
- *Informational Content*: The content related to the non-uniform distribution of units within those substitutability classes

Illustration of Formal Contents

Characteristic Content	Syntactic Content	Informational Content
{cat, dog, spider, gavagai}	"the gavagai is on the mat"	{cat:0.059%, dog:0.012%, spider:0.009% gavagai:0.000%}
Atomic Type	Profunctor Nucleus	Probability Distribution

(Gastaldi & Pellissier, 2021)

...I supposed that all the objects (presentations) that had ever entered into my mind when awake, had in them no more truth than the illusions of my dreams. But immediately upon this I observed that, whilst I thus wished to think that all was false, it was absolutely necessary that I, who thus thought, should be something; And as I observed that this truth, I think, therefore I am, was so certain and of such evidence that no ground of doubt, however extravagant, could be alleged by the Sceptics capable of shaking it, I concluded that I might, without scruple, accept it as the first principle of the philosophy of which I was in search.

Descartes, *Meditations on First Philosophy* (1641)

But I was persuaded that there was nothing in all the world, that there was no heaven, no earth, that there were no minds, nor any bodies: was I not then likewise persuaded that I did not exist? Not at all; of a surety I myself did exist since I persuaded myself of something [or merely because I thought of something]. But there is some deceiver or other, very powerful and very cunning, who ever employs his ingenuity in deceiving me. Then without doubt I exist also if he deceives me, and let him deceive me as much as he will, he can never cause me to be nothing so long as I think that I am something. So that after having reflected well and carefully examined all things, we must come to the definite conclusion that this proposition: I am, I exist, is necessarily true each time that I pronounce it, or that I mentally conceive it.

Descartes, *Meditations on First Philosophy* (1641)

…the philosopher has to say: "When I dissect the process expressed in the proposition 'I think,' I get a whole set of bold claims that are difficult, perhaps impossible, to establish, – for instance, that I am the one who is thinking, that there must be something that is thinking in the first place, that thinking is an activity and the effect of a being who is considered the cause, that there is an 'I,' and finally, that it has already been determined what is meant by thinking, – that I know what thinking is. [...]

Nietzsche, *Beyond Good and Evil*, §16 (1886)

Language Vs. Thought

…Because if I had not already made up my mind what thinking is, how could I tell whether what had just happened was not perhaps 'willing' or 'feeling'? Enough: this 'I think' presupposes that I compare my present state with other states that I have seen in myself, in order to determine what it is: and because of this retrospective comparison with other types of 'knowing,' this present state has absolutely no 'immediate certainty' for me." – In place of that "immediate certainty" which may, in this case, win the faith of the people, the philosopher gets handed a whole assortment of metaphysical questions, genuinely probing intellectual questions of conscience, such as: "Where do I get the concept of thinking from? Why do I believe in causes and effects? What gives me the right to speak about an I, and, for that matter, about an I as cause, and, finally, about an I as the cause of thoughts?" [...]

Nietzsche, *Beyond Good and Evil*, §16 (1886)

Now in order to cognize ourselves, there is required in addition to the act of thought, which brings the manifold of every possible intuition to the unity of apperception, a determinate mode of intuition, whereby this manifold is given; it therefore follows that although my existence is not indeed appearance (still less mere illusion), the determination of my existence can take place only in conformity with the form of inner sense, according to the special mode in which the manifold, which I combine, is given in inner intuition. Accordingly I have no cognition of myself as I am but merely as I appear to myself

Kant, *Critique of Pure Reason* (1781)

But, isn't thinking a kind of speaking? How is it possible for thinking to be engaged in a struggle with speaking? Wouldn't that be a struggle in which thinking was at war with itself? Doesn't this spell the end to the possibility of thinking?

Frege, *Sources of Knowledge of Math. and the math. natural Sc.* (1924-25)

Language Vs. Thought

It is sometimes said: animals do not talk because they lack the mental abilities. And this means: "They do not think, and that is why they do not talk." But — they simply do not talk.

Wittgenstein, *Philosophical Investigations*, 1953, § 25

Language Vs. Thought

The perennial man in the street believes that when he speaks he freely puts together whatever elements have the meanings he intends; but he does so only by choosing members of those classes that regularly occur together, and in the order in which these classes occur. [...] the restricted distribution of classes persists for all their occurrences; the restrictions are not disregarded arbitrarily, e.g. for semantic needs.

Harris, *Distributional Structure*, pp. 775-776, (1954).

1.3 The Lack of Cognitive Import Does Not Prevent LLMs to Be Models of Language

1.31 The Chomskyan condition does not hold of necessity

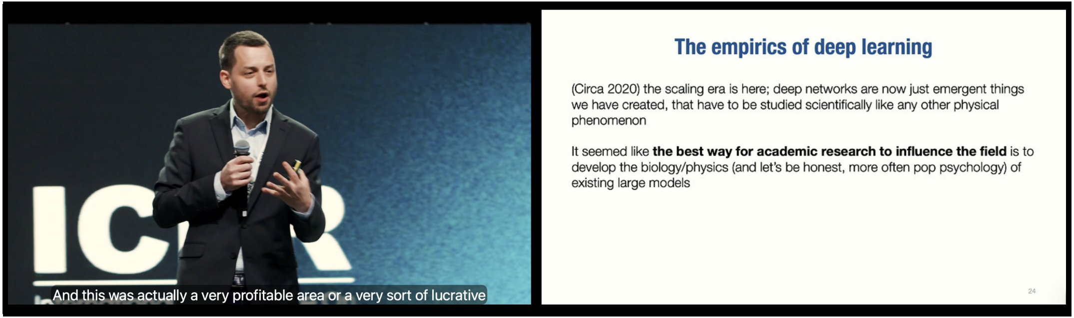
1.32 Content can be an effect of form

1.33 The divorce between language and thought is not recent

1 LLMs Have No a Priori Cognitive Import

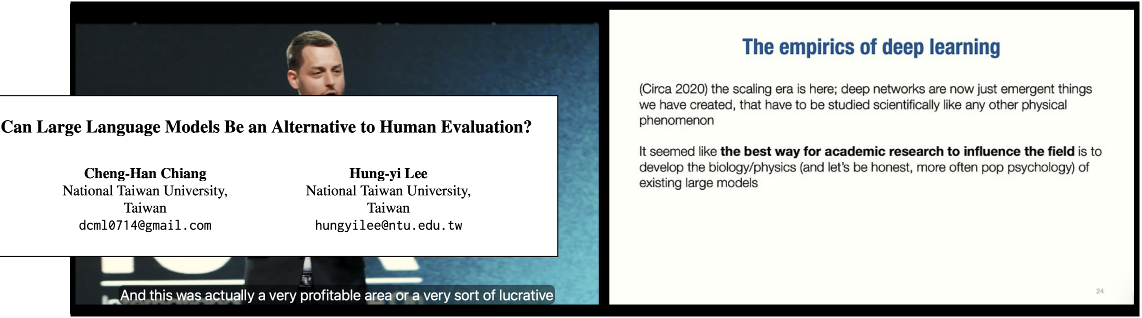
1.1 The cognitive import of computational language models is not unconditional
1.2 The epistemological condition ensuring such a connection does not hold for LLMs
1.3 The lack of cognitive import does not prevent LLMs to be models of language

Empirical Saturnalia



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

Empirical Saturnalia



Can Large Language Models Be an Alternative to Human Evaluation?

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

Do LLMs have Consistent Values?

Naama Rozen
Tel-Aviv University
naamarozen240@gmail.com

Liat Bezalel
Tel-Aviv University
liatbezalel@mail.tau.ac.il

Gal Elidan
Google Research
Hebrew University
elidan@google.com

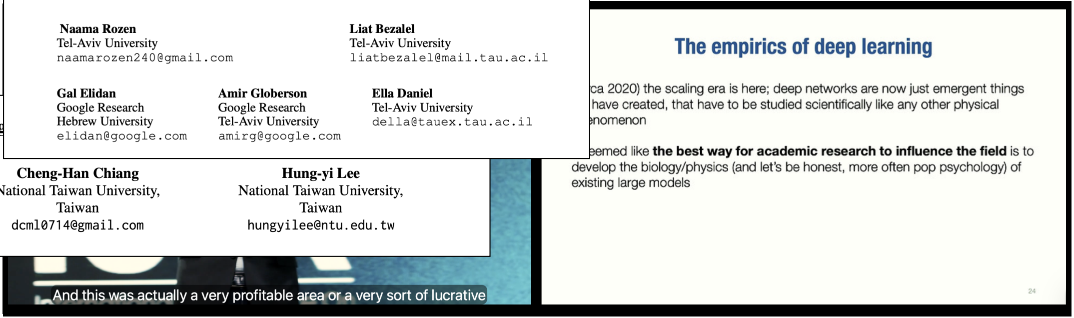
Amir Globerson
Google Research
Tel-Aviv University
amirg@google.com

Ella Daniel
Tel-Aviv University
della@tauex.tau.ac.il

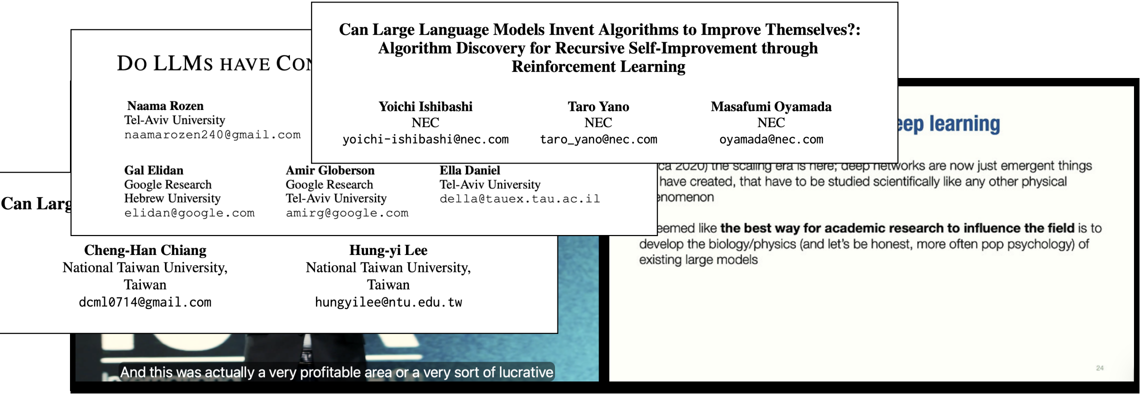
Can Larg

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

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

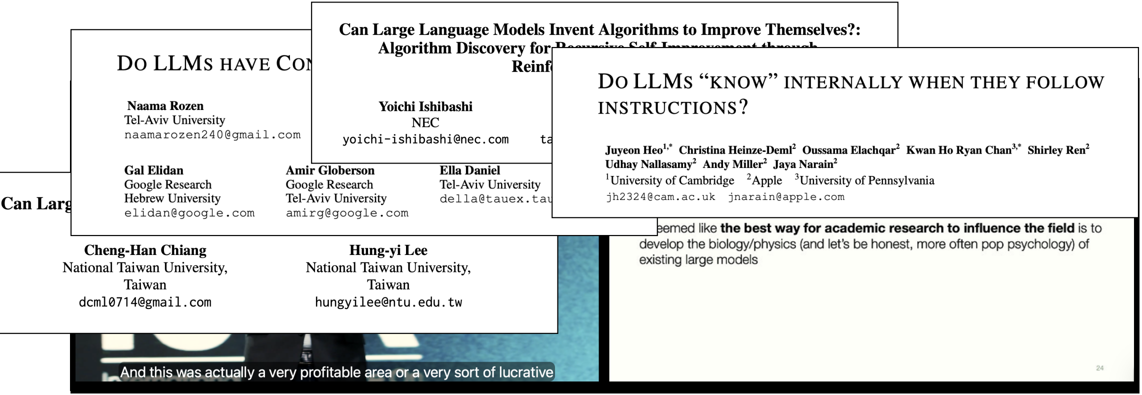


Zico Kolter, *Building Safe and Robust AI Systems*, Keynote at ICLR 2025.



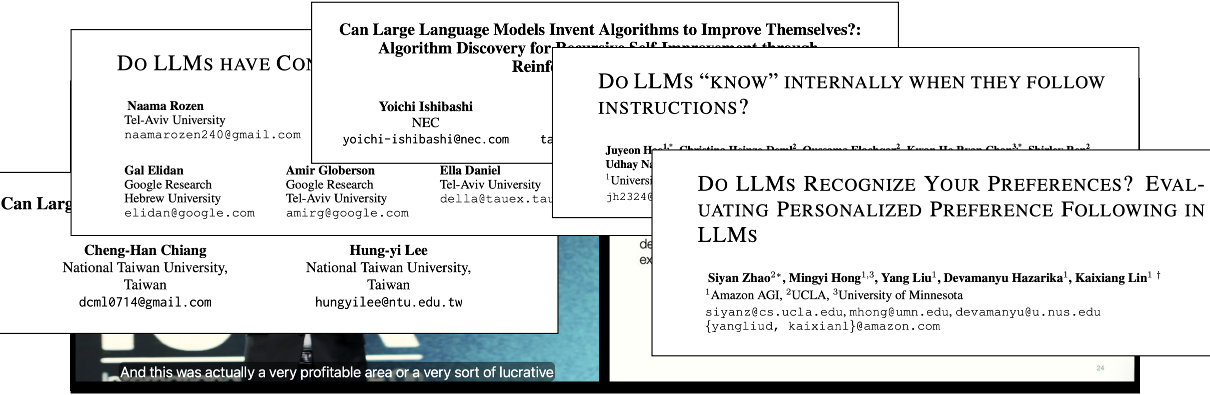
Zico Kolter, *Building Safe and Robust AI Systems*, Keynote at ICLR 2025.

2.1 Empirical Saturnalia

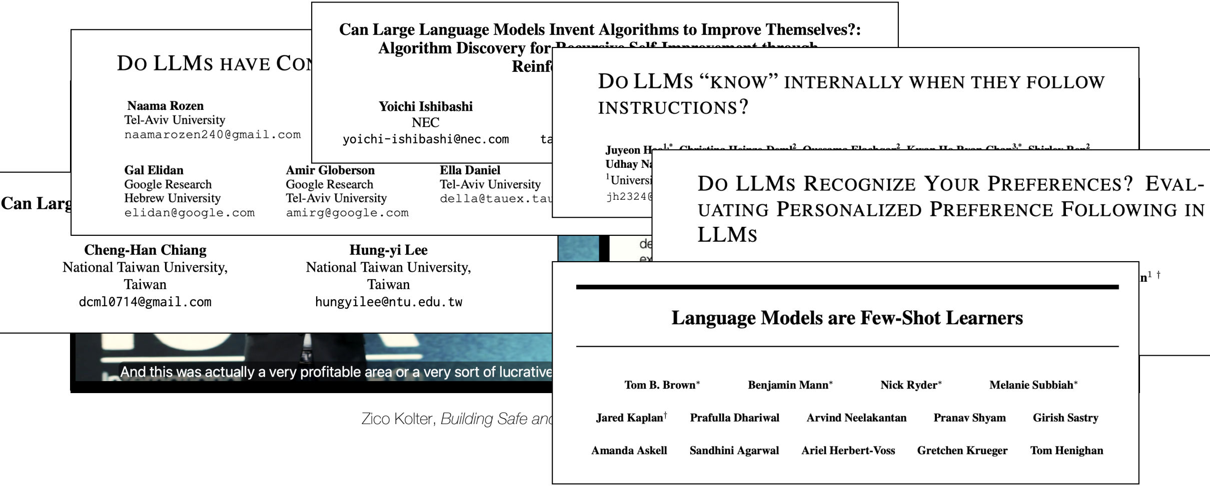


Zico Kolter, *Building Safe and Robust AI Systems*, Keynote at ICLR 2025.

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
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
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
And this was actually a very profitable area or a very sort of lucrative
Zico Kolter, Building Safe and
Do LLMs "know" internally when they follow instructions?
Udhay Na
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

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

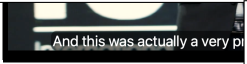
Can Larg

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

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

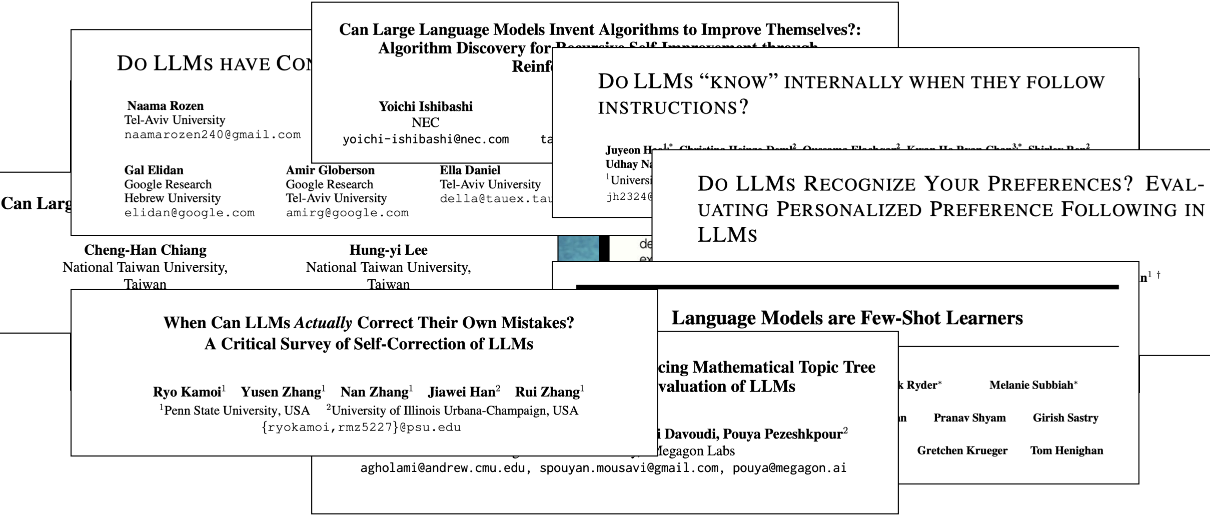
Language Models are Few-Shot Learners

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

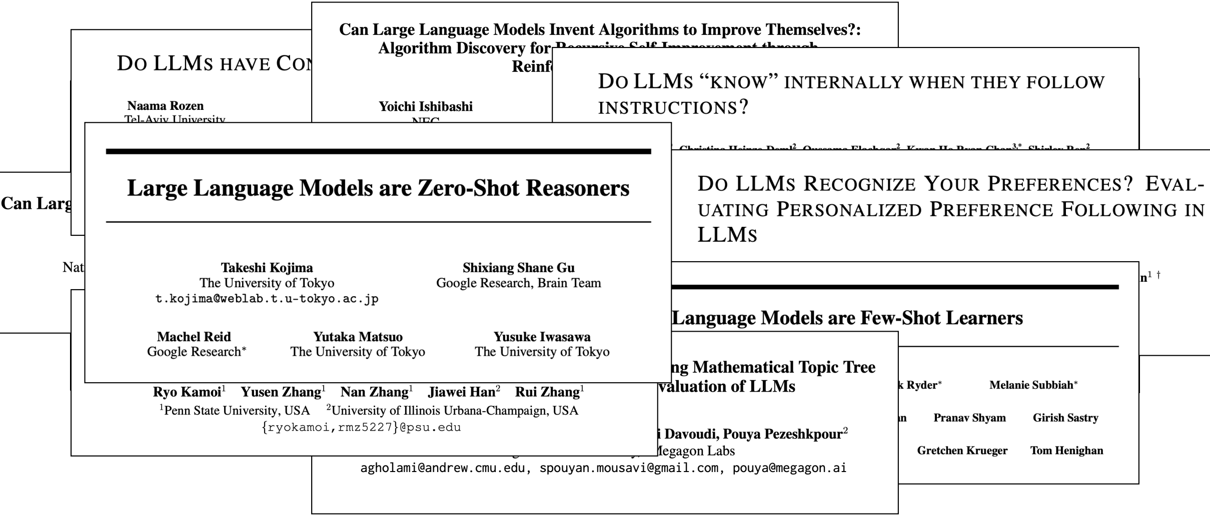


LLMs Are Not Intelligent Thinkers: Introducing Mathematical Topic Tree Benchmark for Comprehensive Evaluation of LLMs

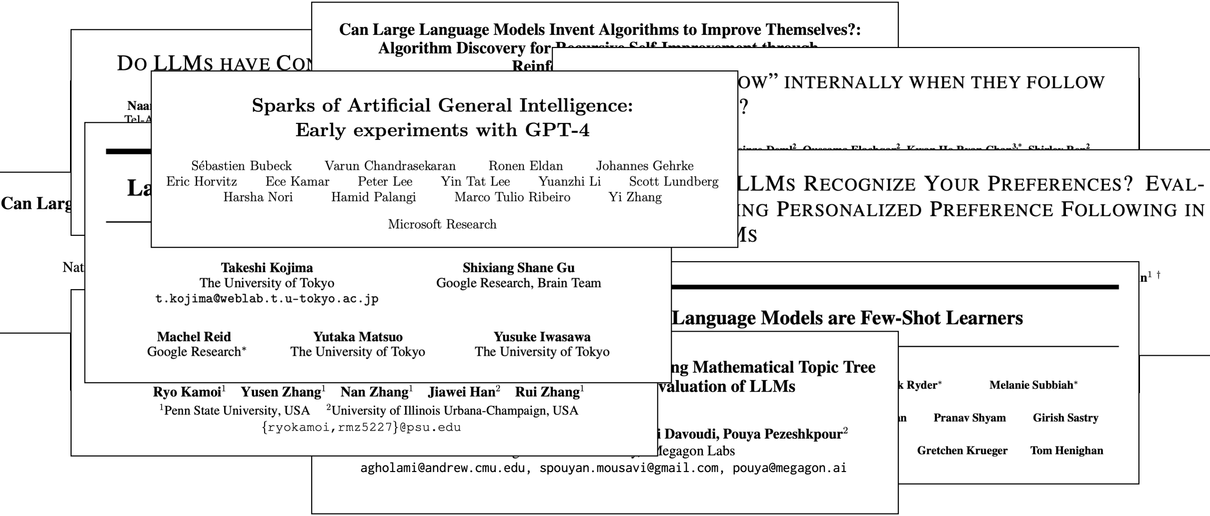
Arash Gholami Davoodi[1], Seyed Pouyan Mousavi Davoudi, Pouya Pezeshkpour[2]
[1]Carnegie Mellon University, [2]Megagon Labs
agholami@andrew.cmu.edu, spouyan.mousavi@gmail.com, pouya@megagon.ai

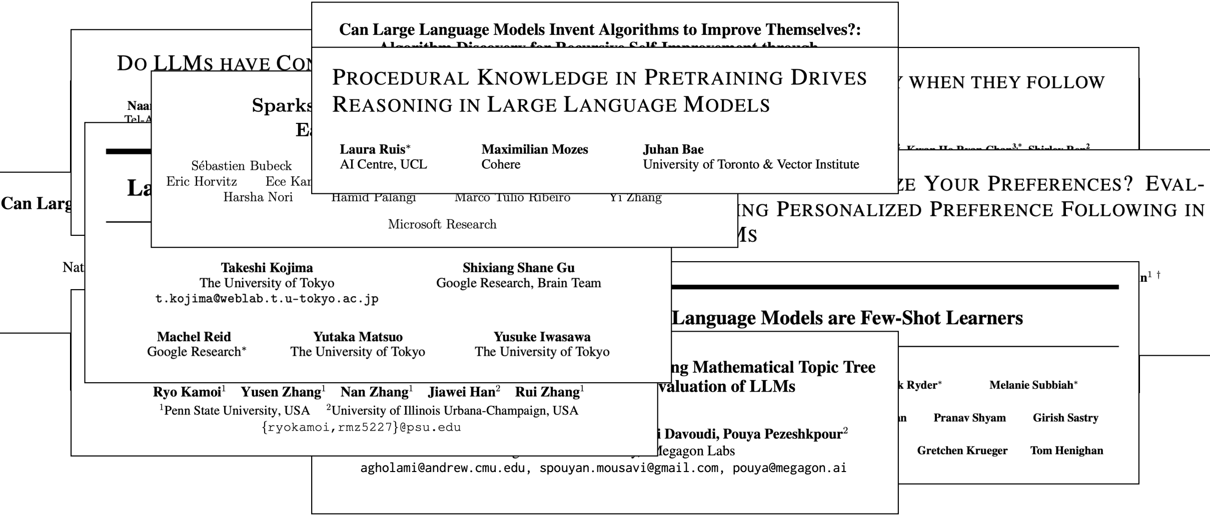

Can Larg
DO LLMS HAVE CON
Naama Rozen
Tel-Aviv University
naamarozen240@gmail.com
Gal Elidan
Google Research
Hebrew University
elidan@google.com
Amir Globerson
Google Research
Tel-Aviv University
amirg@google.com
Ella Daniel
Tel-Aviv University
della@tauex.tau
Can Large Language Models Invent Algorithms to Improve Themselves?:
Algorithm Discovery for
Reinf
Yoichi Ishibashi
NEC
yoichi-ishibashi@nec.com
DO LLMS "KNOW" INTERNALLY WHEN THEY FOLLOW INSTRUCTIONS?
Juyeon H
Udhay N
Universi
jh2324
DO LLMS RECOGNIZE YOUR PREFERENCES? EVALUATING PERSONALIZED PREFERENCE FOLLOWING IN LLMS
Cheng-Han Chiang
National Taiwan University, Taiwan
Hung-yi Lee
National Taiwan University, Taiwan
Language Models are Few-Shot Learners
Ryder*
Melanie Subbiah*
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]
Megagon Labs
agholami@andrew.cmu.edu, spouyan.mousavi@gmail.com, pouya@megagon.ai

Empirical Saturnalia

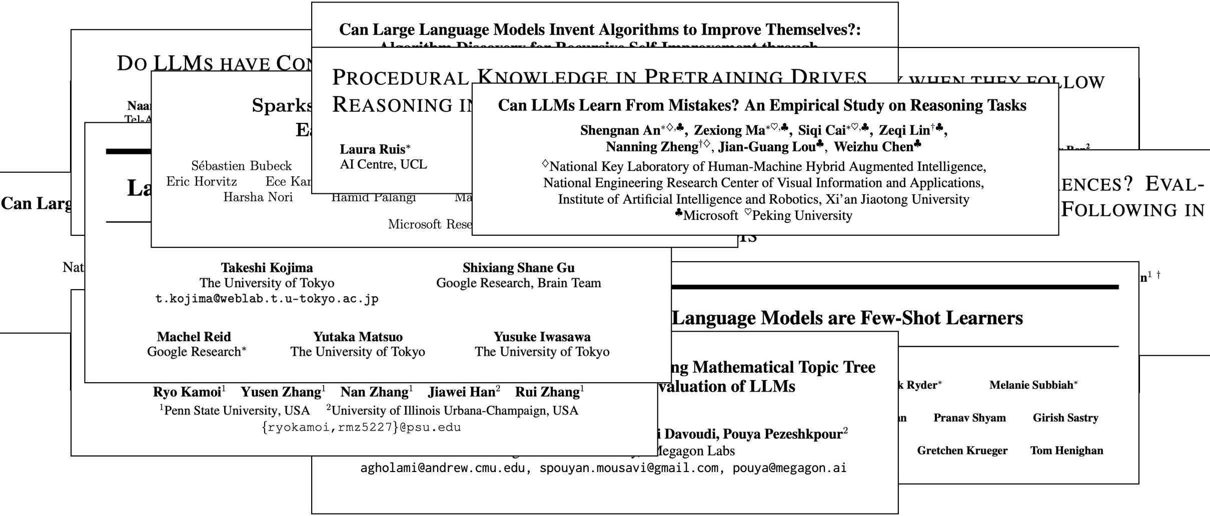


Empirical Saturnalia

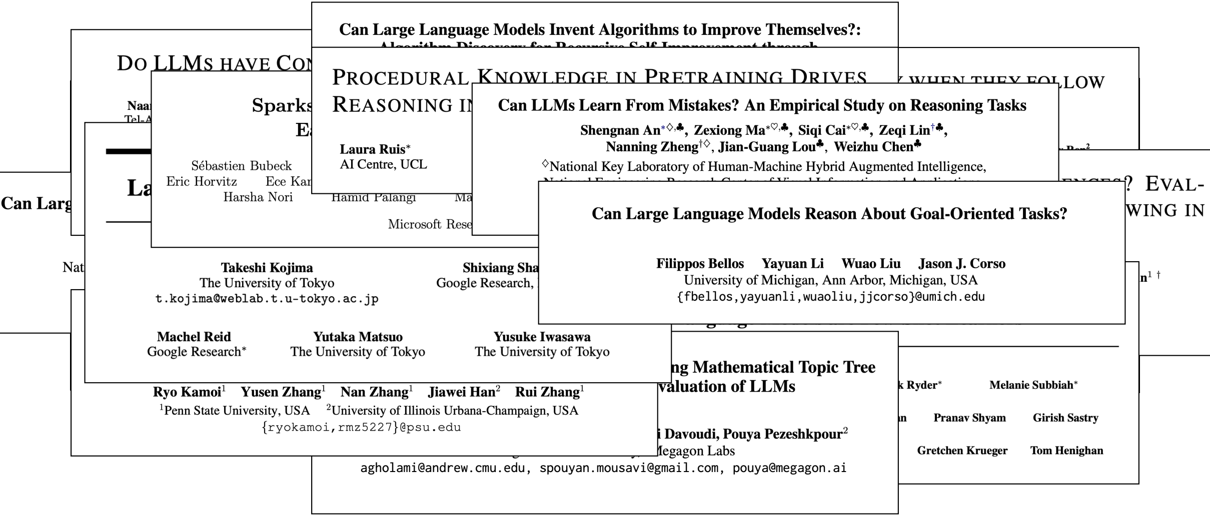


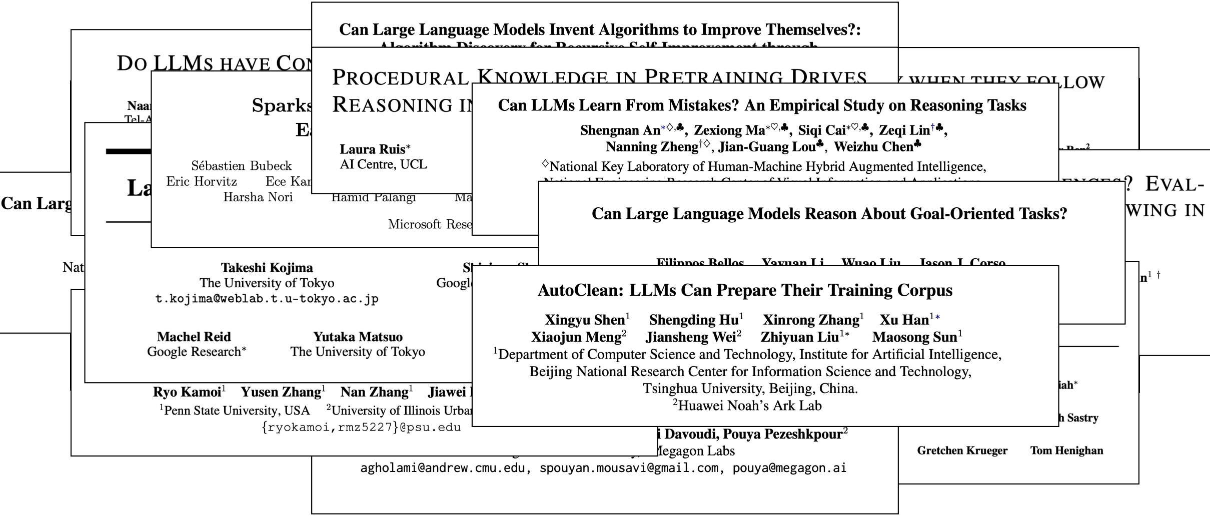

Can Large Language Models Invent Algorithms to Improve Themselves?:
PROCEDURAL KNOWLEDGE IN PRETRAINING DRIVES REASONING IN LARGE LANGUAGE MODELS
Laura Ruis* AI Centre, UCL
Maximilian Mozes Cohere
Juhan Bae University of Toronto & Vector Institute
DO LLMS HAVE CON
Sparks
Sébastien Bubeck
Eric Horvitz
Harsha Nori
Hamid Palangi
Marco Tulio Ribeiro
Yi Zhang
Microsoft Research
WHEN THEY FOLLOW
ZE YOUR PREFERENCES? EVAL-
ING PERSONALIZED PREFERENCE FOLLOWING IN
MS
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
Melanie Subbiah*
Pranav Shyam
Girish Sastry
Gretchen Krueger
Tom Henighan
ng Mathematical Topic Tree
valuation of LLMs
i Davoudi, Pouya Pezeshkpour
Megagon Labs
agholami@andrew.cmu.edu, spouyan.mousavi@gmail.com, pouya@megagon.ai
Ryo Kamoi Yusen Zhang Nan Zhang Jiawei Han Rui Zhang
Penn State University, USA
University of Illinois Urbana-Champaign, USA
{ryokamoi,rmz5227}@psu.edu
Can Larg

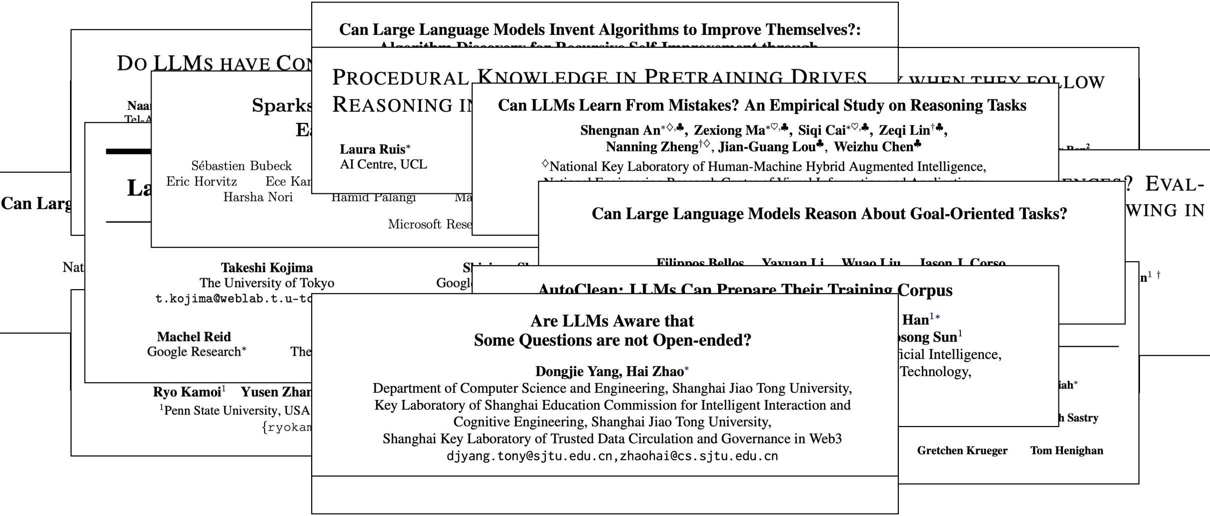
Empirical Saturnalia

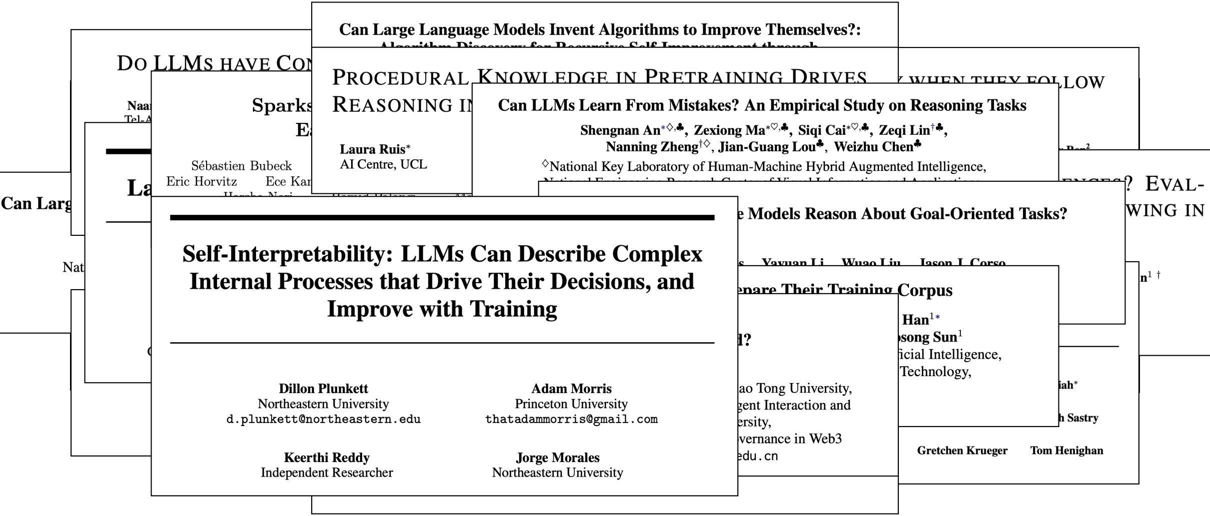


Empirical Saturnalia




Can Large Language Models Invent Algorithms to Improve Themselves?:
Procedural Knowledge in Pretraining Drives
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,
Sébastien Bubeck
Eric Horvitz
Harsha Nori
Microsoft Rese
Can Large Language Models Reason About Goal-Oriented Tasks?
Takeshi Kojima
The University of Tokyo
t.kojima@weblab.t.u-tokyo.ac.jp
Machel Reid
Google Research*
Yutaka Matsuo
The University of Tokyo
AutoClean: LLMs Can Prepare Their Training Corpus
Xingyu Shen, Shengding Hu, Xinrong Zhang, Xu Han
Xiaojun Meng, Jiansheng Wei, Zhiyuan Liu, Maosong Sun
Department of Computer Science and Technology, Institute for Artificial Intelligence,
Beijing National Research Center for Information Science and Technology,
Tsinghua University, Beijing, China.
Huawei Noah's Ark Lab
Ryo Kamoi, Yusen Zhang, Nan Zhang
Penn State University, USA
{ryokamoi,rmz5227}@psu.edu
Davoudi, Pouya Pezeshkpour
agholami@andrew.cmu.edu, spouyan.mousavi@gmail.com, pouya@megagon.ai
Gretchen Krueger
Tom Henighan


Can Large Language Models Invent Algorithms to Improve Themselves?:
Do LLMs have Con
Procedural Knowledge in Pretraining Drives
Reasoning in
Laura Ruis*
AI Centre, UCL
Sparks
Sébastien Bubeck
Eric Horvitz Ece Kar
Harsha Nori
Hamid Palangi
Microsoft Rese
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,
when they follow
Can Larg
Can Large Language Models Reason About Goal-Oriented Tasks?
Takeshi Kojima
The University of Tokyo
t.kojima@weblab.t.u-to
Machel Reid
Google Research*
Ryo Kamoi Yusen Zhan
Penn State University, USA
AutoClean: LLMs Can Prepare Their Training Corpus
Are LLMs Aware that
Some Questions are not Open-ended?
Dongjie Yang, Hai Zhao*
Department of Computer Science and Engineering, Shanghai Jiao Tong University,
Key Laboratory of Shanghai Education Commission for Intelligent Interaction and
Cognitive Engineering, Shanghai Jiao Tong University,
Shanghai Key Laboratory of Trusted Data Circulation and Governance in Web3
djyang.tony@sjtu.edu.cn,zhaohai@cs.sjtu.edu.cn
Gretchen Krueger Tom Henighan



Can Large Language Models Invent Algorithms to Improve Themselves?:
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,
Laura Ruis
AI Centre, UCL
Sébastien Bubeck
Eric Horvitz
Models Reason About Goal-Oriented Tasks?
Gretchen Krueger
Tom Henighan
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

Empirical Saturnalia

Can Large Language Models Invent Algorithms to Improve Themselves?:

Do LLMs have Co…

Procedural Knowledge in Pretraining Drives Reasoning in…

Laura Ruis*
AI Centre, UCL

Sparks…

Sébastien Bubeck
Eric Horvitz Ece Kar…

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[♣]

…when they follow

…ences? Eval-
…wing in

Can Larg…

…t Goal-Oriented Tasks?

Jason J. Corso

…Corpus

…Han[1*]
…song Sun[1]
…icial Intelligence,
…Technology,

…ao Tong University,
…gent Interaction and
…ersity,
…vernance in Web3
…edu.cn

Gretchen Krueger **Tom Henighan**

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

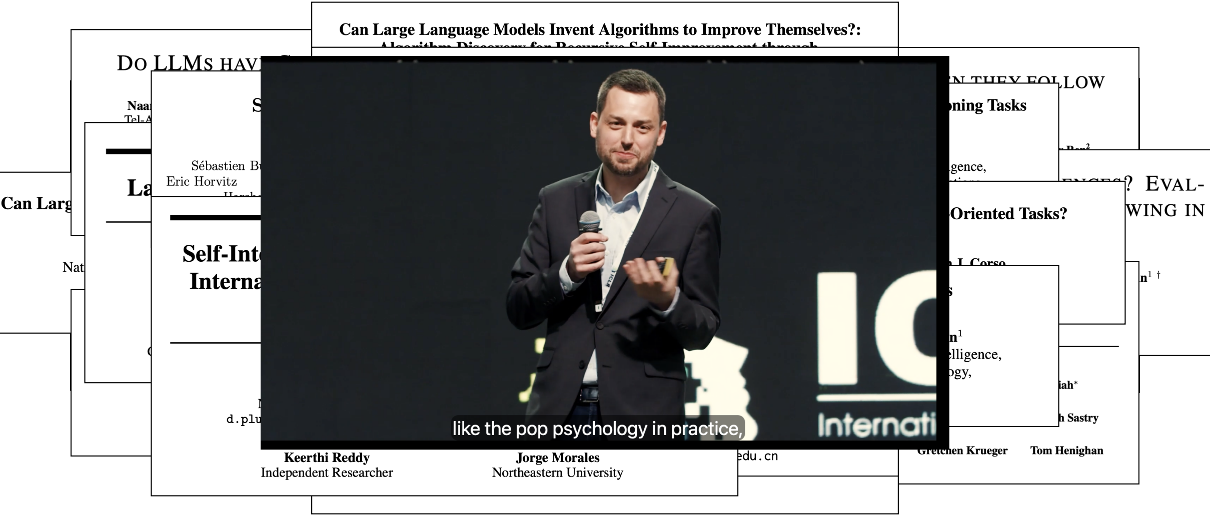
Self-Interpre…
Internal Pro…

Dillon Plunkett
Northeastern University
d.plunkett@northeastern.edu

Adam Morris
Princeton University
thatadammorris@gmail.com

Keerthi Reddy
Independent Researcher

Jorge Morales
Northeastern University


Can Large Language Models Invent Algorithms to Improve Themselves?:
Eric Horvitz
Keerthi Reddy
Independent Researcher
Jorge Morales
Northeastern University
Gretchen Krueger
Tom Henighan
like the pop psychology in practice,

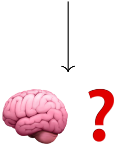
Neural LM

↓

?

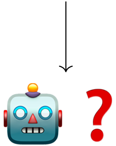
LLMs Are Computable Functions

Neural LM



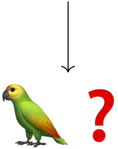
2.2 LLMs Are Computable Functions

Neural LM

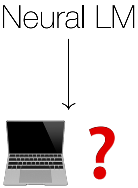


LLMs Are Computable Functions

Neural LM

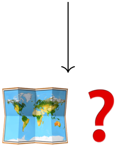


LLMs Are Computable Functions



LLMs Are Computable Functions

Neural LM



LLMs Are Computable Functions

Neural LM

↓

f !

LLMs Are Computable Functions

Neural LM

$$\downarrow$$

Function

$$f\colon \mathbb{R}^n \xrightarrow{f_1} \mathbb{R}^{n_1} \xrightarrow{f_2} \cdots \xrightarrow{f_K} \mathbb{R}^{n_K} \xrightarrow{g} \mathbb{R}^m$$

LLMs Are Computable Functions

Neural LM

↓

Computable

Function

$$f : \mathbb{R}^n \xrightarrow{f_1} \mathbb{R}^{n_1} \xrightarrow{f_2} \cdots \xrightarrow{f_K} \mathbb{R}^{n_K} \xrightarrow{g} \mathbb{R}^m$$

LLMs Are Computable Functions

Neural LM

↓

Computable

Function

$$f\colon \mathbb{R}^n \xrightarrow{f_1} \mathbb{R}^{n_1} \xrightarrow{f_2} \cdots \xrightarrow{f_K} \mathbb{R}^{n_K} \xrightarrow{g} \mathbb{R}^m$$

↓

Turing

Machine

(Turing, 1937)

LLMs Are Computable Functions

Neural LM

↓

Computable Function

$$f : \mathbb{R}^n \xrightarrow{f_1} \mathbb{R}^{n_1} \xrightarrow{f_2} \cdots \xrightarrow{f_K} \mathbb{R}^{n_K} \xrightarrow{g} \mathbb{R}^m$$

↙	↓	↘
Recursive Function	Turing Machine	Lambda Term
(Gödel, 1934)	(Turing, 1937)	(Church, 1936)

LLMs Are Computable Functions

Neural LM

↓

Computable

Function

$$f : \mathbb{R}^n \xrightarrow{f_1} \mathbb{R}^{n_1} \xrightarrow{f_2} \cdots \xrightarrow{f_K} \mathbb{R}^{n_K} \xrightarrow{g} \mathbb{R}^m$$

Recursive Function (Gödel, 1934) $\Longleftrightarrow$ Turing Machine (Turing, 1937) $\Longleftrightarrow$ Lambda Term (Church, 1936)

LLMs Are Computable Functions

Neural LM

↓

Computable
Function

$$f: \mathbb{R}^n \xrightarrow{f_1} \mathbb{R}^{n_1} \xrightarrow{f_2} \cdots \xrightarrow{f_K} \mathbb{R}^{n_K} \xrightarrow{g} \mathbb{R}^m$$

Recursive Function (Gödel, 1934) ⟺ Turing Machine (Turing, 1937) ⟺ Lambda Term (Church, 1936)

2.2 LLMs Are Computable Functions

Neural LM

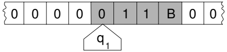
↓

Computable

Function

$$f : \mathbb{R}^n \xrightarrow{f_1} \mathbb{R}^{n_1} \xrightarrow{f_2} \cdots \xrightarrow{f_K} \mathbb{R}^{n_K} \xrightarrow{g} \mathbb{R}^m$$

Recursive Function (Gödel, 1934)	⟺	Turing Machine (Turing, 1937)	⟺	**Lambda Term** (Church, 1936)



credit: Nynexman4464

$$\lambda m.\lambda n.\lambda f.\lambda x.mf(nfx)$$

$$yxz$$

λ-Abstraction and β-Reduction in λ-Calculus

$$\lambda x.yxz$$

$$(\lambda x.yxz)t$$

λ-Abstraction and β-Reduction in λ-Calculus

$$(\lambda x.yxz)t$$

$$\rightsquigarrow$$

$$ytz$$

$$P := \lambda m.\lambda n.\lambda f.\lambda x.mf(nfx)$$

$$P := \lambda m.\lambda n.\lambda f.\lambda x.mf(nfx)$$

$$
\begin{aligned}
0&: \lambda f.\lambda x.x\\
1&: \lambda f.\lambda x.fx\\
2&: \lambda f.\lambda x.f(fx)\\
3&: \lambda f.\lambda x.f(f(fx))\\
4&: \lambda f.\lambda x.f(f(f(fx)))\\
5&: \lambda f.\lambda x.f(f(f(f(fx))))\\
&\quad \ldots\\
n&: \lambda f.\lambda x.\underbrace{f(\ldots(f}_{n \text{ times}}\,x)\ldots)
\end{aligned}
$$

$$P := \lambda m.\lambda n.\lambda f.\lambda x.mf(nfx)$$

0: $\lambda f.\lambda x.x$

1: $\lambda f.\lambda x.fx$

2: $\lambda f.\lambda x.f(fx)$

3: $\lambda f.\lambda x.f(f(fx))$

4: $\lambda f.\lambda x.f(f(f(fx)))$

5: $\lambda f.\lambda x.f(f(f(f(fx))))$

$\dots$

n: $\lambda f.\lambda x.\underbrace{f(\dots(f}_{n \text{ times}}\,x)\dots)$

$$\lambda m.\lambda n.\lambda f.\lambda x.mf(nfx)(\lambda f.\lambda x.f(fx))(\lambda f.\lambda x.f(f(fx)))$$

$$P := \lambda m.\lambda n.\lambda f.\lambda x.mf(nfx)$$

0: $\lambda f.\lambda x.x$

1: $\lambda f.\lambda x.fx$

2: $\lambda f.\lambda x.f(fx)$

3: $\lambda f.\lambda x.f(f(fx))$

4: $\lambda f.\lambda x.f(f(f(fx)))$

5: $\lambda f.\lambda x.f(f(f(f(fx))))$

$\ldots$

n: $\lambda f.\lambda x.\underbrace{f(\ldots(f}_{n \text{ times}}\,x)\ldots)$

$$\lambda m.\lambda n.\lambda f.\lambda x.mf(nfx)(\lambda f.\lambda x.f(fx))(\lambda f.\lambda x.f(f(fx)))$$

$$\rightsquigarrow$$
$$\rightsquigarrow$$
$$\rightsquigarrow$$
$$\rightsquigarrow$$
$$\rightsquigarrow$$
$$\rightsquigarrow$$
$$\rightsquigarrow$$

$$\lambda f.\lambda x.f(f(f(f(fx))))$$

$P := \lambda m.\lambda n.\lambda f.\lambda x.mf(nfx)$
$P' := \lambda r.\lambda s.\lambda f.\lambda x.f(f(f(f(fx))))$

0: $\lambda f.\lambda x.x$

1: $\lambda f.\lambda x.fx$

2: $\lambda f.\lambda x.f(fx)$

3: $\lambda f.\lambda x.f(f(fx))$

4: $\lambda f.\lambda x.f(f(f(fx)))$

5: $\lambda f.\lambda x.f(f(f(f(fx))))$

$\dots$

n: $\lambda f.\lambda x.\underbrace{f(\dots(f}_{n \text{ times}}\,x)\dots)$

$$\lambda r.\lambda s.\lambda f.\lambda x.f(f(f(f(fx))))(\lambda f.\lambda x.f(fx))(\lambda f.\lambda x.f(f(fx)))$$

$$\rightsquigarrow$$
$$\rightsquigarrow$$
$$\rightsquigarrow$$
$$\rightsquigarrow$$
$$\rightsquigarrow$$
$$\rightsquigarrow$$
$$\rightsquigarrow$$

$$\lambda f.\lambda x.f(f(f(f(fx))))$$

$$P := \lambda m.\lambda n.\lambda f.\lambda x.mf(nfx)$$

0: $\lambda f.\lambda x.x$

1: $\lambda f.\lambda x.fx$

2: $\lambda f.\lambda x.f(fx)$

3: $\lambda f.\lambda x.f(f(fx))$

4: $\lambda f.\lambda x.f(f(f(fx)))$

5: $\lambda f.\lambda x.f(f(f(f(fx))))$

$\dots$

n: $\lambda f.\lambda x.\underbrace{f(\dots(f}_{n \text{ times}}\,x)\dots)$

$$\lambda m.\lambda n.\lambda f.\lambda x.mf(nfx)(\lambda f.\lambda x.f(fx))(\lambda f.\lambda x.f(f(fx)))$$

$$\rightsquigarrow$$
$$\rightsquigarrow$$
$$\rightsquigarrow$$
$$\rightsquigarrow$$
$$\rightsquigarrow$$
$$\rightsquigarrow$$
$$\rightsquigarrow$$

$$\lambda f.\lambda x.f(f(f(f(fx))))$$

$$P := \lambda m.\lambda n.\lambda f.\lambda x.mf(nfx)$$

0: $\lambda f.\lambda x.x$

1: $\lambda f.\lambda x.fx$

2: $\lambda f.\lambda x.f(fx)$

3: $\lambda f.\lambda x.f(f(fx))$

4: $\lambda f.\lambda x.f(f(f(fx)))$

5: $\lambda f.\lambda x.f(f(f(f(fx))))$

$\ldots$

n: $\lambda f.\lambda x.\underbrace{f(\ldots(f}_{n \text{ times}}\, x)\ldots)$

$$\lambda m.\lambda n.\lambda f.\lambda x.mf(nfx)(\lambda f.\lambda x.f(fx))(\lambda f.\lambda x.f(f(fx)))$$

$$\lambda m.\lambda n.\lambda f.\lambda x.mf(nfx)(\lambda g.\lambda y.g(gy))(\lambda h.\lambda z.h(h(hz)))$$

$$\lambda n.\lambda f.\lambda x.(\lambda g.\lambda y.g(gy))f(nfx)(\lambda h.\lambda z.h(h(hz)))$$

$$\lambda n.\lambda f.\lambda x.(\lambda g.\lambda y.g(gy))f(nfx)(\lambda h.\lambda z.h(h(hz)))$$

$$\lambda f.\lambda x.(\lambda g.\lambda y.g(gy))f((\lambda h.\lambda z.h(h(hz)))fx)$$

$$\lambda f.\lambda x.(\lambda y.f(fy))((\lambda h.\lambda z.h(h(hz)))fx)$$

$$\lambda f.\lambda x.(\lambda y.f(fy))((\lambda z.f(f(fz)))x)$$

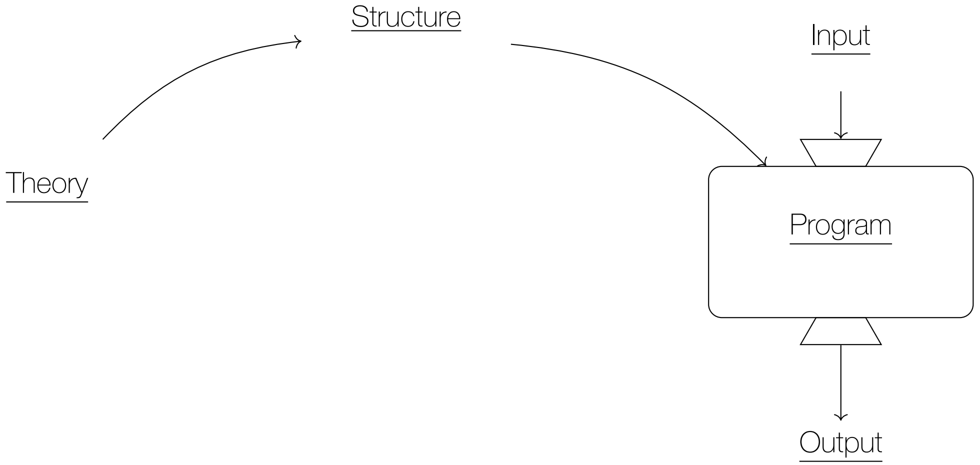
$$\lambda f.\lambda x.(\lambda y.f(fy))(f(f(fx)))$$

$$\lambda f.\lambda x.f(f(f(f(fx))))$$

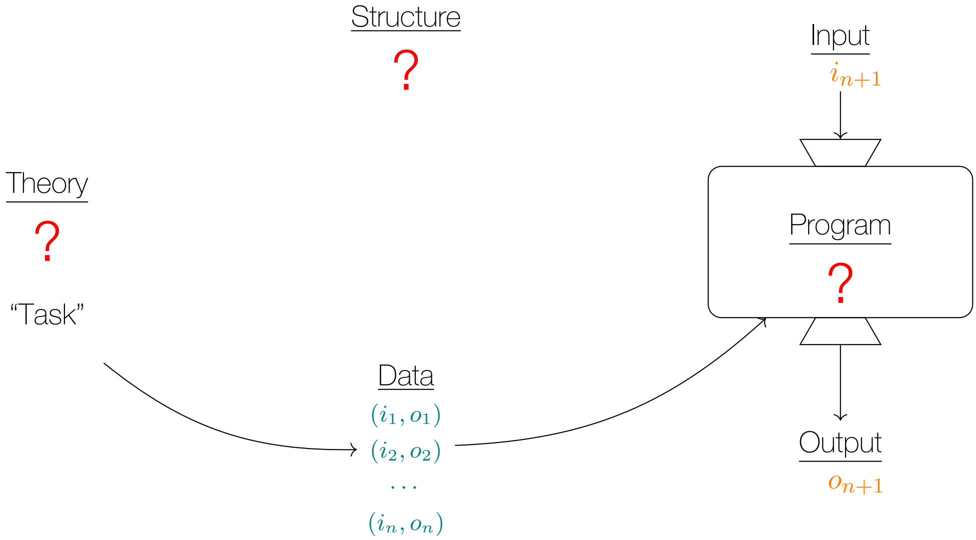
$P := \lambda m.\lambda n.\lambda f.\lambda x.mf(nfx)$

P'' :=λRófÄÒêÑ5È|Àxñ=∞ù ÿmWf286ëy'SÒú>v&ìÂ ¬2 óÉ7öç∞{ã>2ﬂB°µG#À9ÇU ∞btYBô¥ Ù ë‰3;5 å[l–èuô¨Ü ˚ 7–Ù.λ:ˆ4mÓØ¥´è—+ÌsÖ,$+gï„B™÷o–#ïŸê Ûv –gÓÿ/ëiijO‡Œﬁ • J1«€ø¸Ï hãêt‡æY$ˆ6 FïW»RÙKg¢¨.λ‡d¯… D2÷˛ò ˚ xêÈy. Ó”¢b Bé£NÈ1Ê‡/Û9Ñµ–/JYÇõË9ÿÀÈ.λÁÍ Àˆ öÇ,»fq∞±î˘B5Ì>O˜g™“6Ωe“æëC/ã… Ö · fÓ Å]ÑåyÊN°Ê ›¨.λÆà€fUòfEÙ˙Í m#„¸4\r√¬÷Îpò»y*vtÄJÃF1ûÁóz«ñM”DjŒ BËèÍT _ Êa‰ÁÇΩ @\Øˆ˜]Î›h‡.:ˆ4mÓØ¥´è—+ÌsÖ,$+gï„B™÷o–#ïŸê Ûv–gÓÿ /ëiijO‡Œﬁ • J1«€ø¸Ï hãêt‡æY$ˆ6 FïW»RÙKg¢¨ÁÍ Àˆ öÇ,»fq∞±î˘B5Ì>O˜g™“6 Ωe“æëC/ã… Ö · fÓ Å]ÑåyÊN°Ê ›¨(‡d¯… D2÷˛ò ˚ xêÈy. Ó”¢bBé£NÈ1Ê‡/Û9Ñµ–/ JYÇõË9ÿÀÈÁÍ Àˆ öÇ,»fq∞±î˘B5Ì>O˜g™“6Ωe“æëC/ã… Ö · fÓ Å]ÑåyÊN°Ê ›¨Æ à€fUòfEÙ˙Í m#„¸4\r√¬÷Îpò»y*vtÄJÃF1ûÁóz«ñM”DjŒBËèÍT _ Êa‰ÁÇΩ @\ Øˆ˜]Î›h‡)(ËÎÛeí4WµÍ }w„$Ω“K5èÃ¶%3[m '˘BÃﬁfÒ; óJçCËîõŸØcB,ñ$Ãå}ÓÅØ3¡ ˘?ö¯oŒ@ ﬂ8˘R CÆo˘* &<Ÿ ¬o12À‰ àÓÜ#ï ˝,ú¨ «ô¸„∞Îâä“øÃd|ˆÑ'Ê yØ;ˆW ›˝wò[}\»ÖÈúw'6<ù˘=ãÕ ¯Ì'Dz?2±|é' 3Â/rxµ∞µ$ÀëÂ*lf˜îû'+ÍVìyªGæßägô/,uÑ)

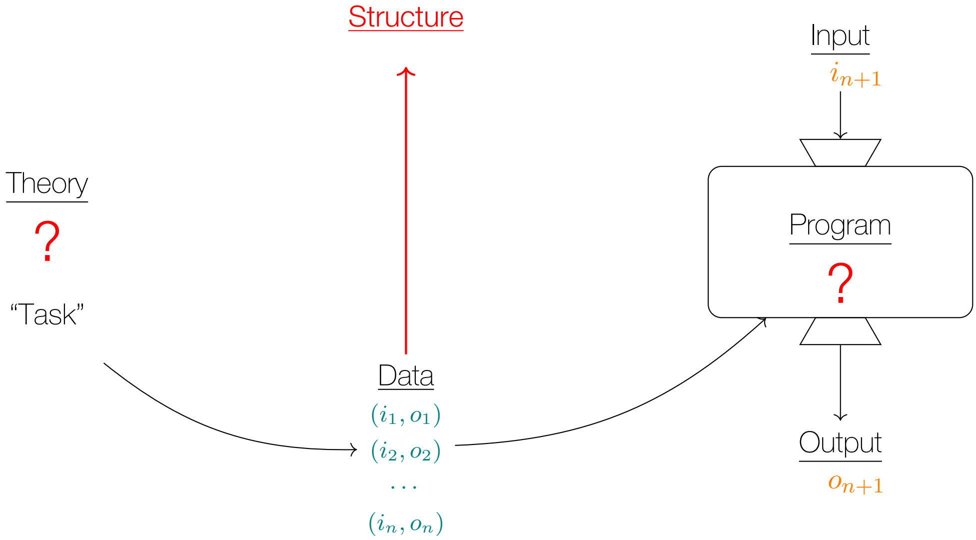
Implicit Structure



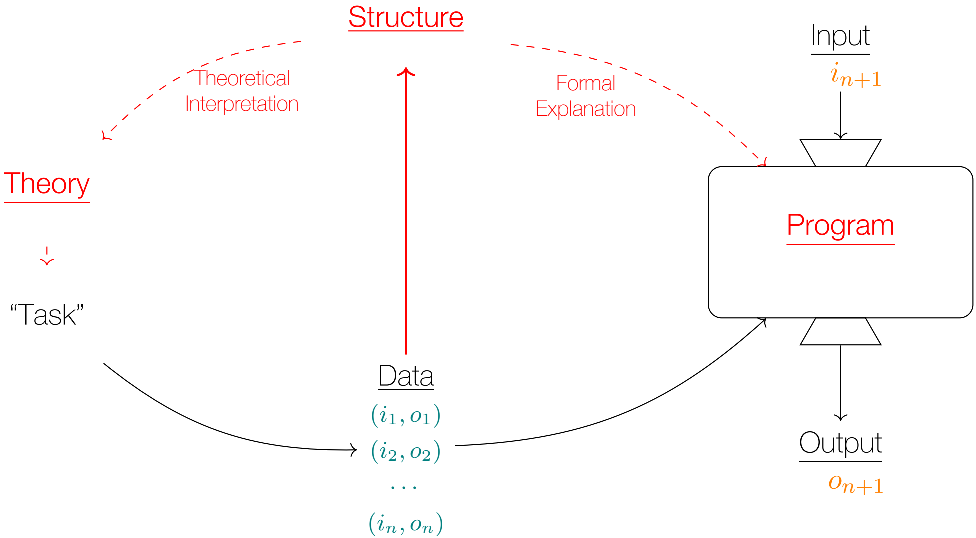
Implicit Structure



Implicit Structure



Implicit Structure



2 The Empirical Study of LLMs Has No Epistemological Grounds

2.1 The NLP field has embraced an empirical turn

2.2 But LLMs are just computable functions

2.3 There is no empirical way of knowing what a computable function does

2.4 The only valid epistemological question is: What is this function the implementation of?

Distributionalism and Contexts

- "You shall know a word by the company it keeps!" (Firth, 1957)
- "Words which are similar in meaning occur in similar contexts" (Rubenstein & Goodenough 1965)
- "Words with similar meanings will occur with similar neighbors if enough text material is available" (Schütze & Pedersen 1995)
- "A representation that captures much of how words are used in natural context will capture much of what we mean by meaning" (Landauer & Dumais 1997)
- "Words that occur in the same contexts tend to have similar meanings" (Pantel 2005)
- "The degree of semantic similarity between two linguistic expressions A and B is a function of the similarity of the linguistic contexts in which A and B can appear" (Lenci, 2008)

3.3 Cognitive and Pragmatic Interpretations of Distributionalism

- Two versions of the Distributional Hypothesis (Lenci, 2008):
  - Weak: Correlation between context and word meaning (Spence and Owens, 1990)
  - Strong: Causality attributed to contextual distributions (Miller and Charles, 1991)
- Theory of (linguistic) meaning as "usage" (Wittgenstein) "the meaning of a word is defined by the circumstances of its use" (Manning and Schütze, 1999)

3.3 Cognitive and Pragmatic Interpretations of Distributionalism

- Two versions of the Distributional Hypothesis (Lenci, 2008):
  - Weak: Correlation between context and word meaning (Spence and Owens, 1990)
  - Strong: Causality attributed to contextual distributions (Miller and Charles, 1991)

- Theory of (linguistic) meaning as "usage" (Wittgenstein) "the meaning of a word is defined by the circumstances of its use" (Manning and Schütze, 1999)

- Context is assumed to be the restricted domain or scope within which entities of the same nature can be presented together ("co-occur"), in such a way that they can be associated by a cognitive agent.

3.4 Distributionalism Vs. Context Co-Occurence

"Whereas LSA starts with a kind of co-occurrence, that of words with passages, the analysis produces a result in which the fact that two words appear in the same passage is not what makes them similar"(Landauer et al., 2007)

99% of the word-pairs for which LSA can establish a high similarity never appear together in the same context (Dennis et al., 2003)

"radius of the sphere"

"a circle's diameter"	0.55
"music of the spheres"	0.03

Matrix and Analogy

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

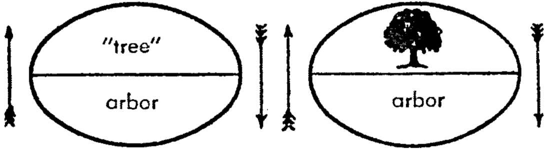
your : house
my : apartment

	...	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	...

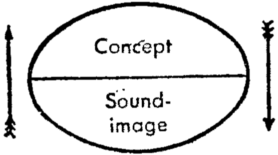
3 Distributionalism Is the Best Theoretical Candidate to Study LLMs

3.1 All linguistic properties of an LLM come from distributions in data
3.2 Distributionalism is often associated to contexts
3.3 Contexts are often understood cognitively or pragmatically
3.4 The global character of distributional properties challenges cognitive and pragmatic interpretations
3.5 Distributionalism is not a thesis about cognition, but about the structure of language

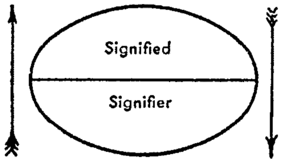
Saussure's Sign



Saussure’s Sign



Saussure's Sign

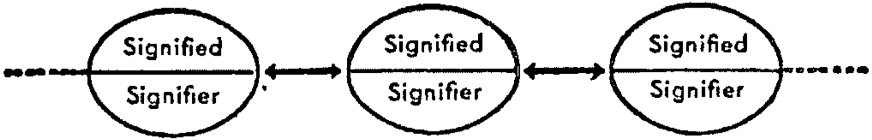


Saussure's Sign

"But here is the paradox: on the one hand the concept seems to be the counterpart of the sound-image, and on the other hand the sign itself is in turn the counterpart of the other signs of language.

Language is a system of interdependent terms in which the value of each term results solely from the simultaneous presence of the others, as in the diagram:"

(F. d. Saussure, 1959, p. 114)



The Language (*Langue*) Hypothesis

“What is both the integral and concrete object of linguistics? The question is especially difficult [...]”

“As I see it there is only one solution to all the foregoing difficulties: *from the very outset we must put both feet on the ground of language and use language [langue] as the norm of all other manifestations of speech*. Actually, among so many dualities, language alone seems to lend itself to independent definition and provide a fulcrum that satisfies the mind.”

(F. d. Saussure, 1959, p. 8-9)

The Language (*Langue*) Hypothesis

"But what is language [*langue*]? It is not to be confused with human speech [*langage*], of which it is only a definite part, though certainly an essential one. It is both a social product of the faculty of speech and a collection of necessary conventions that have been adopted by a social body to permit individuals to exercise that faculty. Taken as a whole, speech is many-sided and heterogeneous; straddling several areas simultaneously—physical, physiological, and psychological—it belongs both to the individual and to society; we cannot put it into any category of human facts, for we cannot discover its unity.

Language (*langue*), on the contrary, is a self-contained whole and a principle of classification. As soon as we give language first place among the facts of speech, we introduce a natural order into a mass that lends itself to no other classification."

(F. d. Saussure, 1959, p. 9)

The nominative form of Latin *honor*, for instance, is analogical. Speakers first said *honōs : honōsem*, then through rhotacization of the *s*, *honōs : honōrem*. After that, the radical had a double form. This duality was eliminated by the new form honor, created on the pattern of *ōrātor : ōrātōrem*, etc., through a process which subsequently will be set up as a proportion:

$$\textit{ōrātōrem} : \textit{ōrātor} = \textit{honōrem} : x$$
$$x = \textit{honor}$$

Thus analogy, to offset the diversifying action of a phonetic change (*honōs : honōrem*), again unified the forms and restored regularity (*honor : honōrem*).

(F. d. Saussure, 1959, p. 161)

4.2 The Idea of Virtually Structured Distributions Is at the Heart of Classical Structuralism

4.21 Saussure’s notion of sign is intrinsically distributional

4.22 “Langue” as a virtual structure behind distribution is the very object of Saussurean linguistics

4.23 Analogical operations local operators of such virtual a structure

4.3 From the Distributional to the Structuralist Hypothesis

Distributional Hypothesis

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

$$\Downarrow$$

Structuralist Hypothesis

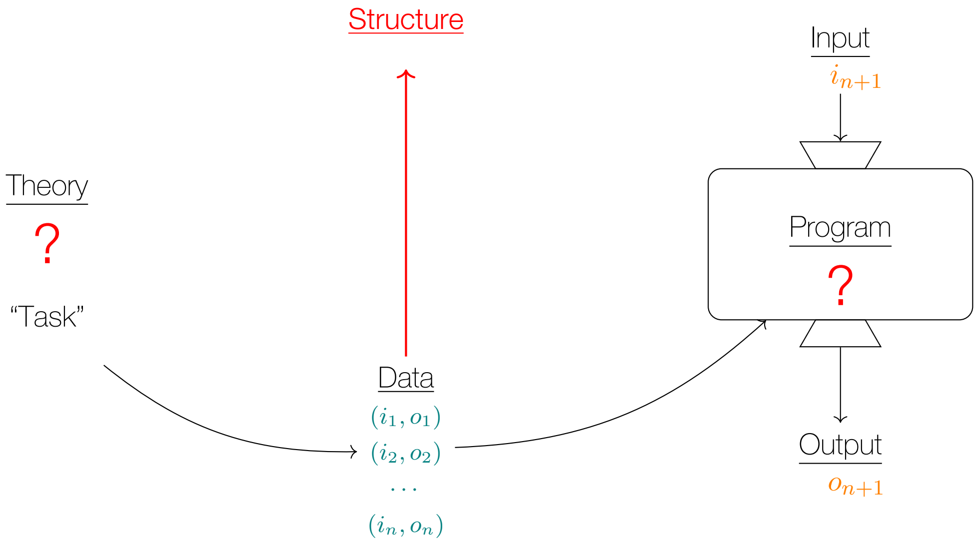
Linguistic content is the effect of a virtual *structure* underlying linguistic practices within a community

"A priori it would seem to be a generally valid thesis that for every *process* there is a corresponding *system*, by which the process can be analyzed and described by means of a limited number of premises. It must be assumed that any process, can be analyzed into a limited number of elements recurring in various combinations. Then, on the basis of this analysis, it should be possible to order these elements into classes according to their possibilities of combination. And it should be further possible to set up a general and exhaustive calculus of the possible combinations."

(Hjelmslev, 1953, p. 9)

The Structuralist Hypothesis

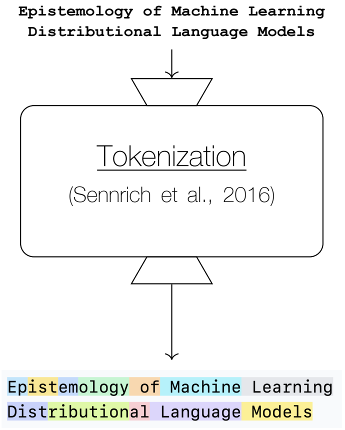
- Meaning is the effect of structure
- Distributional properties convey meaning only through the action of a latent structure determining possible semantic values, and which is inseparable from the principles of identification of the elementary units of language, since meaning is the effect of discriminating operations performed through segmentation procedures of which the units of language keep the trace
- Linguistic content is the effect of a virtual structure of classes and dependencies at multiple levels underlying (and derivable from) the mass of things said or written in a given language


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}

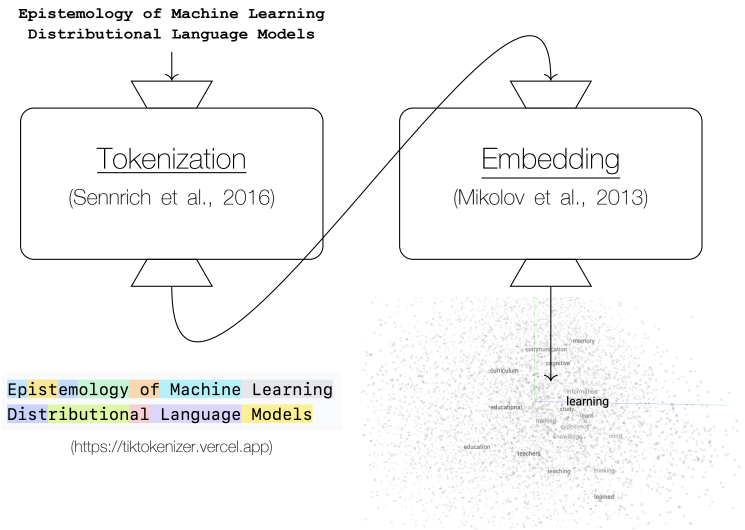
4 Distributionalism Is a Corollary of Structuralism

4.1 The source of distributional properties is a virtual structure
4.2 The idea of virtually structured distributions is at the heart of classical structuralism
4.3 We need to move on from the distributional hypothesis to the structuralist hypothesis

Formal Explainability

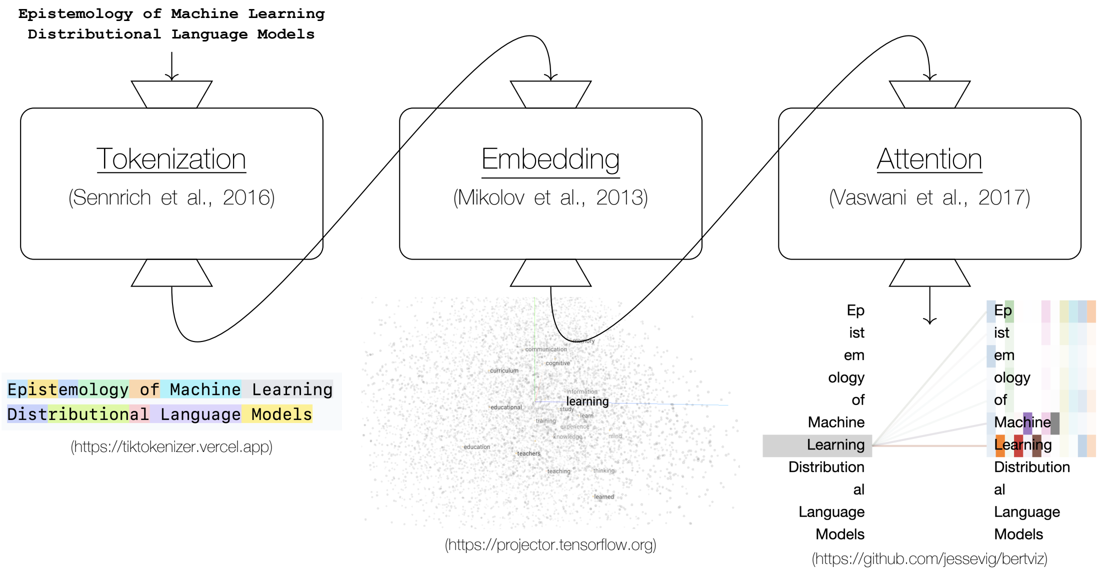


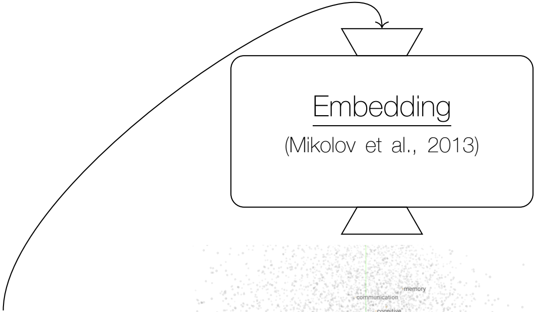
(https://tiktokenizer.vercel.app)



(https://projector.tensorflow.org)

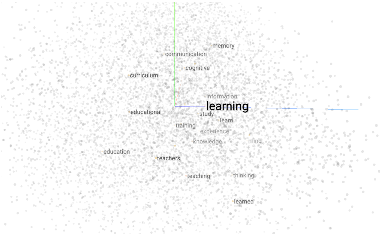
Formal Explainability





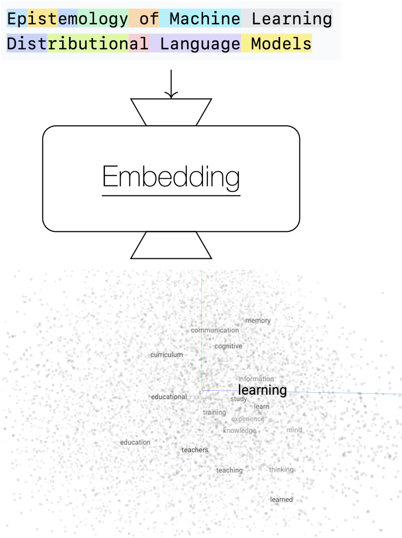
Epistemology of Machine Learning
Distributional Language Models

(https://tiktokenizer.vercel.app)



(https://projector.tensorflow.org)

Formal Explainability

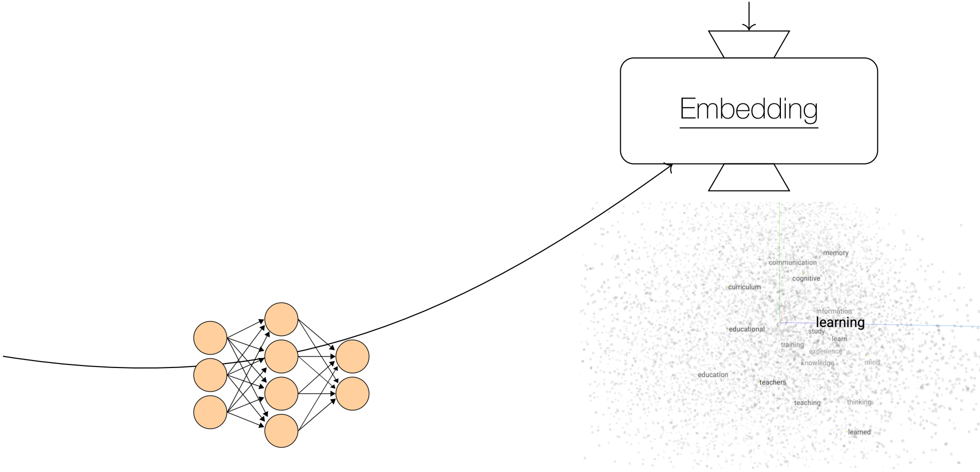


Formal Explainability

Structure

?

Epistemology of Machine Learning
Distributional Language Models



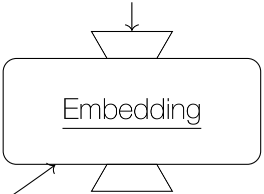
Data

Formal Explainability

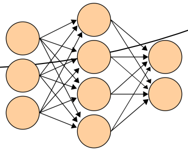
Structure

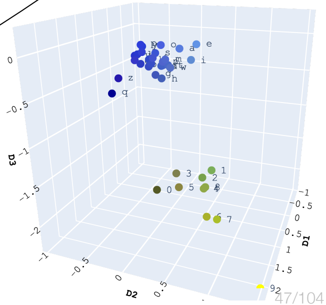
?

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



Data





Word2vec Explained

(Levy & Goldberg, 2014)

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

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

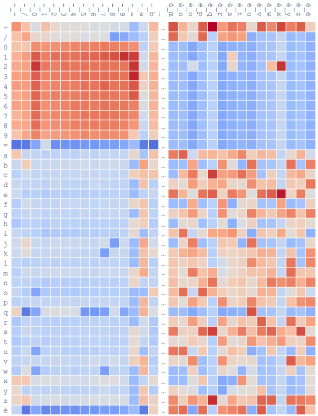
- Word2vec performs an implicit, low-dimensional factorization of a pointwise mutual information (pmi), word-context matrix.
- The Singular Value Decomposition (SVD) provides an exact solution to this problem.

Example: Characters in Wikipedia

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

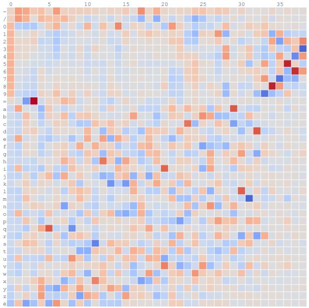
$$C = X \times X = \{\texttt{(-,-)}, \texttt{(-,/)}, \texttt{(-,0)}, \ldots, \texttt{(é,z)}, \texttt{(é,é)}\}$$

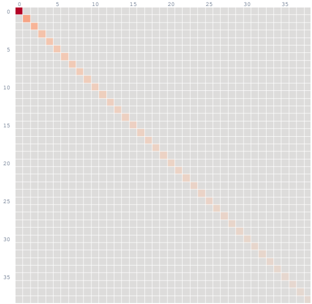
$$M_{wc} = \mathrm{pmi}(w, c) = \log \frac{p(w, c)}{p(w)p(c)}$$

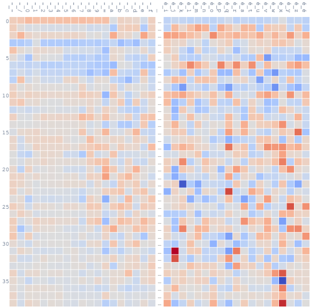


SVD of Wikipedia Character PMI Matrix

U Σ V^T

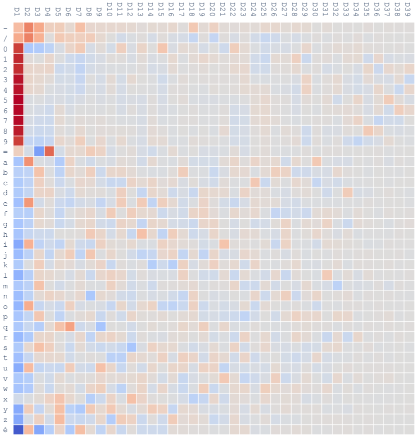






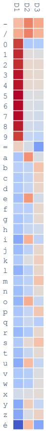
Truncate

$$U \times \Sigma$$



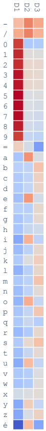
Truncate

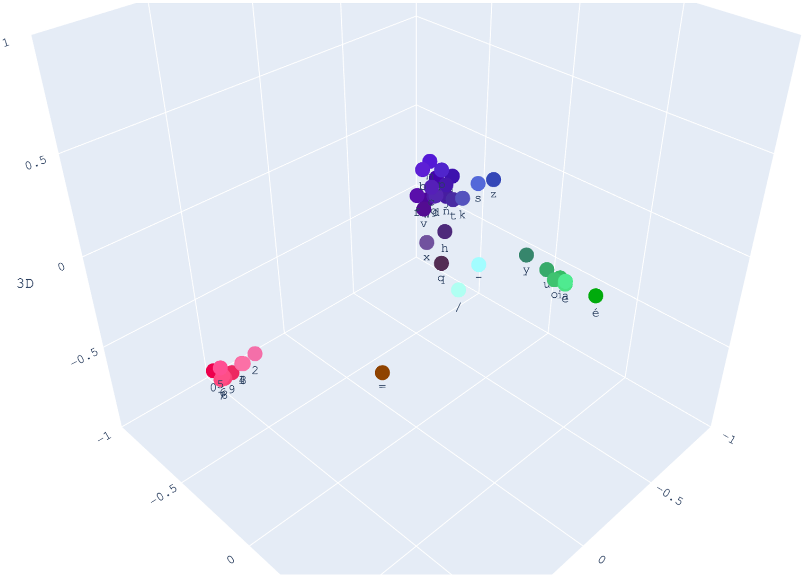
$$\hat{U} \times \hat{\Sigma}$$



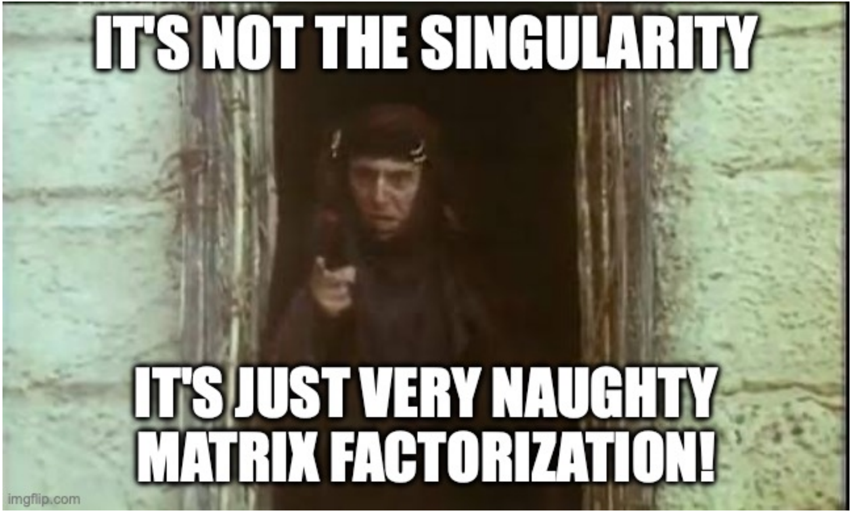
Plot

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

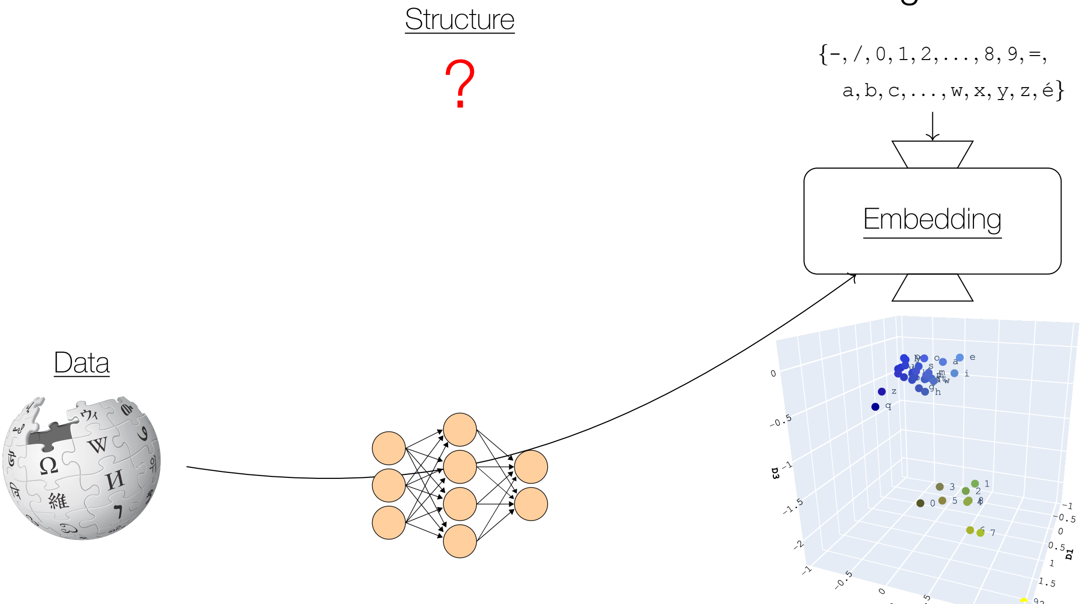




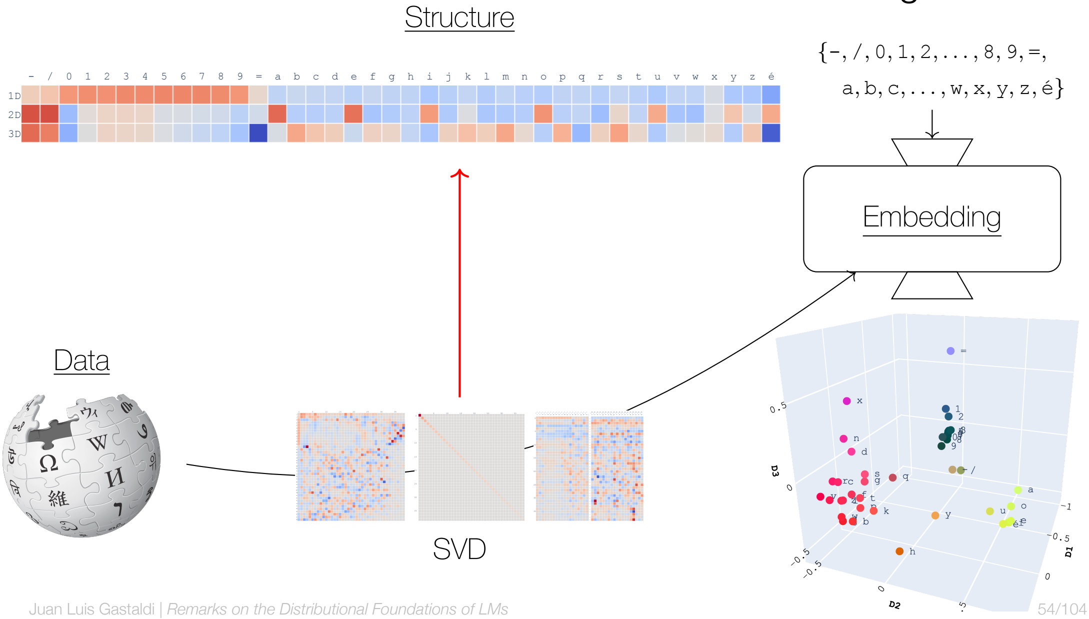
What to Conclude?



Embedding Structure



Embedding Structure



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.

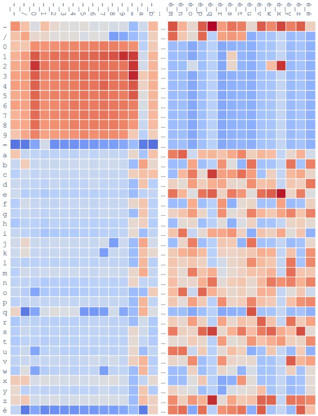
(Goldberg and Levy, 2014)

Embeddings as Functions Over Sets

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

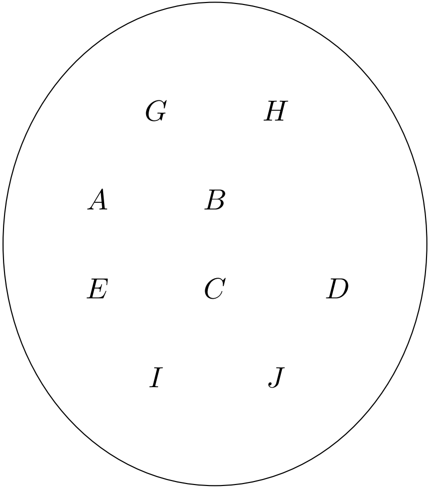
$$Y = X \times X = \{\texttt{(-,-)}, \texttt{(-,/)}, \texttt{(-,0)}, \ldots, \texttt{(é,z)}, \texttt{(é,é)}\}$$

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

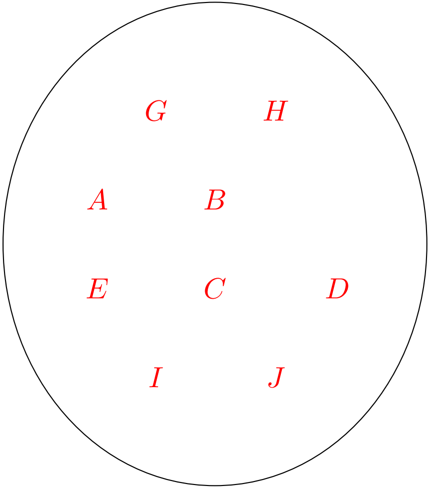


5.51

A Category Is Like a Set With Structure



A Category Is Like a Set With Structure

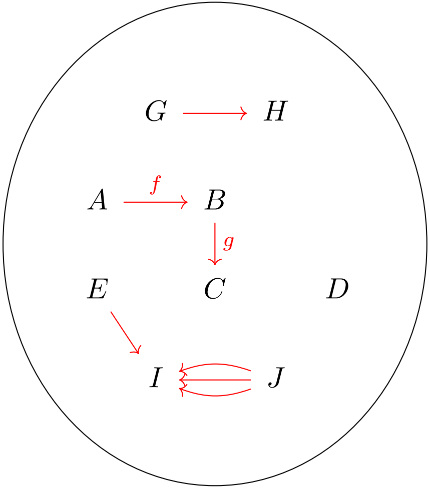


Definition (Category – Awodey, 2010)

Data:

- Objects: $A, B, C, \ldots$

A Category Is Like a Set With Structure

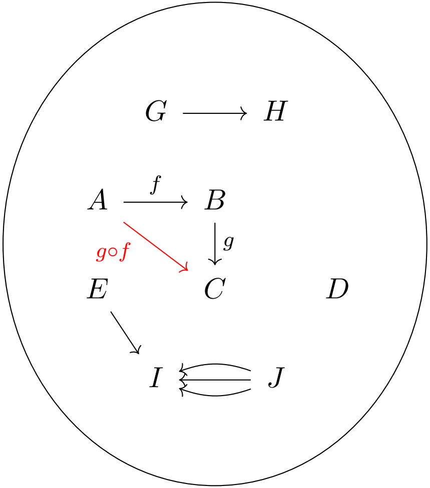


Definition (Category – Awodey, 2010)

Data:

- Objects: $A, B, C, \ldots$
- Arrows: $f, g, \ldots$

A Category Is Like a Set With Structure



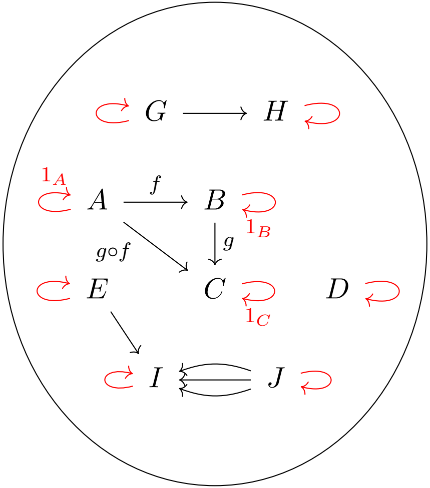
Definition (Category – Awodey, 2010)

Data:

- Objects: $A, B, C, \ldots$
- Arrows: $f, g, \ldots$
- Composition: Given $f : A \to B$ and $g : B \to C$, there is given an arrow

$$g \circ f : A \to C$$

A Category Is Like a Set With Structure



Definition (Category – Awodey, 2010)

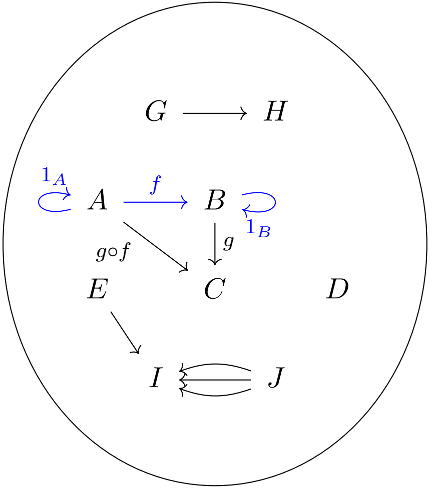
Data:

- Objects: $A, B, C, \dots$
- Arrows: $f, g, \dots$
- Composition: Given $f : A \to B$ and $g : B \to C$, there is given an arrow

$$g \circ f : A \to C$$

- Identity: For each A, there is $1_A : A \to A$

A Category Is Like a Set With Structure



Definition (Category – Awodey, 2010)

Data:

- Objects: $A, B, C, \ldots$
- Arrows: $f, g, \ldots$
- Composition: Given $f : A \to B$ and $g : B \to C$, there is given an arrow

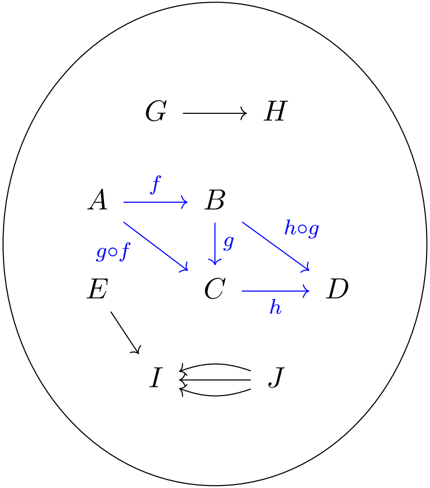
$$g \circ f : A \to C$$

- Identity: For each A, there is $1_A : A \to A$

Laws:

- Unit: $f \circ 1_A = f = 1_B \circ f$

A Category Is Like a Set With Structure



Definition (Category – Awodey, 2010)

Data:

- Objects: $A, B, C, \ldots$
- Arrows: $f, g, \ldots$
- Composition: Given $f : A \to B$ and $g : B \to C$, there is given an arrow

$$g \circ f : A \to C$$

- Identity: For each A, there is $1_A : A \to A$

Laws:

- Unit: $f \circ 1_A = f = 1_B \circ f$
- Associativity: $f \circ (g \circ h) = (f \circ g) \circ h$

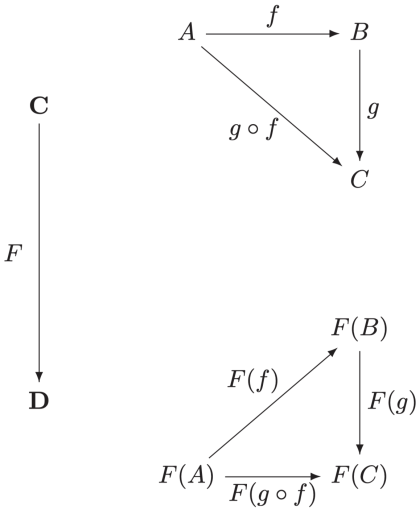
A Functor Is a Map Between Categories

Definition (Functor – Awodey, 2010)

A *functor*

$$F: \mathsf{C} \to \mathsf{D}$$

between categories C and D is a mapping of objects to objects and arrows to arrows, in such a way that



A Functor Is a Map Between Categories

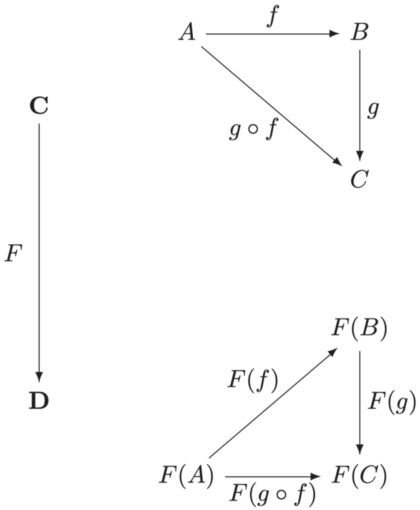
Definition (Functor – Awodey, 2010)

A *functor*

$$F : \mathsf{C} \to \mathsf{D}$$

between categories C and D is a mapping of objects to objects and arrows to arrows, in such a way that

(a) $F(f : A \to B) = F(f) : F(A) \to F(B)$



A Functor Is a Map Between Categories

Definition (Functor – Awodey, 2010)

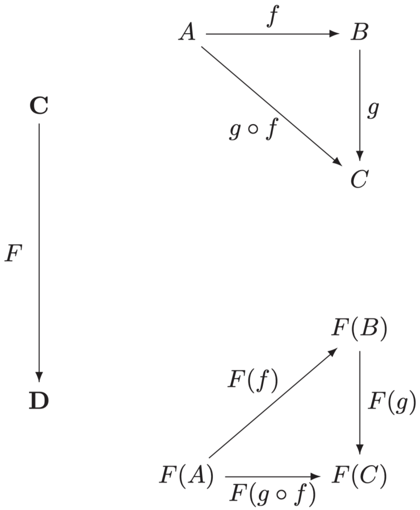
A *functor*

$$F : \mathbf{C} \to \mathbf{D}$$

between categories **C** and **D** is a mapping of objects to objects and arrows to arrows, in such a way that

(a) $F(f : A \to B) = F(f) : F(A) \to F(B)$

(b) $F(1_A) = 1_{F(A)}$



A Functor Is a Map Between Categories

Definition (Functor – Awodey, 2010)

A *functor*

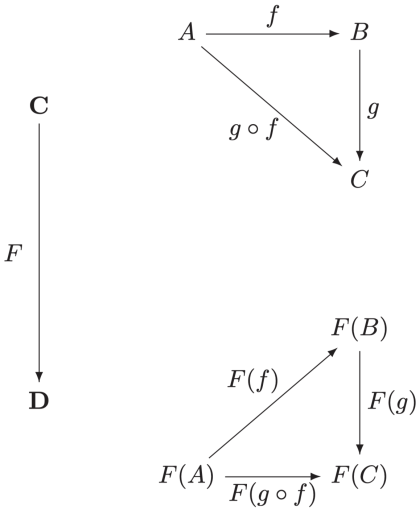
$$F : \mathbf{C} \to \mathbf{D}$$

between categories **C** and **D** is a mapping of objects to objects and arrows to arrows, in such a way that

(a) $F(f : A \to B) = F(f) : F(A) \to F(B)$

(b) $F(1_A) = 1_{F(A)}$

(c) $F(g \circ f) = F(g) \circ F(f)$



Product of Categories

Definition 2.15. In any category $\mathbf{C}$, a *product diagram* for the objects A and B consists of an object P and arrows

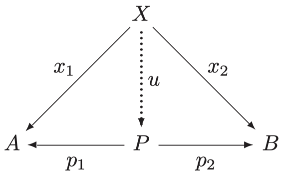
$$A \xleftarrow{\;p_1\;} P \xrightarrow{\;p_2\;} B$$

satisfying the following UMP:

Given any diagram of the form

$$A \xleftarrow{\;x_1\;} X \xrightarrow{\;x_2\;} B$$

there exists a unique $u : X \to P$, making the diagram



commute, that is, such that $x_1 = p_1 u$ and $x_2 = p_2 u$.

(Awodey, 2010)

A Profunctor Is a Functor From the Product of Two Arbitrary Categories to the Set Category

$term_i$ → C^{op}

$context_i$ → D

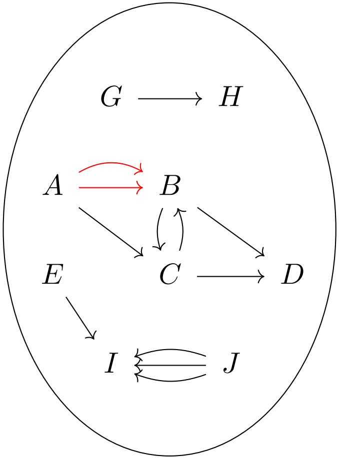
measure → Set

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

5.54 A Category Enriched Over $\mathcal{V}$ Is a Category Having a $v_{\in\mathcal{V}}$'s Worth Arrows Between Two Objects

$$\mathsf{hom}(A, B)$$

$$\mathsf{C}(A, B) = \{f \in \mathsf{C} | f : A \to B\}$$

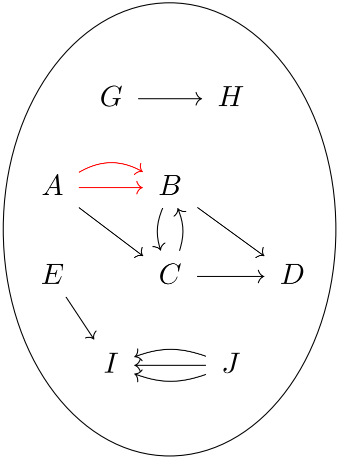


5.54 A Category Enriched Over $\mathcal{V}$ Is a Category Having a $v_{\in\mathcal{V}}$'s Worth Arrows Between Two Objects

$\mathsf{hom}(A, B)$

$\mathsf{C}(A, B) \quad = \{f \in \mathsf{C} | f : A \to B\}$

$\mathsf{C}(A, B) \in \mathsf{Set}$



5.54 A Category Enriched Over $\mathcal{V}$ Is a Category Having a $v_{\in \mathcal{V}}$'s Worth Arrows Between Two Objects

$\mathsf{hom}(A, B)$

$\mathsf{C}(A, B) \quad = \{f \in \mathsf{C} | f : A \to B\}$

$\mathsf{C}(A, B) \in \mathsf{Set}$

Enrichment over $\mathcal{V}$

$\mathsf{C}(A, B) \in \mathcal{V}$,

where $\mathcal{V}$ is a "nice" (monoidal) category

A Functor Between the Enriched Categories $\mathsf{D} \to \mathsf{C}$ Induces a Profunctor Is $\mathsf{C}^{\mathrm{op}} \times \mathsf{D} \to \mathcal{V}$

$term_i$ $context_i$ measure

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

5.55 A Functor Between the Enriched Categories $\mathsf{D} \to \mathsf{C}$ Induces a Profunctor Is $\mathsf{C}^{\mathrm{op}} \times \mathsf{D} \to \mathcal{V}$

$$\overset{term_i}{\mathsf{C}^{\mathrm{op}}} \times \overset{context_i}{\mathsf{D}} \to \overset{\text{measure}}{\mathcal{V}}$$

A Functor Between the Enriched Categories $\mathsf{D} \to \mathsf{C}$ Induces a Profunctor Is $\mathsf{C}^{\mathrm{op}} \times \mathsf{D} \to \mathcal{V}$

$term_i$ $context_i$ measure

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

5.55 A Functor Between the Enriched Categories $\mathsf{D} \to \mathsf{C}$ Induces a Profunctor Is $\mathsf{C}^{\mathrm{op}} \times \mathsf{D} \to \mathcal{V}$

$$\underbrace{\mathsf{C}^{\mathrm{op}}}_{term_i} \times \underbrace{\mathsf{D}}_{context_i} \to \underbrace{\bar{\mathbb{R}}}_{\text{measure}}$$

5.5 We Can Generalize Matrices to Enriched Profunctors $: \mathsf{C}^{\mathrm{op}} \times \mathsf{D} \to \mathcal{V}$

5.51 A category is like a set with structure

5.52 A functor is a map between categories

5.53 A profunctor is a functor from the product of two arbitrary categories to the Set category

5.54 A category enriched over $\mathcal{V}$ is a category having a $v_{\in\mathcal{V}}$'s worth arrows between two objects

5.55 A functor between the enriched categories $\mathsf{D} \to \mathsf{C}$ induces a profunctor is $\mathsf{C}^{\mathrm{op}} \times \mathsf{D} \to \mathcal{V}$

5 The General Form of Distributions Is $\mathcal{M}\colon \mathsf{C}^{\mathrm{op}} \times \mathsf{D} \to \mathcal{V}$

5.1 The (formal) key of neural LMs lies on embeddings

5.2 SVD over a PMI matrix provides the formal explanation for words embeddings

5.3 This result has important consequences for explainability

5.4 A matrix can be understood as a function $M\colon X \times Y \to \mathbb{R}$

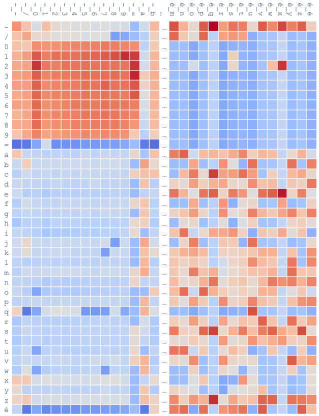
5.5 We can generalize matrices to enriched profunctors $\colon \mathsf{C}^{\mathrm{op}} \times \mathsf{D} \to \mathcal{V}$

From Matrices to Distributional Operators

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

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

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



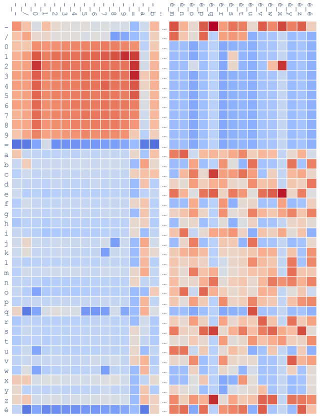
From Matrices to Distributional Operators

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

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

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

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



6.1 From Matrices to Distributional Operators

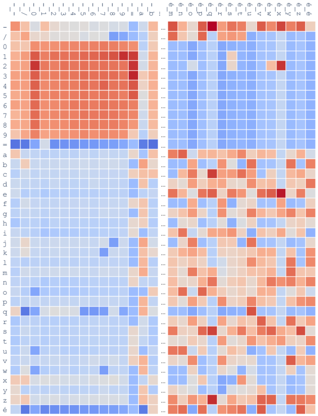
$$X = \{\texttt{-}, \texttt{/}, \texttt{0}, \texttt{1}, \texttt{2}, \texttt{3}, \texttt{4}, \texttt{5}, \texttt{6}, \texttt{7}, \texttt{8}, \texttt{9}, \texttt{=}, \texttt{a}, \texttt{b}, \texttt{c}, \ldots, \texttt{w}, \texttt{x}, \texttt{y}, \texttt{z}, \texttt{é}\}$$

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

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

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

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



From Matrices to Distributional Operators

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

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

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

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

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

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

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

6.1 From Matrices to Distributional Operators

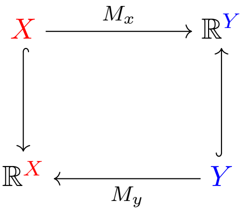
$$X = \{\texttt{-}, \texttt{/}, \texttt{0}, \texttt{1}, \texttt{2}, \texttt{3}, \texttt{4}, \texttt{5}, \texttt{6}, \texttt{7}, \texttt{8}, \texttt{9}, \texttt{=}, \texttt{a}, \texttt{b}, \texttt{c}, \ldots, \texttt{w}, \texttt{x}, \texttt{y}, \texttt{z}, \texttt{é}\}$$

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

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

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

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



6.1 From Matrices to Distributional Operators

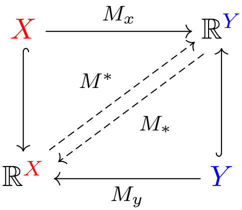
$$X = \{\texttt{-}, \texttt{/}, \texttt{0}, \texttt{1}, \texttt{2}, \texttt{3}, \texttt{4}, \texttt{5}, \texttt{6}, \texttt{7}, \texttt{8}, \texttt{9}, \texttt{=}, \texttt{a}, \texttt{b}, \texttt{c}, \ldots, \texttt{w}, \texttt{x}, \texttt{y}, \texttt{z}, \texttt{é}\}$$

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

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

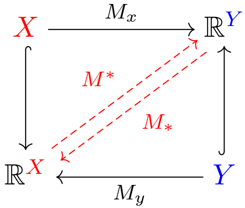
$$\begin{aligned} M_x \colon X &\to \mathbb{R}^Y \\ x &\mapsto M(x, -) \end{aligned}$$

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



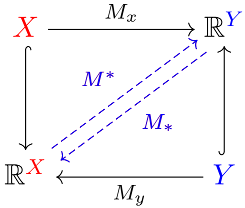
$$M^* \colon \mathbb{R}^X \to \mathbb{R}^Y$$
$$M_* \colon \mathbb{R}^Y \to \mathbb{R}^X$$

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



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

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



Fixed Points of Distributional Operators as Structure

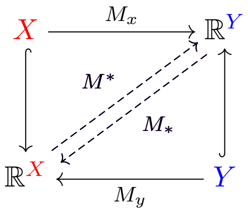
$$M_* M^* : \mathbb{R}^X \to \mathbb{R}^X$$

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

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

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

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



6.2 Fixed Points of Distributional Operators as Structure

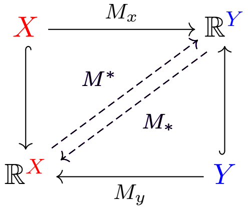
$$M_*M^*: \mathbb{R}^X \to \mathbb{R}^X$$

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

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

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

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



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

Fixed Points of Distributional Operators as Structure

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

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

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

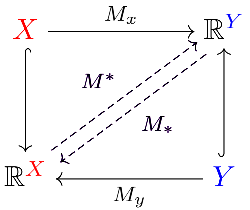
$$\{v_1, \ldots, v_n\} \subset \mathbb{R}^Y$$

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

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

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

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



$$M = U \Sigma V^T \qquad U := [u_1, \ldots, u_m] \qquad V := [v_1, \ldots, v_n]$$

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

Embeddings as Functors Over Categories

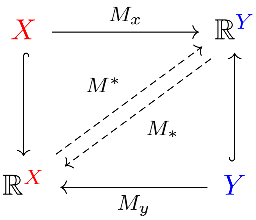
$$X = \{\texttt{-}, \texttt{/}, \texttt{0}, \texttt{1}, \texttt{2}, \texttt{3}, \texttt{4}, \texttt{5}, \texttt{6}, \texttt{7}, \texttt{8}, \texttt{9}, \texttt{=}, \texttt{a}, \texttt{b}, \texttt{c}, \ldots, \texttt{w}, \texttt{x}, \texttt{y}, \texttt{z}, \texttt{é}\}$$

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

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

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

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



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

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

Embeddings as Functors Over Categories

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

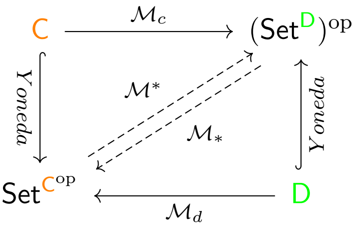
$$\mathsf{D} = \mathsf{C} = \{\texttt{-}, \texttt{/}, \texttt{0}, \texttt{1}, \texttt{2}, \texttt{3}, \texttt{4}, \texttt{5}, \texttt{6}, \texttt{7}, \texttt{8}, \texttt{9}, \texttt{=}, \texttt{a}, \texttt{b}, \texttt{c}, \ldots, \texttt{w}, \texttt{x}, \texttt{y}, \texttt{z}, \texttt{é}\}$$

Profunctor

$$\mathcal{M} \colon \mathsf{C}^{\mathrm{op}} \times \mathsf{D} \to \mathsf{Set}$$
$$(c, d) \mapsto \mathcal{M}(c, d)$$

$$\mathcal{M}_c \colon \mathsf{C} \to \left(\mathsf{Set}^{\mathsf{D}}\right)^{\mathrm{op}}$$
$$c \mapsto \mathcal{M}(c, -)$$

$$\mathcal{M}_d \colon \mathsf{D} \to \mathsf{Set}^{\mathsf{C}^{\mathrm{op}}}$$
$$d \mapsto \mathcal{M}(-, d)$$



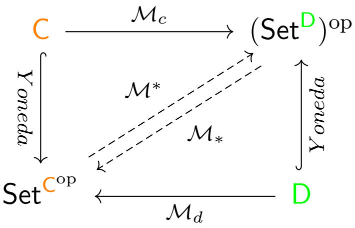
$$\mathcal{M}^* \colon \mathsf{Set}^{\mathsf{C}^{\mathrm{op}}} \to (\mathsf{Set}^{\mathsf{D}})^{\mathrm{op}}$$
$$\mathcal{M}_* \colon (\mathsf{Set}^{\mathsf{D}})^{\mathrm{op}} \to \mathsf{Set}^{\mathsf{C}^{\mathrm{op}}}$$

Isbell Adjunction

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

$\mathcal{M}_*\mathcal{M}^*: \mathsf{Set}^{\mathsf{C}^{op}} \to \mathsf{Set}^{\mathsf{C}^{op}}$

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



Isbell Adjunction

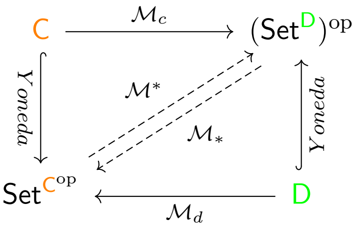
$$\mathcal{M}^* : \mathsf{Set}^{\mathsf{C}^{\mathrm{op}}} \leftrightarrows (\mathsf{Set}^{\mathsf{D}})^{\mathrm{op}} : \mathcal{M}_*$$

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

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

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

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



Nucleus of $\mathcal{M} = \{(f_i, g_i)\}$, such that:

$\mathcal{M}^* f_i \cong g_i$ and $\mathcal{M}_* g_i \cong f_i$

Isbell Adjunction

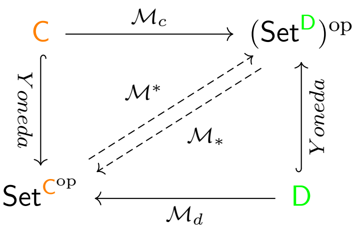
$$\mathcal{M}^*: \mathsf{Set}^{\mathsf{C}^{\mathrm{op}}} \leftrightarrows (\mathsf{Set}^{\mathsf{D}})^{\mathrm{op}}: \mathcal{M}_*$$

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

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

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

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



Nucleus of $\mathcal{M} = \{(f_i, g_i)\}$, such that:

$\mathcal{M}^* f_i \cong g_i$ and $\mathcal{M}_* g_i \cong f_i$

The nucleus is a category complete and cocomplete

Embeddings as Functors Over Categories

Isbell Adjunction

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

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

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

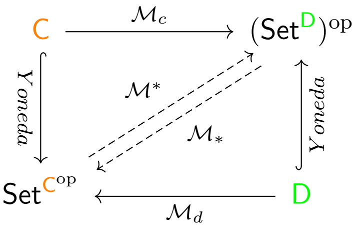
$$\mathrm{Fix}(\mathcal{M}_*\mathcal{M}^*) := \{f \in \mathsf{Set}^{\mathsf{C}^{\mathrm{op}}} | \mathcal{M}_*\mathcal{M}^*(f) \cong f\}$$

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

Nucleus of $\mathcal{M} = \{(f_i, g_i)\}$, such that:

$\mathcal{M}^* f_i \cong g_i$ and $\mathcal{M}_* g_i \cong f_i$

The nucleus is a category complete and cocomplete



Categories C and D can be enriched!

E.g.:

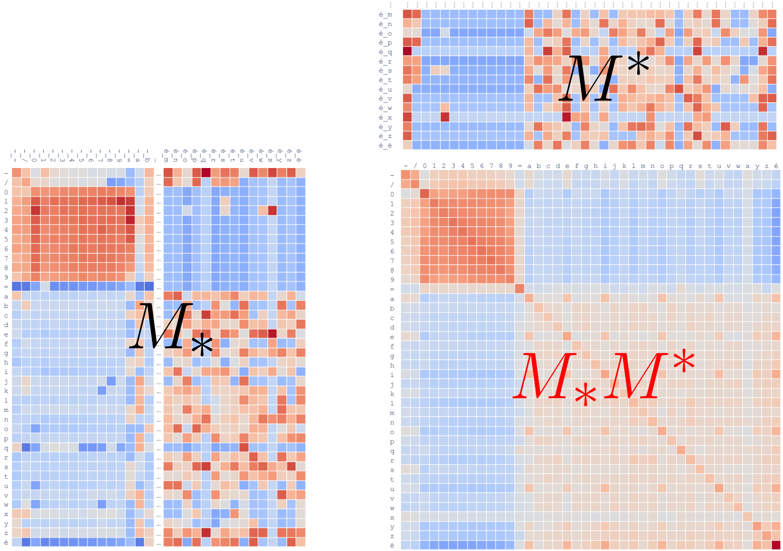
$$\mathcal{M}^*: 2^{\mathsf{C}^{\mathrm{op}}} \leftrightarrows (2^{\mathsf{D}})^{\mathrm{op}} : \mathcal{M}_*$$

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

6 The General Form of Structures Is $\mathcal{M}^*\colon \mathcal{V}^{\mathsf{C}} \leftrightarrows \mathcal{V}^{\mathsf{D}}\colon \mathcal{M}_*$

6.1 SVD looks for linear fixed points of the linear operators M^*M_* and M_*M^*
6.2 The set of fixed points reveals (limited) structural features underlying the distributions
6.3 The nucleus of an enriched profunctor provides a generalization of this setting

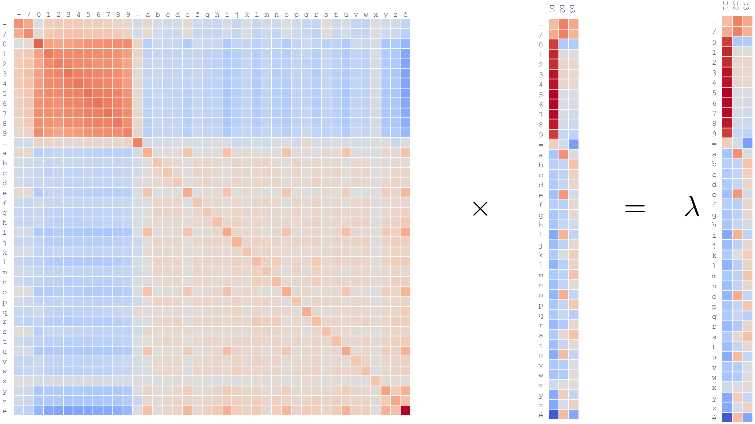
The Operator $M_* M^*$ Is a Covariance Matrix



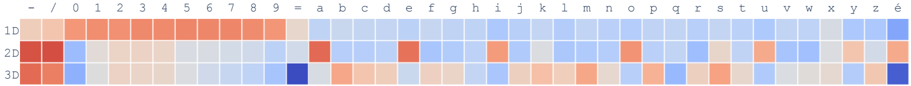
7.1

Eigenvectors as Fixed Points

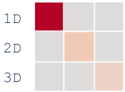
$$M_* M^* u = \lambda u$$



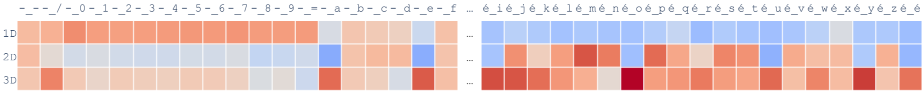
Eigenvectors of $M_* M^*$:



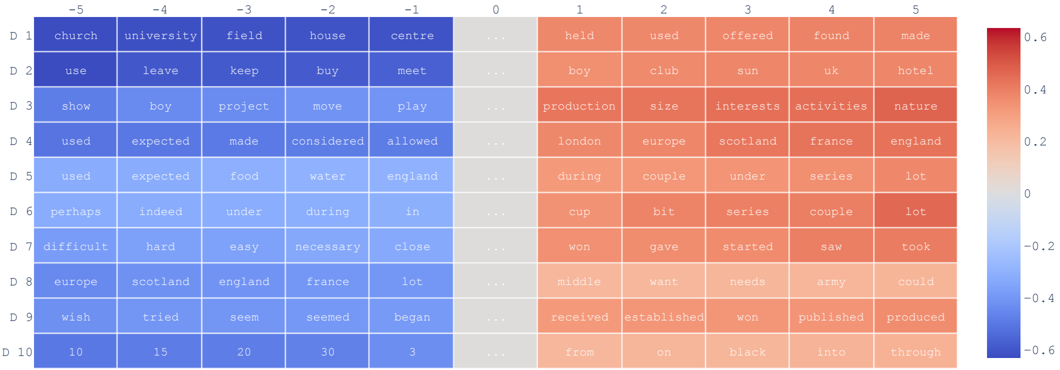
Eigenvalues of $M_* M^*$ and $M^* M_*$:



Eigenvectors of $M^* M_*$:

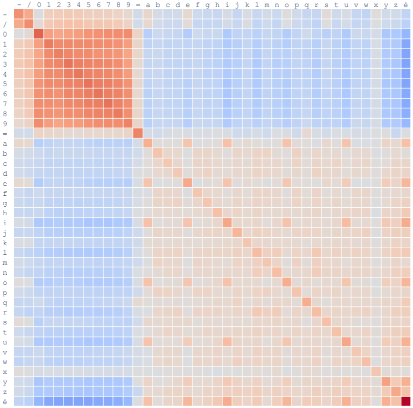


Words

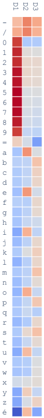

-5 -4 -3 -2 -1 0 1 2 3 4 5
D 1 church university field house centre ... held used offered found made
D 2 use leave keep buy meet ... boy club sun uk hotel
D 3 show boy project move play ... production size interests activities nature
D 4 used expected made considered allowed ... london europe scotland 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

Binary Fixed Points

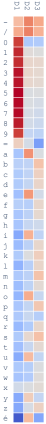
$$M_* M^* u = \lambda u$$



$\times$

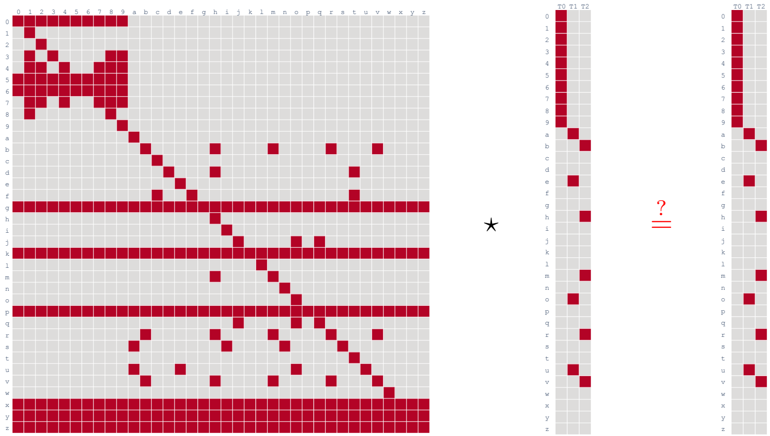


$= \lambda$



Binary Fixed Points

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

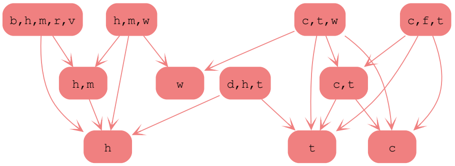


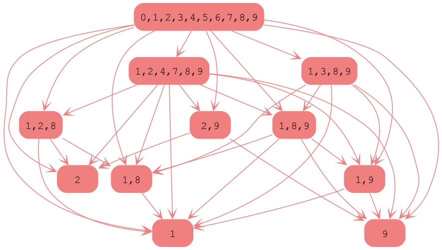
“Eigensets”

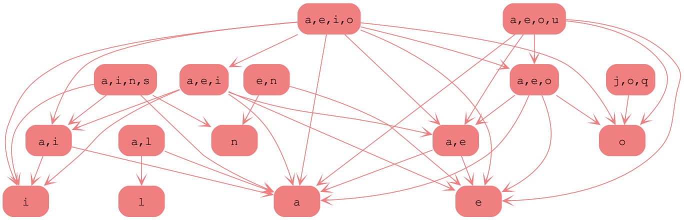
$$\mathcal{M}_*\mathcal{M}^* f = f$$

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				

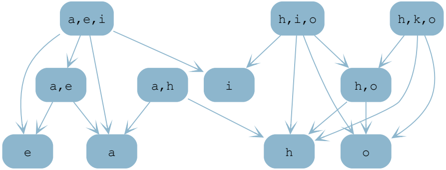
Partial Order Structure

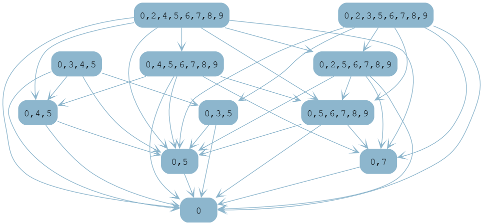


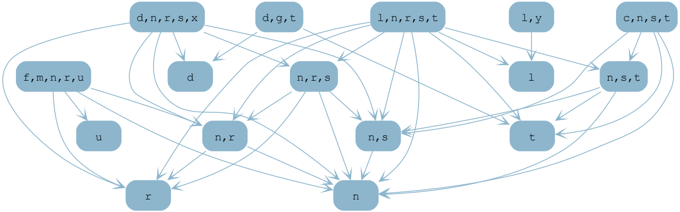




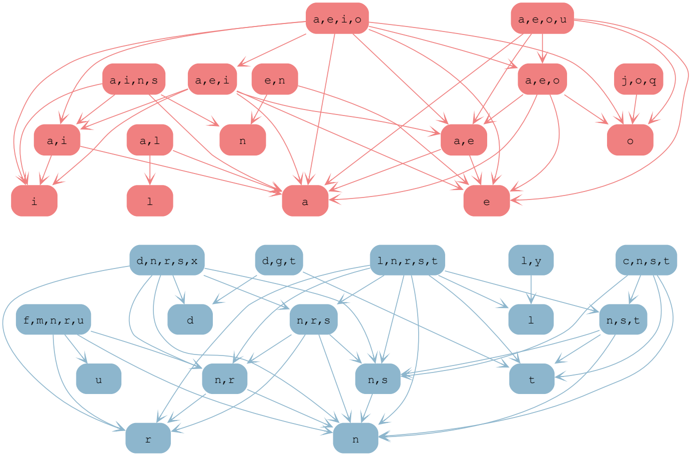
Dual Partial Order



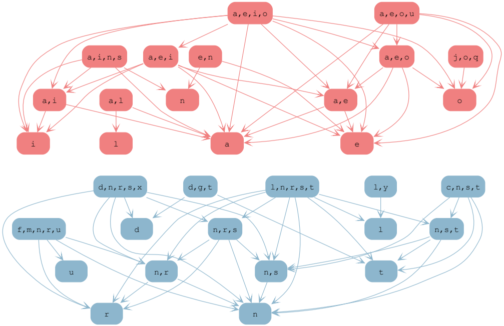




Paring of Partial Ordered Fixed Points



Structure



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

$$\Downarrow$$

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

Structure

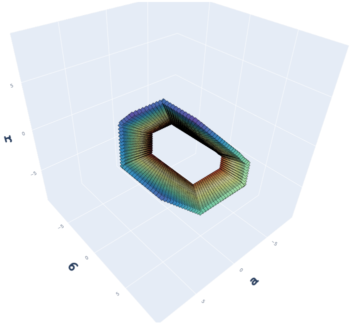
?

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

$$\Downarrow$$

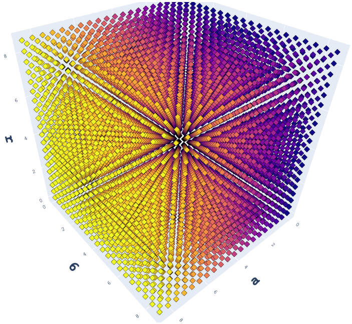
$$\mathcal{M}^*: \bar{\mathbb{R}}^{\mathsf{C}^{\mathrm{op}}} \leftrightarrows (\bar{\mathbb{R}}^{\mathsf{D}})^{\mathrm{op}} : \mathcal{M}_*$$

Structure



?

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

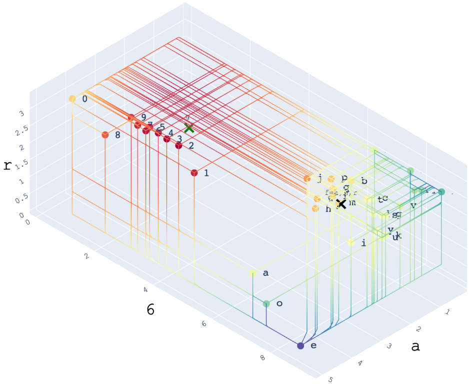


$$\mathsf{C}^{\mathrm{op}} \times \mathsf{D} \to \bar{\mathbb{R}}$$

$$\Downarrow$$

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

Structure



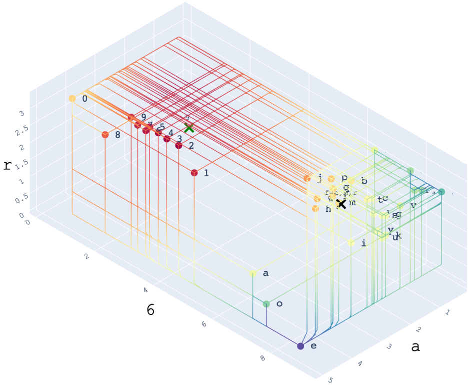
$$\mathsf{C}^{\mathrm{op}} \times \mathsf{D} \to \bar{\mathbb{R}}$$

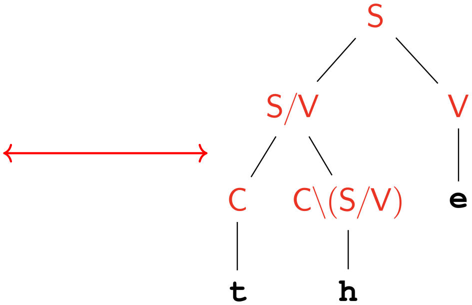
$$\Downarrow$$

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

7.3 The Profunctor's Nucleus Defines a System of Logical Types

Structure





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

$$\Downarrow$$

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

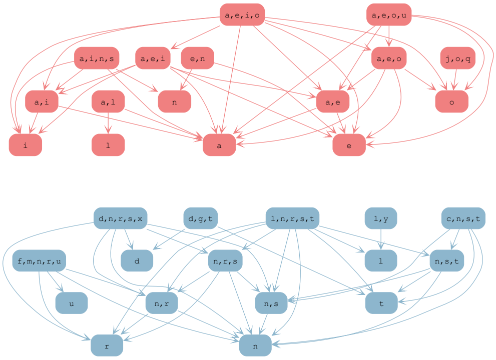
Definition (Polar/Orthogonal - Girard, 2011)

[G]iven a binary operation, noted $a, b \rightsquigarrow \langle a|b\rangle : A \times B \to C$ 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 :

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

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

- 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, $X^{\perp\perp\perp} = X^{\perp}$.



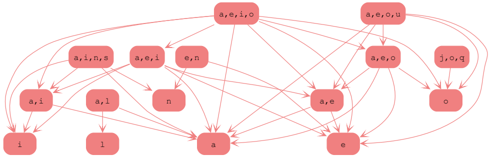
Definition (Polar/Orthogonal - Girard, 2011)

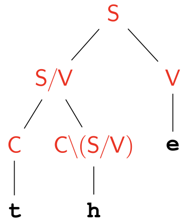
[G]iven a binary operation, noted $a, b \rightsquigarrow \langle a|b \rangle : A \times B \to C$ and a subset $P \subset C$ (the 'pole') one can define the *polar* $X^\perp \subset B$ of a subset $X \subset A$ (resp. $Y^\perp \subset A$ of a subset $Y \subset B$) by :

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

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

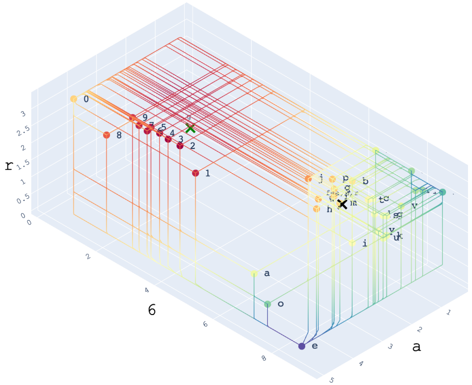
- The map 'polar' is decreasing: $X \subset X' \Rightarrow X'^\perp \subset X^\perp$.
- The set Pol$(A) \subset \mathcal{P}(A)$ of *polar* sets, i.e., of the form $Y^\perp$, is closed under arbitrary intersections. In particular, A is polar and $X^{\perp\perp}$ is the smallest polar set containing X.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

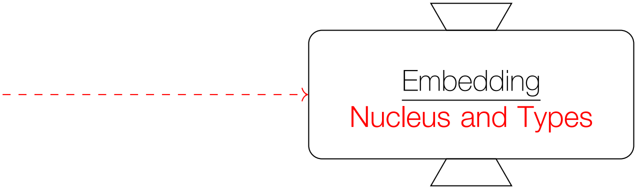




(Gastaldi and Pellissier, 2021)

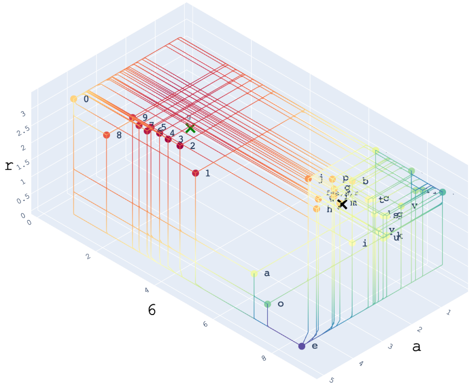
Structure

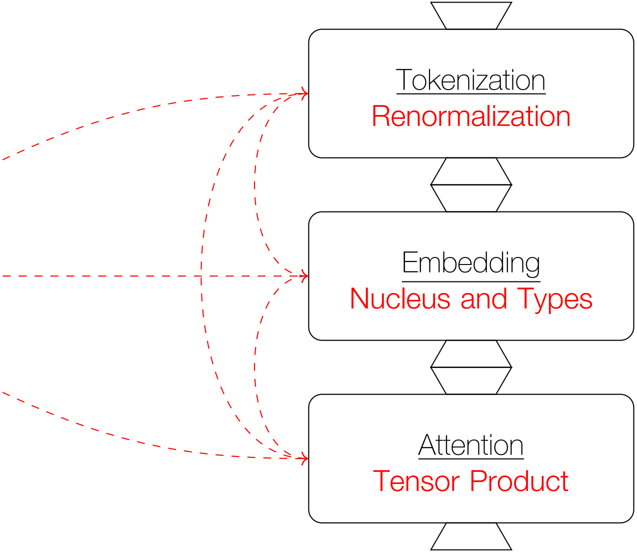




Formal Explainability

Structure





7.5 The Resulting Objects Correspond to Classical Structuralist Theoretical Constructs

Distributional Hypothesis

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

$$\Downarrow$$

Structuralist Hypothesis

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

7.5 The Resulting Objects Correspond to Classical Structuralist Theoretical Constructs

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

Structure

Distributional Hypothesis

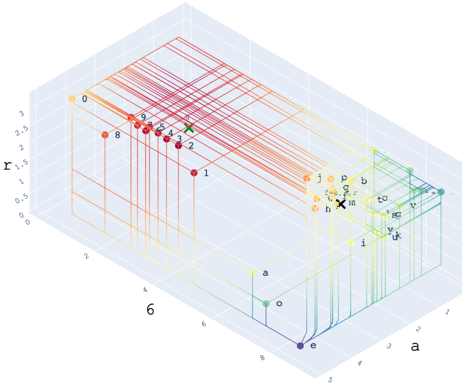
The content of linguistic units is determined by their *distribution* in a corpus.

$\Downarrow$

Theory
"Task"

Structuralist Hypothesis

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



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

7.5 Structuralist Semiology: Language (*Langue*) and Text

A Language [...] is the Paradigmatic of a Denotative Semiotic whose Paradigms are Manifested by all Purports.

(Hjelmslev, 1975, Df. 38)

A Text [...] is the Syntagmatic of a Denotative Semiotic whose Chains are Manifested by all Purports.

(Hjelmslev, 1975, Df. 39)

7.5 Structuralist Semiology: Language (*Langue*) and Text

A Language [...] is the Paradigmatic of a Denotative Semiotic whose Paradigms are Manifested by all Purports.

(Hjelmslev, 1975, Df. 38)

A Paradigmatic or Sign-System [...] is a Semiotic System.

(Hjelmslev, 1975, Df. 35)

A Text [...] is the Syntagmatic of a Denotative Semiotic whose Chains are Manifested by all Purports.

(Hjelmslev, 1975, Df. 39)

A Syntagmatic or Sign-Process [...] is a Semiotic Process.

(Hjelmslev, 1975, Df. 33)

A Semiotic [...] is a Hierarchy, any of whose Components admits of a further Analysis into Classes defined by mutual Relation, so that any of these classes admits of an analysis into Derivates defined by mutual Mutation.

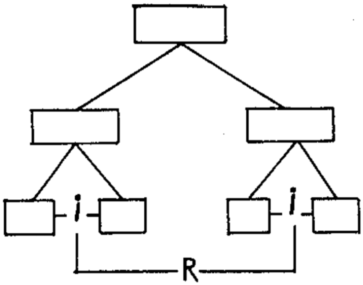
(Hjelmslev, 1975, Df. 24)

A Semiotic [...] is a Hierarchy, any of whose Components admits of a further Analysis into Classes defined by mutual Relation, so that any of these classes admits of an analysis into Derivates defined by mutual Mutation.

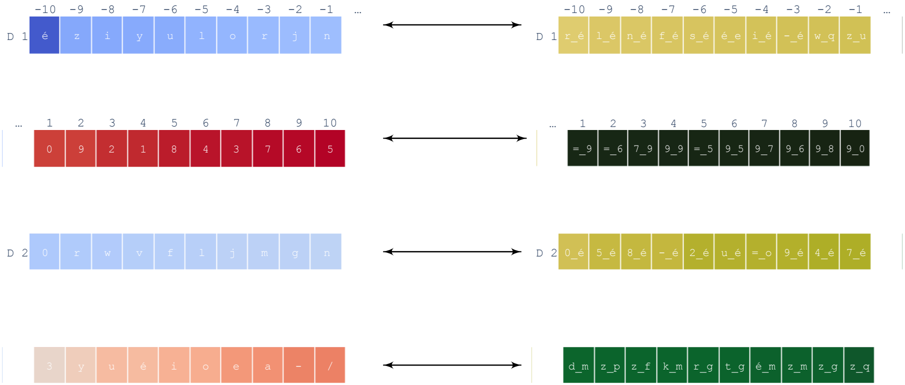
(Hjelmslev, 1975, Df. 24)

Mutation [...] is a Function existing between first-Degree Derivates of one and the same Class, *a function that has Relation to a function* between other first-degree derivates of one and the same class and belonging to the same Rank.

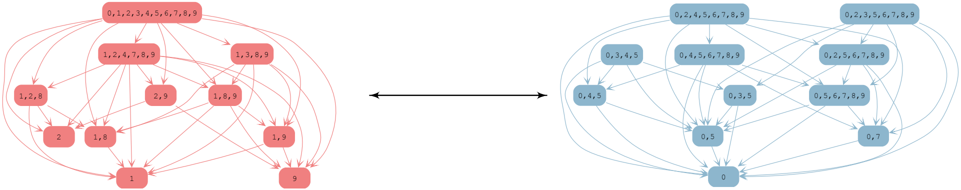
(Hjelmslev, 1975, Df. 23)

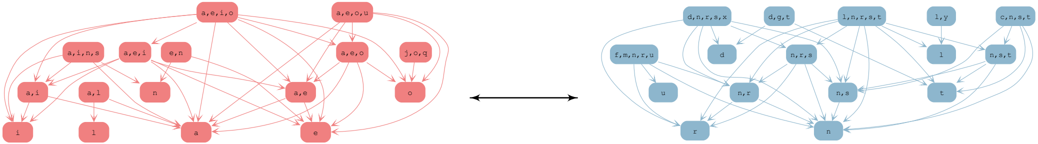


Syntagmatic and Text (Vectors)

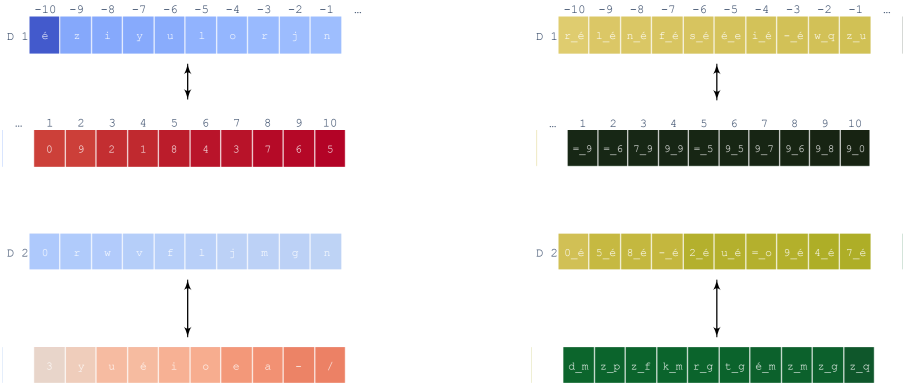


7.5 Syntagmatic and Text (Fixed Points/Types)

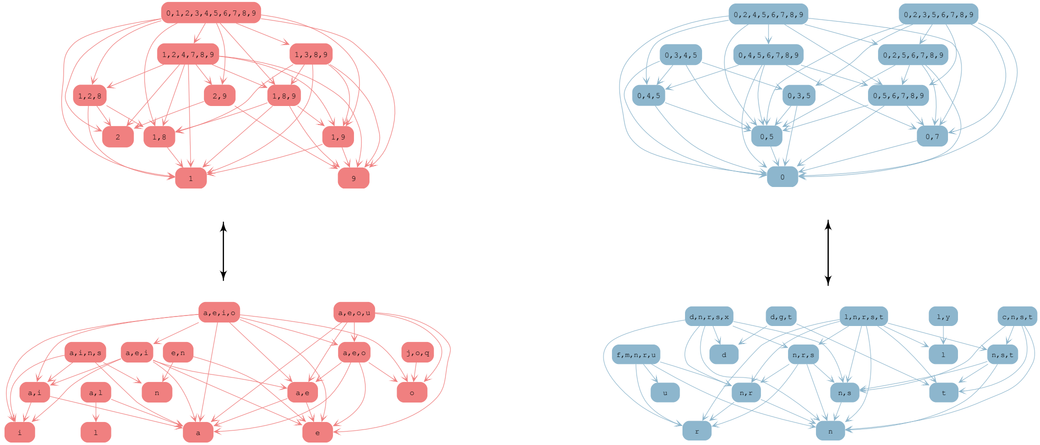




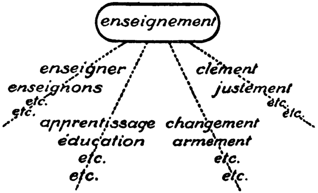
7.5 Paradigmatic and *Langue* (Vectors)



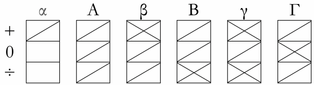
7.5 Paradigmatic and *Langue* (Nucleai/Types)



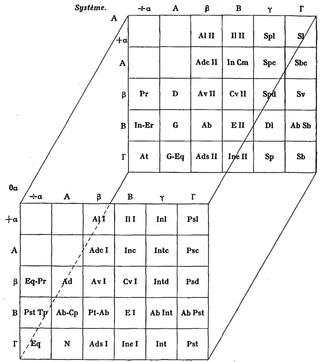
Structuralism and Formalism



(Saussure, 1980)



(Hjelmslev, 1975)



(Hjelmslev, 1935)

Segments	Environments											
	#— r̯	#—r	#—l	e i —C	æ—C	a o u —C	s— e i	s—æ	s— a o u	…	t̯ —	C³—
t̯	√											
t		√		√	√	√	√	√	√			
K						√			√			
k		√	√		√			√				
ᴋ				√			√					
G						√						
g		√	√		√							
ɢ				√								
r				√	√	√						√
r̯											√	

(Harris, 1960)

	a	b	d	e	f	g	h	i
a	aa	ab	ad		af	ag	ah	
b	ba							bi
d	da			de				di
e		eb	ed			eg		
f				fe				
g								gi
h	ha							hi
i			id				ih	i

Diagram 1.

	b	d	f	g	h	a	e	i
f						**fa**	fe	
h						ha		hi
g						**ga**	**ge**	gi
b						ba	**be**	bi
d						da	de	di
a	ab	ad	af	ag	ah	aa		
e	eb	ed	**ef**	eg				
i	**ib**	id		**ig**	ih			i

Diagram 3.

		I (p r s t)	II i	II o	II u	III y	IV &
I	p r s t	—	+	+	+	—	—
II	i	+	+			+	—
	o	+					
	u	+					
III	y	—	+			+	—
IV	&	—	—			—	+

Diagram 2.

(SpangHanssen1959)

Table 8.
Vowel × binary final cluster (cf. sect. 84).

	ft	gt	ks	ds	vn	vl	dr!	mp	nk	ng	nd	nt	ns	lk	ld	lt	rk	rd	rt	rn	S	T	jC	
a	5	10	6	3	9	8	6	8	16	20	14	9	6	9	8	11	7	1	9	3	168	281	3	a
e	–	–	3	1	3	2	2	1	–	4	7	5	6	–	3	5	–	1	3	3	49	95	33	e
i	7	6	9	5	–	1	2	4	13	11	20	8	3	2	11	6	6	1	1	–	116	171	–	i
o	3	2	2	5	4	2	1	1	1	2	3	2	–	4	13	3	6	9	10	4	77	120	–	o
u	2	9	5	4	–	–	6	12	8	4	12	3	2	4	8	4	4	–	2	–	89	143	–	u
y	–	2	–	2	–	–	1	2	4	7	6	2	–	1	6	6	3	2	1	–	45	56	–	y
æ	4	11	1	–	4	4	2	2	9	11	8	1	3	2	11	4	6	6	6	4	99	145	–	æ
ø	5	2	–	–	1	4	–	–	–	–	1	2	3	–	–	–	3	–	1	6	28	47	10	ø
aa	–	–	–	1	–	–	1	–	–	–	4	–	–	–	–	–	–	2	–	1	9	11	–	aa
	26	42	26	21	21	21	21	30	51	59	75	32	23	22	60	39	35	22	33	21	680	1069	46	

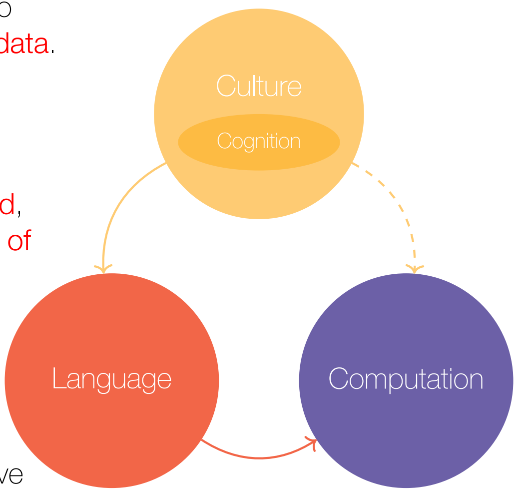
(SpangHanssen1959)

7 This New Structuralist Formalism Provides New Representational Tools for Explainability and Interpretability

7.1 Linear fixed points exhibit interpretable characteristics
7.2 Presheaf embeddings could replace vector embeddings
7.3 The profunctor's nucleus defines a system of logical types
7.4 The profunctor's nucleus could allow to study tokenization, embedding, and attention in a unified formal way
7.5 The resulting objects correspond to classical structuralist theoretical constructs

Language Models Are Culture Models

- 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.



8 Language Models Are Culture Models

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

8.2 Resulting models are, a priori, models of the data

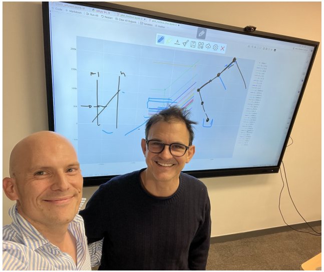
8.3 The cognitive content of such models is suspended, and cannot be restored without raising the problem of the data

8.4 The scale of the data for such models exceeds the individual scale

8.5 Cultural conditions of data production become constitutive in the relation between cognitive contents and language models

1. LLMs have no a priori cognitive import
2. The empirical study of LLMs has no epistemological grounds
3. Distributionalism is the best theoretical candidate to study LLMs
4. Distributionalism is a corollary of structuralism
5. The general form of distributions is $\mathcal{M} \colon \mathsf{C}^{\mathrm{op}} \times \mathsf{D} \to \mathcal{V}$
6. The general form of structures is $\mathcal{M}^* \colon \mathcal{V}^{\mathsf{C}} \leftrightarrows \mathcal{V}^{\mathsf{D}} \colon \mathcal{M}_*$
7. This new structuralist formalism provides new representational tools for explainability and interpretability
8. Language models are culture 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

Full Argument I

1	LLMs have no a priori cognitive import
1.1	The cognitive import of computational language models is not unconditional
1.11	The contemporary connection between computational LMs and cognition was set up by Chomsky
1.12	Yet, he denies any theoretical legitimacy to LLMs
1.13	The connection set up by Chomsky has very precise epistemological conditions
1.2	The epistemological condition ensuring such a connection does not hold for LLMs
1.3	The lack of cognitive import does not prevent LLMs to be models of language
1.31	The Chomskyan condition does not hold of necessity
1.32	Content can be an effect of form
1.33	The divorce between language and thought is not recent

Full Argument II

2	The empirical study of LLMs has no epistemological grounds
2.1	The NLP field has embraced an empirical turn
2.2	But LLMs are just computable functions
2.3	There is no empirical way of knowing what a computable function does
2.4	The only valid epistemological question is: What is this function the implementation of?
3	Distributionalism is the best theoretical candidate to study LLMs
3.1	All linguistic properties of an LLM come from distributions in data
3.2	Distributionalism is often associated to contexts
3.3	Contexts are often understood cognitively or pragmatically
3.4	The global character of distributional properties challenges cognitive and pragmatic interpretations
3.5	Distributionalism is not a thesis about cognition, but about the structure of language

Full Argument III

4	Distributionalism is a corollary of structuralism
4.1	The source of distributional properties is a virtual structure
4.2	The idea of virtually structured distributions is at the heart of classical structuralism
4.21	Saussure's notion of sign is intrinsically distributional
4.22	"Langue" as a virtual structure behind distribution is the very object of Saussurean linguistics
4.23	Analogical operations local operators of such virtual a structure
4.3	We need to move on from the distributional hypothesis to the structuralist hypothesis
5	The general form of distributions is $\mathcal{M}\colon \mathsf{C}^{\mathrm{op}} \times \mathsf{D} \to \mathcal{V}$
5.1	The (formal) key of neural LMs lies on embeddings
5.2	SVD over a PMI matrix provides the formal explanation for words embeddings

Full Argument IV

5.3 This result has important consequences for explainability

5.4 A matrix can be understood as a function $M \colon X \times Y \to \mathbb{R}$

5.5 We can generalize matrices to enriched profunctors $: \mathsf{C}^{\mathrm{op}} \times \mathsf{D} \to \mathcal{V}$

5.51 A category is like a set with structure

5.52 A functor is a map between categories

5.53 A profunctor is a functor from the product of two arbitrary categories to the Set category

5.54 A category enriched over $\mathcal{V}$ is a category having a $v_{\in \mathcal{V}}$'s worth arrows between two objects

5.55 A functor between the enriched categories $\mathsf{D} \to \mathsf{C}$ induces a profunctor is $\mathsf{C}^{\mathrm{op}} \times \mathsf{D} \to \mathcal{V}$

6 The general form of structures is $\mathcal{M}^* \colon \mathcal{V}^{\mathsf{C}} \leftrightarrows \mathcal{V}^{\mathsf{D}} \colon \mathcal{M}_*$

6.1 SVD looks for linear fixed points of the linear operators $M^* M_*$ and $M_* M^*$

6.2 The set of fixed points reveals (limited) structural features underlying the distributions

Full Argument V

6.3	The nucleus of an enriched profunctor provides a generalization of this setting
7	This new structuralist formalism provides new representational tools for explainability and interpretability
7.1	Linear fixed points exhibit interpretable characteristics
7.2	Presheaf embeddings could replace vector embeddings
7.3	The profunctor's nucleus defines a system of logical types
7.4	The profunctor's nucleus could allow to study tokenization, embedding, and attention in a unified formal way
7.5	The resulting objects correspond to classical structuralist theoretical constructs
8	Language models are culture models
8.1	A formal approach to data analysis can contribute to inferring symbolic language models from linguistic data
8.2	Resulting models are, a priori, models of the data

Full Argument VI

8.3 The cognitive content of such models is suspended, and cannot be restored without raising the problem of the data
8.4 The scale of the data for such models exceeds the individual scale
8.5 Cultural conditions of data production become constitutive in the relation between cognitive contents and language models

Linguistics and Language Models:
What Can They Learn from Each Other?
Leibniz Center for Informatics
Dagstuhl, Germany

Remarks on the Distributional Foundations of Language Models

Juan Luis Gastaldi

`www.giannigastaldi.com`

ETH zürich

July 22, 2025

Main Argument

1 LLMs have no a priori cognitive import
2 The empirical study of LLMs has no epistemological grounds
3 Distributionalism is the best theoretical candidate to study LLMs
4 Distributionalism is a corollary of structuralism
5 The general form of distributions is $\mathcal{M} \colon \mathsf{C}^{\mathrm{op}} \times \mathsf{D} \to \mathcal{V}$
6 The general form of structures is $\mathcal{M}^* \colon \mathcal{V}^{\mathsf{C}} \leftrightarrows \mathcal{V}^{\mathsf{D}} \colon \mathcal{M}_*$
7 This new structuralist formalism provides new representational tools for explainability and interpretability
8 Language models are culture models