

NeuroMod Annual Meeting
Université Côte d'Azur
Antibes, France

What are Neural Language Models the Model of?
Epistemological and Theoretical Perspectives on LLMs

Juan Luis Gastaldi
www.giannigastaldi.com

ETH zürich

July 8, 2025

Introduction

Epistemological Perspectives

Theoretical Perspectives

- The Algebra Behind the Embeddings

- The Structure Behind the Algebra

- The Categories Behind the Structure

Take Aways

Introduction

Epistemological Perspectives

Theoretical Perspectives

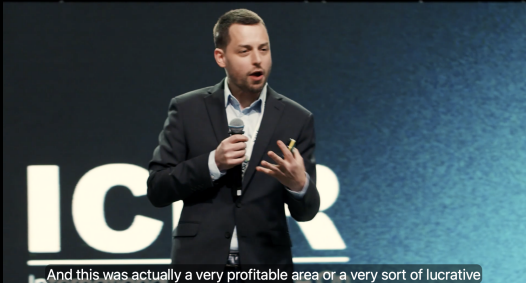
- The Algebra Behind the Embeddings

- The Structure Behind the Algebra

- The Categories Behind the Structure

Take Aways

Empirical Saturnalia



And this was actually a very profitable area or a very sort of lucrative

The empirics of deep learning

(Circa 2020) the scaling era is here; deep networks are now just emergent things we have created, that have to be studied scientifically like any other physical phenomenon

It seemed like **the best way for academic research to influence the field** is to develop the biology/physics (and let's be honest, more often pop psychology) of existing large models

24

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

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

And this was actually a very profitable area or a very sort of lucrative

The empirics of deep learning

(Circa 2020) the scaling era is here; deep networks are now just emergent things we have created, that have to be studied scientifically like any other physical phenomenon

It seemed like **the best way for academic research to influence the field** is to develop the biology/physics (and let's be honest, more often pop psychology) of existing large models

24

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

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

Cheng-Han Chiang

National Taiwan University,
Taiwan
dcml0714@gmail.com

Hung-yi Lee

National Taiwan University,
Taiwan
hungyilee@ntu.edu.tw

The empirics of deep learning

ca 2020) the scaling era is here; deep networks are now just emergent things we have created, that have to be studied scientifically like any other physical phenomenon

seemed like **the best way for academic research to influence the field** is to develop the biology/physics (and let's be honest, more often pop psychology) of existing large models

And this was actually a very profitable area or a very sort of lucrative

24

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

DO LLMs HAVE CONSCIOUSNESS?

Naama Rozen
Tel-Aviv University
naamarozen240@gmail.com

Gal Elidan
Google Research
Hebrew University
elidan@google.com

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

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

Yoichi Ishibashi
NEC
yoichi-ishibashi@nec.com

Taro Yano
NEC
taro_yano@nec.com

Masafumi Oyamada
NEC
oyamada@nec.com

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

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

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

deep learning

ca 2020) the scaling era is here, deep networks are now just emergent things
have created, that have to be studied scientifically like any other physical
phenomenon

seemed like **the best way for academic research to influence the field** is to
develop the biology/physics (and let's be honest, more often pop psychology) of
existing large models

And this was actually a very profitable area or a very sort of lucrative

24

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

Empirical Saturnalia

DO LLMs HAVE CONSCIOUSNESS?

Naama Rozen
Tel-Aviv University
naamarozen240@gmail.com

Gal Elidan
Google Research
Hebrew University
elidan@google.com

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

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

Yoichi Ishibashi
NEC
yoichi-ishibashi@nec.com

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

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

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

DO LLMs “KNOW” INTERNALLY WHEN THEY FOLLOW INSTRUCTIONS?

Juyeon Heo^{1,*} Christina Heinze-Deml² Oussama Elachqar² Kwan Ho Ryan Chan^{3,*} Shirley Ren²
Udhay Nallasamy² Andy Miller² Jaya Narain²

¹University of Cambridge ²Apple ³University of Pennsylvania
jh2324@cam.ac.uk jnarain@apple.com

permeated like **the best way for academic research to influence the field** is to develop the biology/physics (and let's be honest, more often pop psychology) of existing large models

24

And this was actually a very profitable area or a very sort of lucrative

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

Empirical Saturnalia

DO LLMs HAVE CONSCIOUSNESS?

Naama Rozen
Tel-Aviv University
naamarozen240@gmail.com

Gal Elidan
Google Research
Hebrew University
elidan@google.com

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

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

Yoichi Ishibashi
NEC
yoichi-ishibashi@nec.com

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

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

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

DO LLMs “KNOW” INTERNALLY WHEN THEY FOLLOW INSTRUCTIONS?

Juyeon Han^{1,*}, Chaitin Hui², Dongyue Hu², Kaituo Hu², Yuxin Hu², Jiahao Hu², Jiahao Hu², Jiahao Hu², Jiahao Hu², Jiahao Hu²
¹University of Illinois at Chicago, ²University of Illinois at Chicago
jh2324@uic.edu

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

Siyan Zhao^{2*}, Mingyi Hong^{1,3}, Yang Liu¹, Devamanyu Hazarika¹, Kaixiang Lin¹ †
¹Amazon AGI, ²UCLA, ³University of Minnesota
siyanz@cs.ucla.edu, mhong@umn.edu, devamanyu@u.nus.edu
{yangliud, kaixianl}@amazon.com

And this was actually a very profitable area or a very sort of lucrative

24

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

Empirical Saturnalia

DO LLMs HAVE CONSCIOUSNESS?

Naama Rozen
Tel-Aviv University
naamarozen240@gmail.com

Gal Elidan
Google Research
Hebrew University
elidan@google.com

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

Yoichi Ishibashi
NEC
yoichi-ishibashi@nec.com

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

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

DO LLMs “KNOW” INTERNALLY WHEN THEY FOLLOW INSTRUCTIONS?

Juyeon Hwang^{1*}, Chaitin Hui², Dongyue Hu², Elad Nitzan², Kevin H. Poon^{3*}, Shih-Wei
Udhay Narayanan¹
¹University of California, San Diego
jh2324@ucsd.edu

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

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

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

n¹ †

And this was actually a very profitable area or a very sort of lucrative

Zico Kolter, *Building Safe and*

DO LLMs HAVE CONSCIOUSNESS?

Naama Rozen
Tel-Aviv University
naamarozen240@gmail.com

Gal Elidan
Google Research
Hebrew University
elidan@google.com

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

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

Yoichi Ishibashi
NEC
yoichi-ishibashi@nec.com

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

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

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

DO LLMs “KNOW” INTERNALLY WHEN THEY FOLLOW INSTRUCTIONS?

Juyeon Hwang^{1*}, Chaitin Hui¹, Dongyue Huo², Qianqian Huo², Kaituo Huo², Hui Ren^{3*}, Shihong Ren³,
Udhay Narayanan¹
¹University of Michigan, ²Google, ³Google Research

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

Language Models are Few-Shot Learners

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

Arash Gholami Davoodi¹, Seyed Pouyan Mousavi Davoudi, Pouya Pezeshkpour²
¹Carnegie Mellon University, ²Megagon Labs
agholami@andrew.cmu.edu, spouyan.mousavi@gmail.com, pouya@megagon.ai

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

Can Large

Can Large

Can Large

Can Large

Can Large

Can Large

Can Large

Can Large

Can Large

Can Large

Can Large

Can Large

Can Large

Can Large

Can Large

Can Large

Can Large

Can Large

Can Large

Can Large

Can Large

Can Large

Can Large

Can Large

Can Large

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

DO LLMs HAVE CONSCIOUSNESS?

Naama Rozen
Tel Aviv University

Yoichi Ishibashi
NEC

Do LLMs “KNOW” INTERNALLY WHEN THEY FOLLOW INSTRUCTIONS?

Large Language Models are Zero-Shot Reasoners

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

Shixiang Shane Gu
Google Research, Brain Team

Machel Reid
Google Research*

Yutaka Matsuo
The University of Tokyo

Yusuke Iwasawa
The University of Tokyo

Ryo Kamoi¹ Yusen Zhang¹ Nan Zhang¹ Jiawei Han² Rui Zhang¹
¹Penn State University, USA ²University of Illinois Urbana-Champaign, USA
{ryokamoi, rmz5227}@psu.edu

Chaojie Huang¹, Danyang Qu², Geyan Ebrahimi², Kaituo Hui¹, Peng Chen^{3*}, Shihang Ren³

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

Language Models are Few-Shot Learners

Scaling Mathematical Topic Tree Evaluation of LLMs

Mehdi Davoudi, Pouya Pezeshkpour²
Megagon Labs

agholami@andrew.cmu.edu, spouyan.mousavi@gmail.com, pouya@megagon.ai

Mark Ryder*

Melanie Subbiah*

Jon

Pranav Shyam

Girish Sastry

Gretchen Krueger

Tom Henighan

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

DO LLMs HAVE CONSCIOUSNESS?

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

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

Microsoft Research

HOW” INTERNALLY WHEN THEY FOLLOW
?

LLMs RECOGNIZE YOUR PREFERENCES? EVALUATING
PERSONALIZED PREFERENCE FOLLOWING IN LLMs

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

Ryo Kamoi¹ Yusen Zhang¹ Nan Zhang¹ Jiawei Han² Rui Zhang¹
¹Penn State University, USA ²University of Illinois Urbana-Champaign, USA
{ryokamoi, rmz5227}@psu.edu

Language Models are Few-Shot Learners

Long Mathematical Topic Tree
Evaluation of LLMs

Davoudi, Pouya Pezeshkpour²
Megagon Labs

agholami@andrew.cmu.edu, spouyan.mousavi@gmail.com, pouya@megagon.ai

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

Can Large Language Models Invent Algorithms to Improve Themselves?: Algorithms Discovered for Reasoning Self-Improvement through

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

Hamid Palangi Marco Tulio Ribeiro Yi Zhang

Microsoft Research

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

Ryo Kamoi¹

Yusen Zhang¹

Nan Zhang¹

Jiawei Han²

Rui Zhang¹

¹Penn State University, USA ²University of Illinois Urbana-Champaign, USA

{ryokamoi, rmz5227}@psu.edu

Language Models are Few-Shot Learners

Long Mathematical Topic Tree Evaluation of LLMs

Davoudi, Pouya Pezeshkpour²
Megagon Labs

agholami@andrew.cmu.edu, spouyan.mousavi@gmail.com, pouya@megagon.ai

WHEN THEY FOLLOW

YOUR PREFERENCES? EVALUATING
PERSONALIZED PREFERENCE FOLLOWING IN
LLMs

n¹†

Ryder*

Melanie Subbiah*

on

Pranav Shyam

Girish Sastry

Gretchen Krueger

Tom Henighan

Can Large Language Models Invent Algorithms to Improve Themselves?: Algorithms Discovered for Reasoning Self-Improvement through

PROCEDURAL KNOWLEDGE IN PRETRAINING DRIVES REASONING IN

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,
National Engineering Research Center of Visual Information and Applications,
Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University

[♣]Microsoft [♡]Peking University

Laura Ruis^{*}
AI Centre, UCL

Sébastien Bubeck
Eric Horvitz Ece Kar
Harsha Nori

Hamid Palangi
Microsoft Rese

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

Ryo Kamoi¹

Yusen Zhang¹

Nan Zhang¹

Jiawei Han²

Rui Zhang¹

¹Penn State University, USA ²University of Illinois Urbana-Champaign, USA

{ryokamoi, rmz5227}@psu.edu

Language Models are Few-Shot Learners

ing Mathematical Topic Tree valuation of LLMs

i Davoudi, Pouya Pezeshkpour²
Megagon Labs

agholami@andrew.cmu.edu, spouyan.mousavi@gmail.com, pouya@megagon.ai

k Ryder*

Melanie Subbiah*

on

Pranav Shyam

Girish Sastry

Gretchen Krueger

Tom Henighan

Empirical Saturnalia

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

PROCEDURAL KNOWLEDGE IN PRETRAINING DRIVES REASONING IN

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,
National Engineering Research Center for Software Engineering

Can Large Language Models Reason About Goal-Oriented Tasks?

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

Long Mathematical Topic Tree Evaluation of LLMs

Davoudi, Pouya Pezeshkpour²
Megagon Labs

agholami@andrew.cmu.edu, spouyan.mousavi@gmail.com, pouya@megagon.ai

Laura Ruis^{*}
AI Centre, UCL

Sébastien Bubeck
Eric Horvitz Ece Kar
Harsha Nori

Hamid Palangi

Microsoft Rese

Takeshi Kojima

The University of Tokyo

t.kojima@weblab.t.u-tokyo.ac.jp

Shixiang Sha

Google Research,

Machel Reid

Google Research*

Yutaka Matsuo

The University of Tokyo

Yusuke Iwasawa

The University of Tokyo

Ryo Kamoi¹

Yusen Zhang¹

Nan Zhang¹

Jiawei Han²

Rui Zhang¹

¹Penn State University, USA

²University of Illinois Urbana-Champaign, USA

{ryokamoi,rmz5227}@psu.edu

K Ryder*

Melanie Subbiah*

on

Pranav Shyam

Garish Sastry

Gretchen Krueger

Tom Henighan

Empirical Saturnalia

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

PROCEDURAL KNOWLEDGE IN PRETRAINING DRIVES REASONING IN

Laura Ruis*
AI Centre, UCL

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

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

[◇]National Key Laboratory of Human-Machine Hybrid Augmented Intelligence,
National Engineering Research Center for Software Engineering, Tsinghua University

Can Large Language Models Reason About Goal-Oriented Tasks?

AutoClean: LLMs Can Prepare Their Training Corpus

Xingyu Shen¹ Shengding Hu¹ Xinrong Zhang¹ Xu Han^{1*}
Xiaojun Meng² Jiansheng Wei² Zhiyuan Liu^{1*} Maosong Sun¹

¹Department of Computer Science and Technology, Institute for Artificial Intelligence,
Beijing National Research Center for Information Science and Technology,
Tsinghua University, Beijing, China.

²Huawei Noah's Ark Lab

David Davoudi, Pouya Pezeshkpour²
Megagon Labs

agholami@andrew.cmu.edu, spouyan.mousavi@gmail.com, pouya@megagon.ai

DO LLMs HAVE COMMON SENSE?

Sparks

Sébastien Bubeck
Eric Horvitz Ece Kar
Harsha Nori

Takeshi Kojima

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

Machel Reid
Google Research*

Yutaka Matsuo
The University of Tokyo

Ryo Kamoi¹ Yusen Zhang¹ Nan Zhang¹ Jiawei Li²

¹Penn State University, USA ²University of Illinois Urbana-Champaign

{ryokamoi, rmz5227}@psu.edu

Empirical Saturnalia

Can Large Language Models Invent Algorithms to Improve Themselves?:

Algorithm: Diagnosis for Decreasing Self-Improvement through

DO LLMs HAVE CON

Sparks

PROCEDURAL KNOWLEDGE IN PRETRAINING DRIVES

WHEN THEY FOLLOW

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 Ece Kar
Harsha Nori

Hamid Palangi Ma

Microsoft Rese

Can Large Language Models Reason About Goal-Oriented Tasks?

? EVAL- WING IN

Takeshi Kojima

The University of Tokyo

t.kojima@weblab.t.u-tokyo.ac.jp

Machel Reid
Google Research*

Ryo Kamoi¹ Yusen Zhar
¹Penn State University, USA
 {ryokan, yzhar}@psu.edu

Google

Filippos Bellos Yanyan Li Wuzao Liu Jason I. Corso

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
diyang.tony@sjtu.edu.cn, zhaohai@cs.sjtu.edu.cn

Han^{1*}
Song Sun¹
Artificial Intelligence,
Technology,

Gretchen Krueger Tom Henighan

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

DO LLMs HAVE COGNITIVE

PROCEDURAL KNOWLEDGE IN PRETRAINING DRIVES

WHEN THEY FOLLOW

Sparks

REASONING IN

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 Ece Kar

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
thatadamorris@gmail.com

Keerthi Reddy
Independent Researcher

Jorge Morales
Northeastern University

Can Models Reason About Goal-Oriented Tasks?

Yayuan Li Wuzao Liu Jason L. Corso

Can Large Language Models Reason About Their Training Corpus?

Han^{1*}
Song Sun¹
Artificial Intelligence,
Technology,

Tao Tong University,
Agent Interaction and
University,
Governance in Web3
edu.cn

Gretchen Krueger Tom Henighan

Empirical Saturnalia

Can Large Language Models Invent Algorithms to Improve Themselves?:

Algorithm: Diagnosis for Determining Self-Immersion through

PROCEDURAL KNOWLEDGE IN PRETRAINING DRIVES

REASONING IN

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[♣]

Laura Ruis*

ALCentre UCL

◆National Key Laboratory of Human-Machine Hybrid Augmented Intelligence.

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

Dillon Plunkett

Northeastern University

d.plunkett@northeastern.edu

Adam Morris

Princeton University

thatadamorris@gmail.com

Keerthi Reddy

Independent Researcher

Jorge Morales

Northeastern University

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

Corpus

Han^{1*}song Sun¹

ificial Intelligence,
Technology,

Gretchen Krueger

Tom Henighan

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

DO LLMs HAVE

Naam
Tel. A

Sébastien B
Eric Horvitz

La

Self-Int
Internat

d.plu



like the pop psychology in practice,

Keerthi Reddy
Independent Researcher

Jorge Morales
Northeastern University

edu.cn

Gretchen Krueger

Tom Henighan

IN THEY FOLLOW
oning Tasks

gence,

Oriented Tasks?

L Corso

n¹
elligence,
ogy,

iah*

h Sastry

EVAL-
WING IN

Introduction

Epistemological Perspectives

Theoretical Perspectives

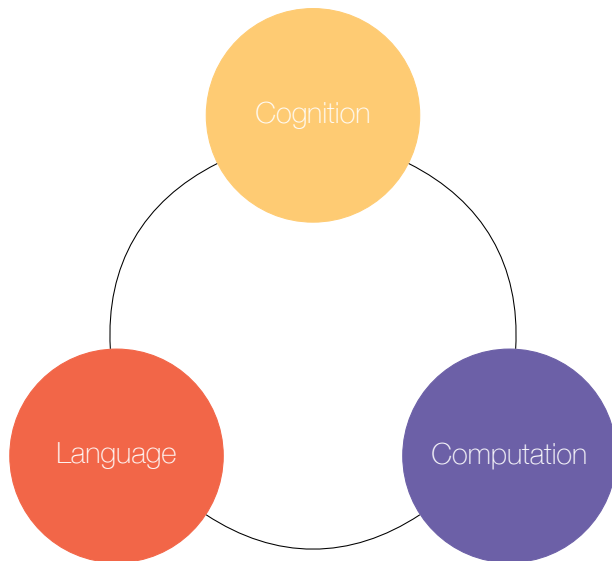
- The Algebra Behind the Embeddings

- The Structure Behind the Algebra

- The Categories Behind the Structure

Take Aways

Chomsky's Generativist Program and the Cognitive Revolution



The New York Times

OPINION
GUEST ESSAY

Noam Chomsky: The False Promise of ChatGPT

March 8, 2023



Chomsky against Abstraction in Principle

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

(Chomsky, 1992)



Chomsky against Abstraction in Principle

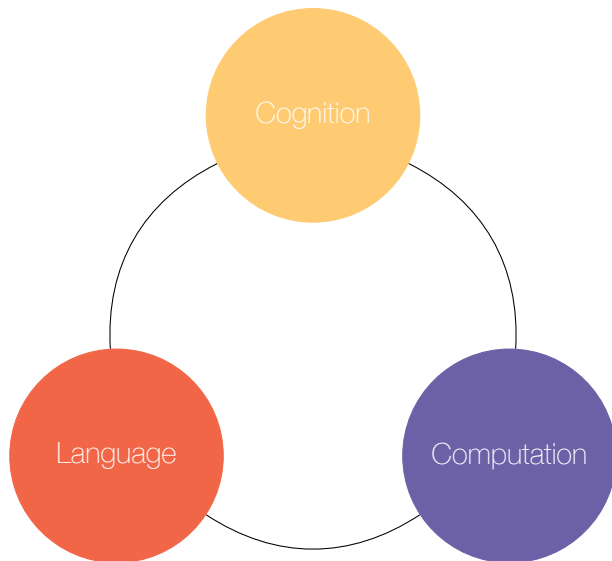
“There’s nothing wrong with principled abstraction. In fact, one might think of large areas of mathematics as that. But here we have something new, principled abstraction in an empirical discipline.”

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

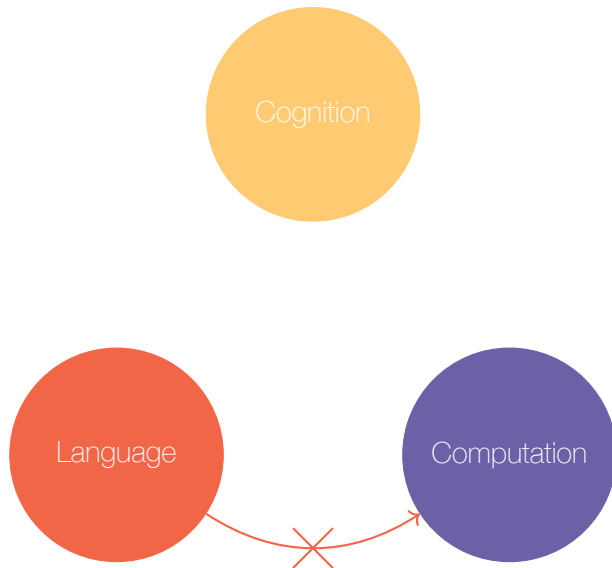
(Chomsky, 1992)



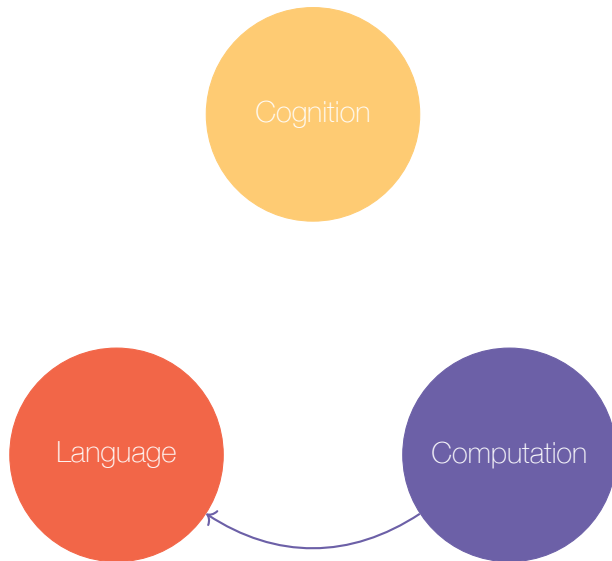
The Condition of Chomsky's Cognitive Foundations



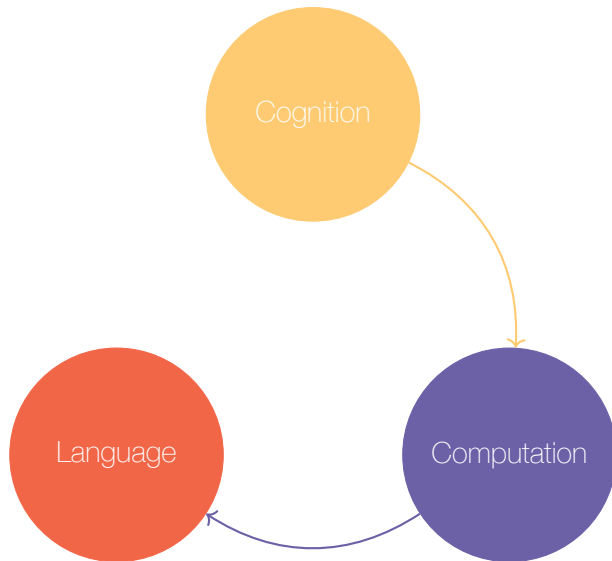
The Condition of Chomsky's Cognitive Foundations



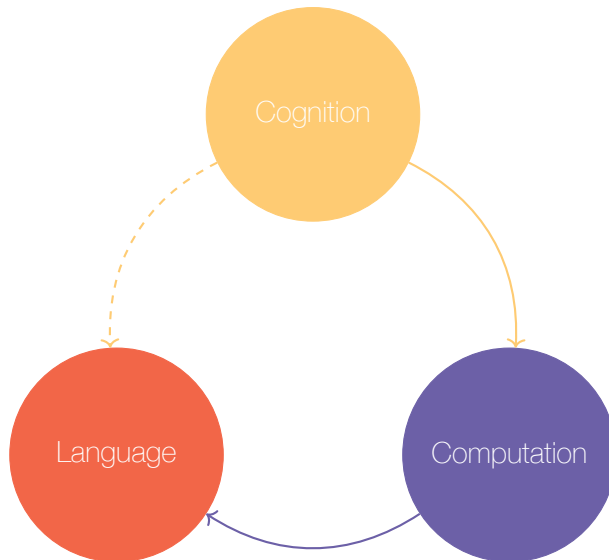
The Condition of Chomsky's Cognitive Foundations



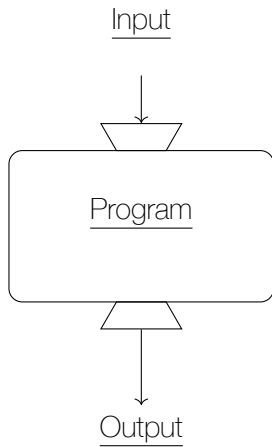
The Condition of Chomsky's Cognitive Foundations



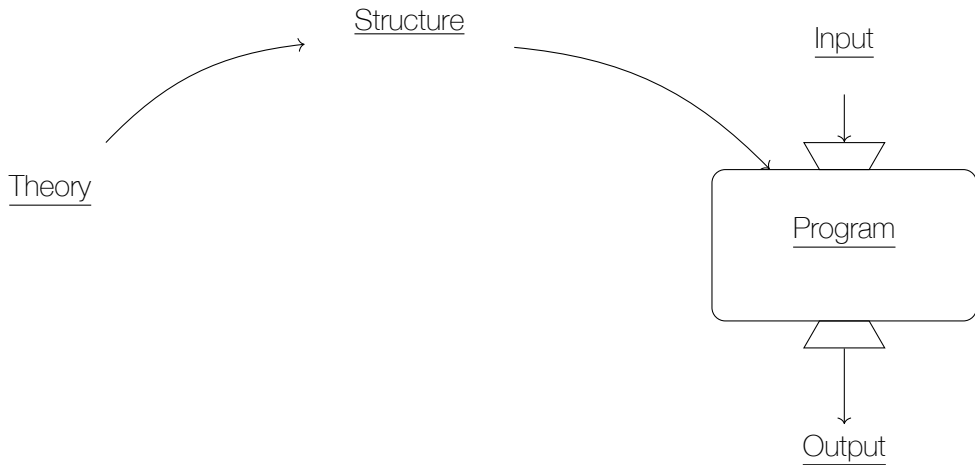
The Condition of Chomsky's Cognitive Foundations



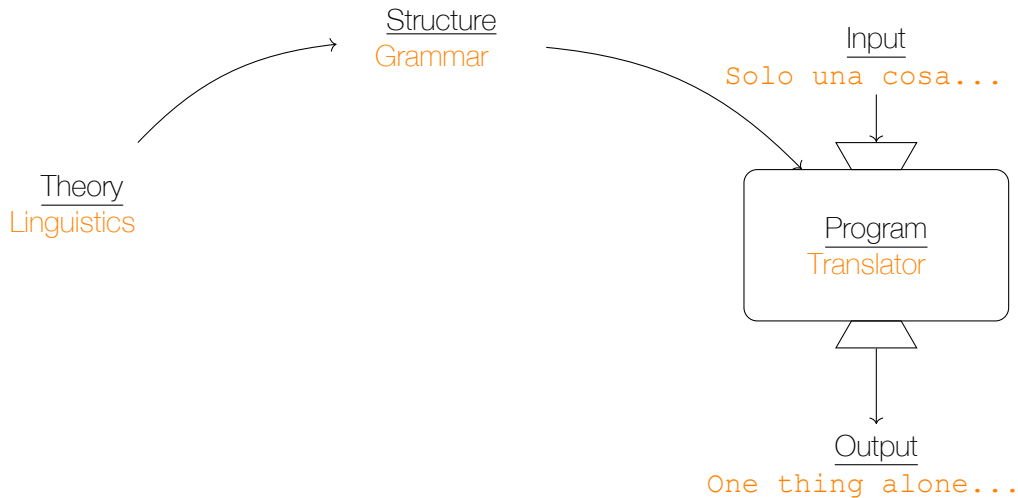
The Trap



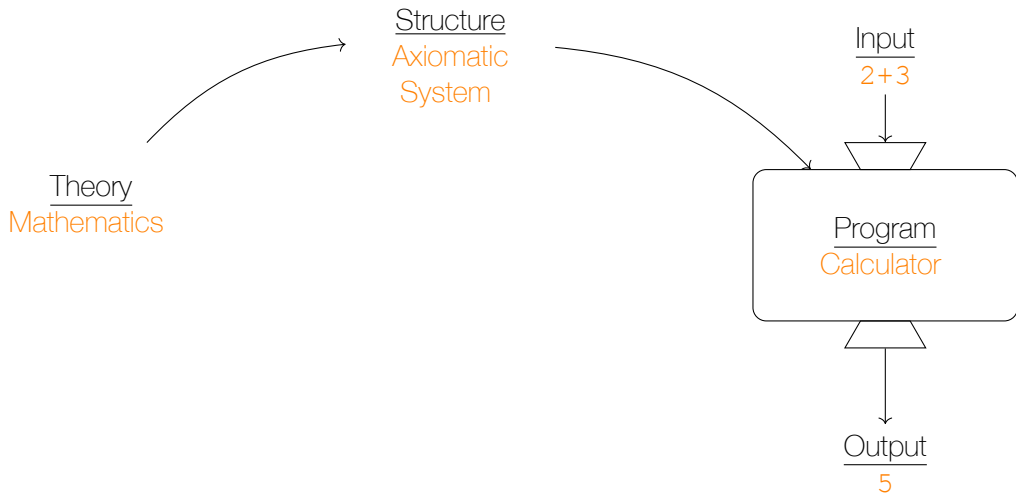
The Trap



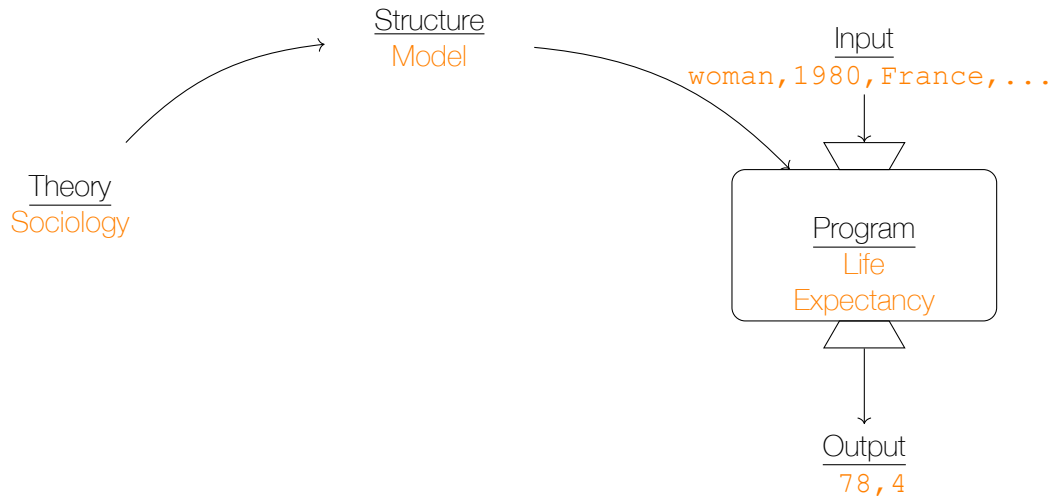
The Trap



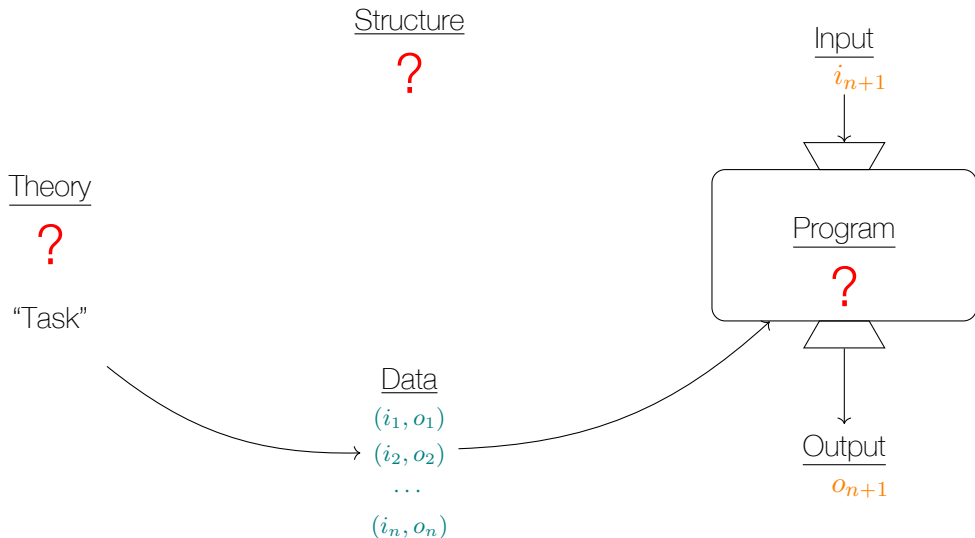
The Trap



The Trap

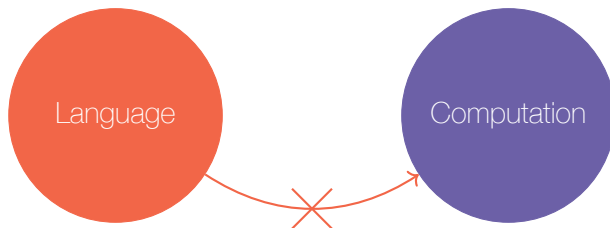


The Trap



Necessary Condition?

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



Necessary Condition?

- ◇ Inadequacy of distributional models (Chomsky, 1953)

Inconclusive

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

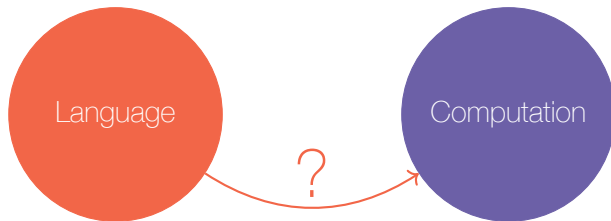
Empirically challenged

- ◇ Limited expressive power of FSAs (Chomsky, 1956)

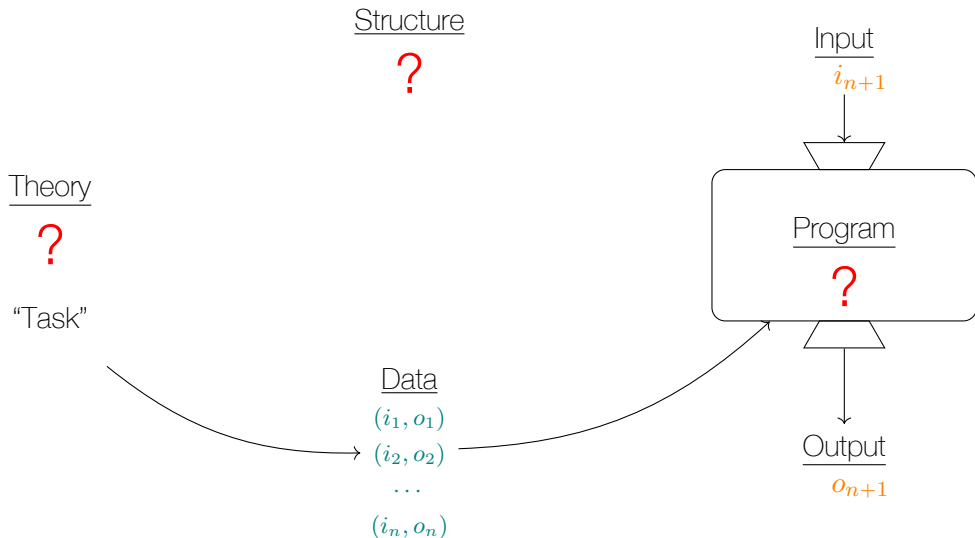
The relevance is unclear

- ◇ Poverty of stimulus (Chomsky, 1959)

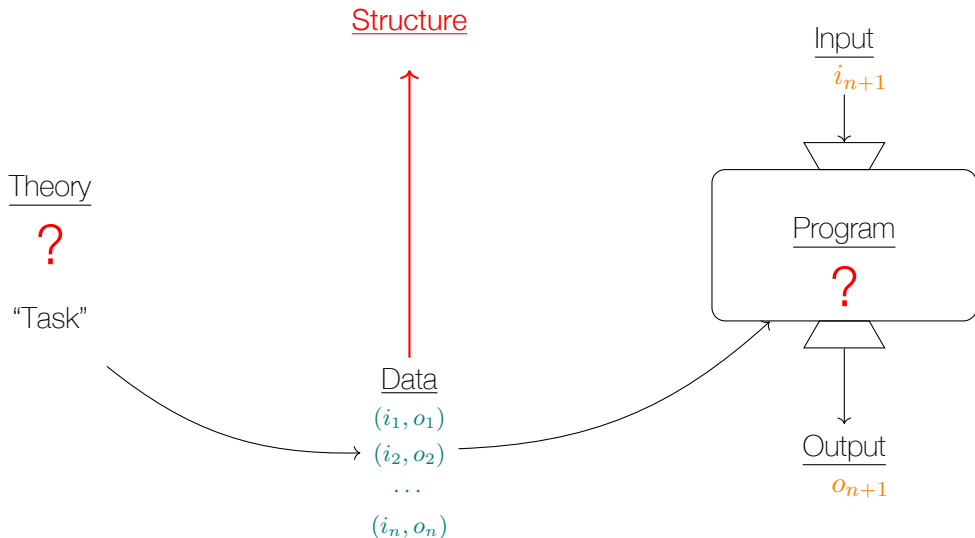
Assumes what is to be proved



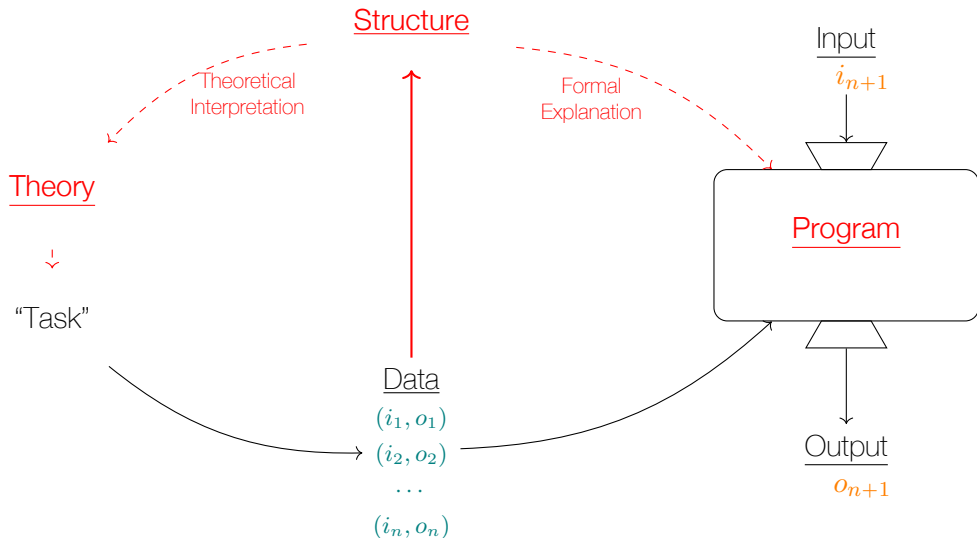
Making It Explicit



Making It Explicit



Making It Explicit



Introduction

Epistemological Perspectives

Theoretical Perspectives

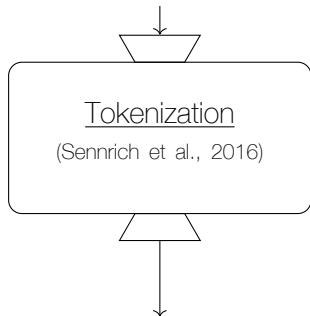
- The Algebra Behind the Embeddings

- The Structure Behind the Algebra

- The Categories Behind the Structure

Take Aways

Epistemology of Machine Learning
Distributional Language Models

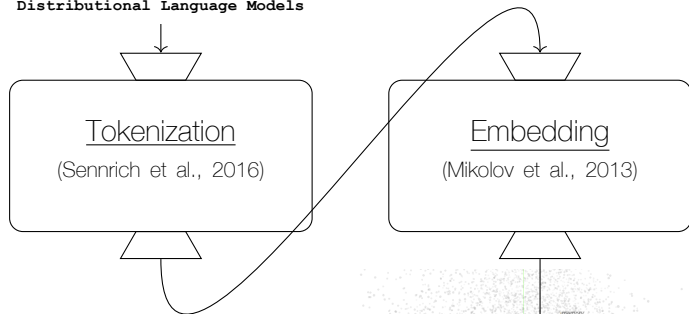


Epistemology of Machine Learning
Distributional Language Models

(<https://tiktokenizer.vercel.app>)

Formal Explainability

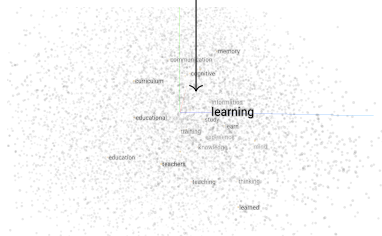
Epistemology of Machine Learning Distributional Language Models



Epistemology of Machine Learning

Distributional Language Models

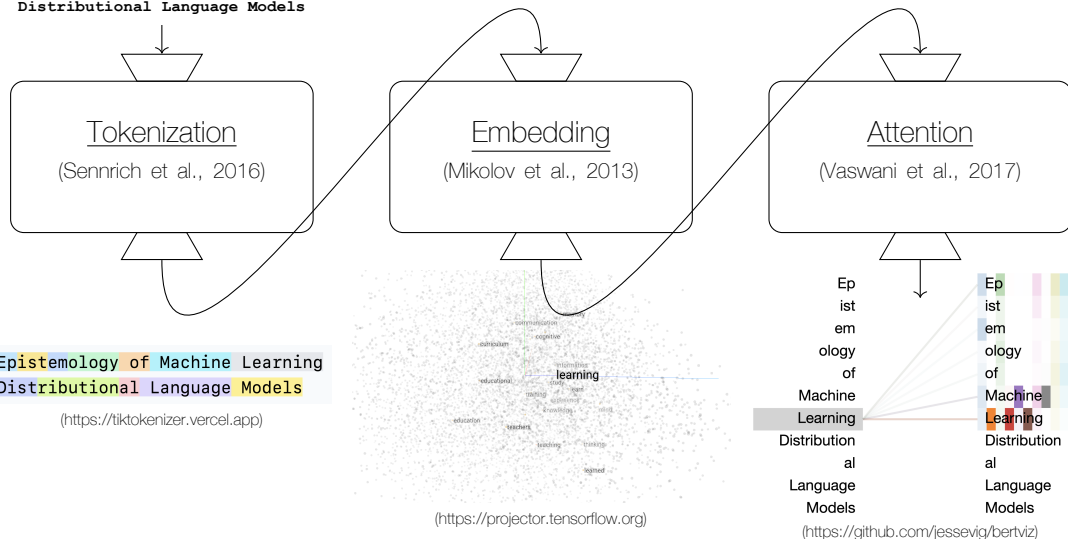
(<https://tiktokenizer.vercel.app>)

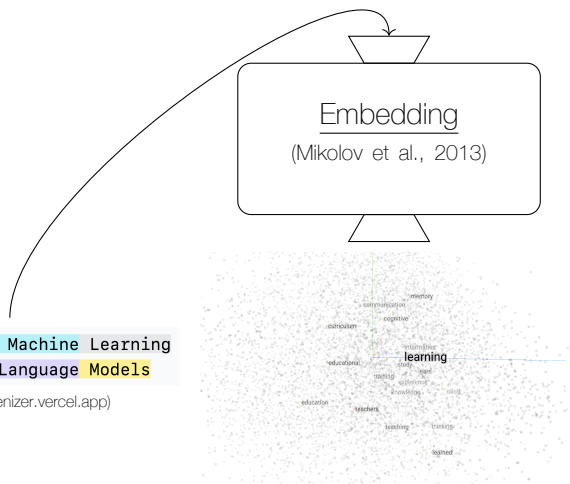


(<https://projector.tensorflow.org>)

Formal Explainability

Epistemology of Machine Learning
Distributional Language Models





The diagram illustrates the process of formal explainability in machine learning. It starts with a sentence, "Epistemology of Machine Learning Distributional Language Models", which is tokenized. An arrow points from this tokenized sentence to a box labeled "Embedding (Mikolov et al., 2013)". From the embedding box, another arrow points to a word cloud. The word cloud contains various words related to the original sentence, with "learning" being the most prominent word in the center.

Embedding
(Mikolov et al., 2013)

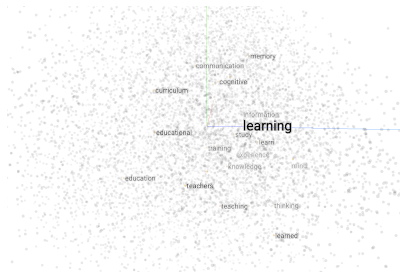
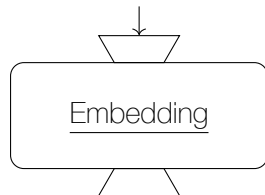
Epistemology of Machine Learning
Distributional Language Models

(<https://tiktokenizer.vercel.app>)

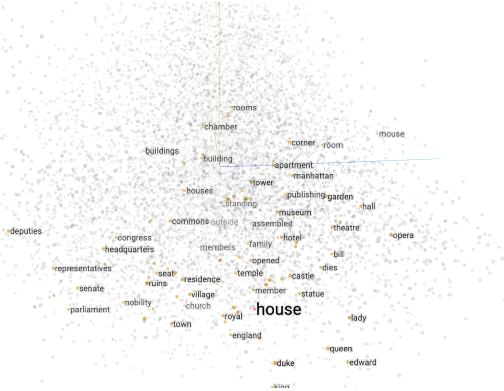
(<https://projector.tensorflow.org>)

The Structure of Embeddings

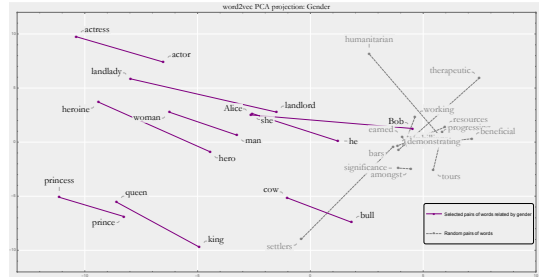
Epistemology of Machine Learning
Distributional Language Models



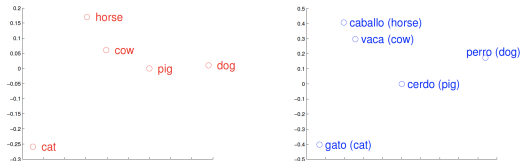
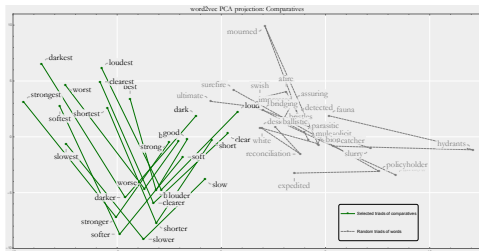
Embeddings: Similarity and Analogy



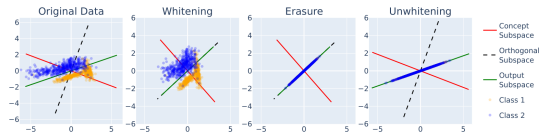
(<https://projector.tensorflow.org>)



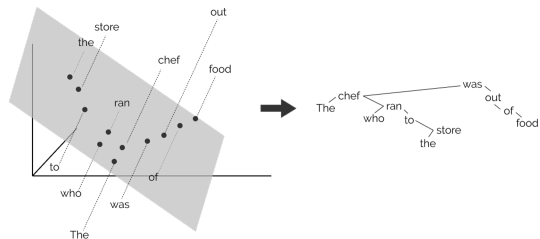
Embeddings: Other Applications



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



(Belrose et al., 2024)



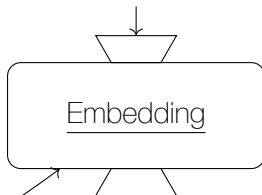
(<https://nlp.stanford.edu/~johnhew/structural-probe.html>)

The Structure of Embeddings

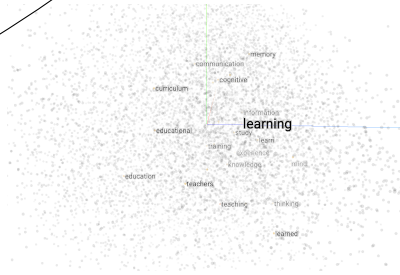
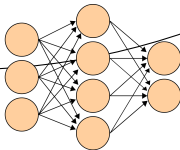
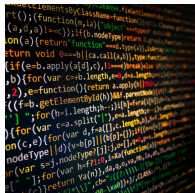
Structure

?

Epistemology of Machine Learning
Distributional Language Models



Data

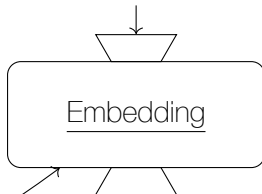


The Structure of Embeddings

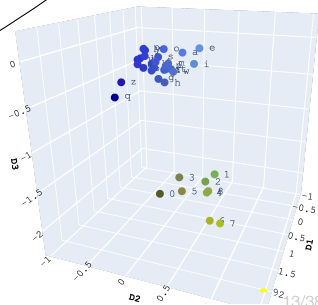
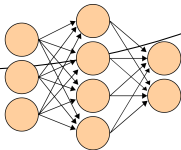
Structure

?

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



Data



Introduction

Epistemological Perspectives

Theoretical Perspectives

- The Algebra Behind the Embeddings

- The Structure Behind the Algebra

- The Categories Behind the Structure

Take Aways

word2vec Explained

(Levy and Goldberg, 2014)

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

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

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

word2vec Explained

(Levy and Goldberg, 2014)

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

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

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

word2vec Explained

(Levy and Goldberg, 2014)

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

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

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

word2vec Explained

(Levy and Goldberg, 2014)

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

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

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

word2vec Explained (Levy and Goldberg, 2014)

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

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

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

word2vec Explained

(Levy and Goldberg, 2014)

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

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

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

word2vec Explained

(Levy and Goldberg, 2014)

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

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

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

word2vec Explained (Levy and Goldberg, 2014)

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

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

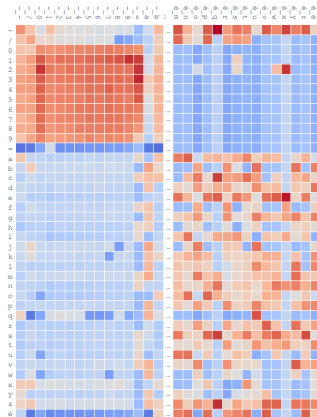
- ◊ Word2vec performs an **implicit, low-dimensional factorization** of a **pointwise mutual information (pmi), word-context matrix**.
- ◊ The **Singular Value Decomposition (SVD)** provides an **exact solution** to this problem.

Example: Characters in Wikipedia

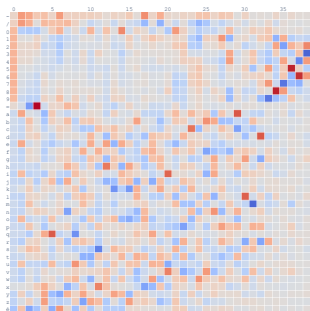
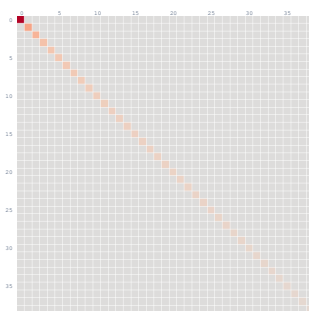
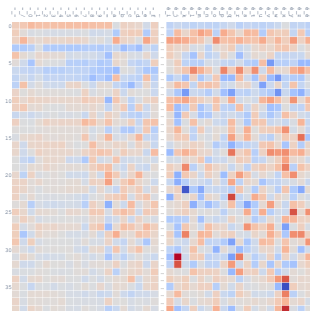
$W = \{-, /, 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, =, a, b, c, \dots, w, x, y, z, \acute{e}\}$

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

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

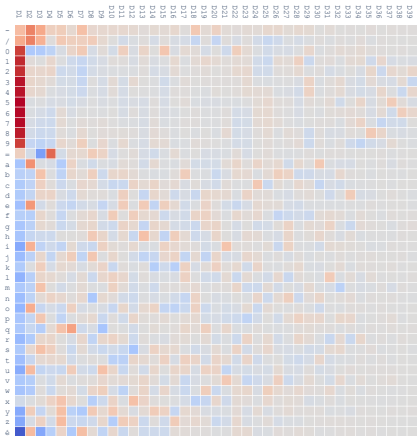


SVD of Wikipedia Character PMI Matrix

 U  Σ  V^T 

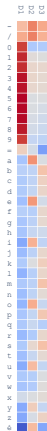
Truncate

$$U \times \Sigma$$

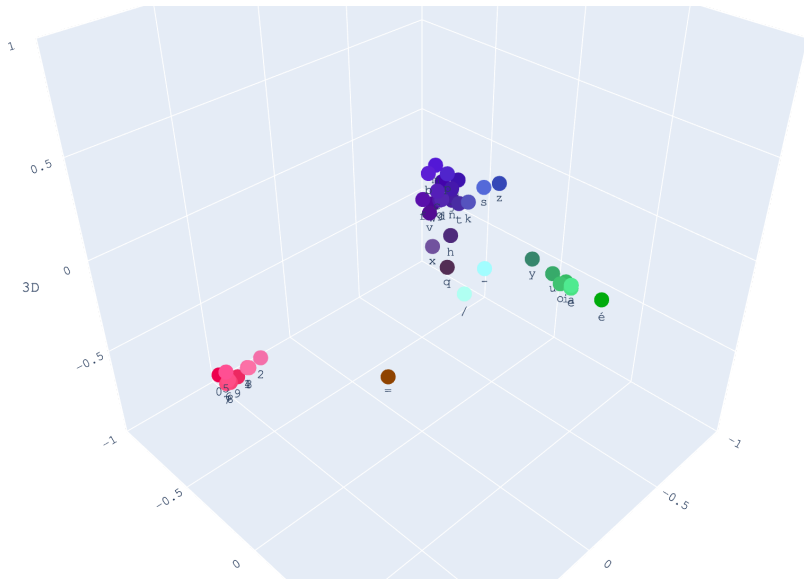
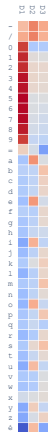


Truncate

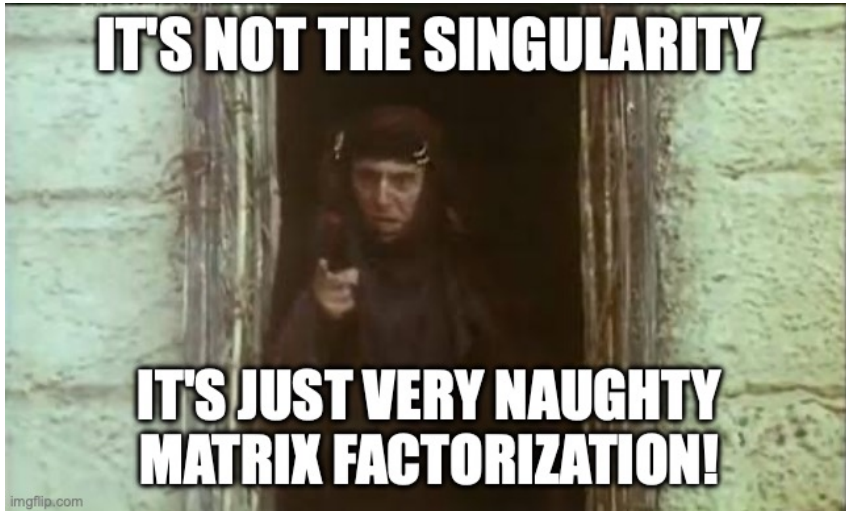
$$\hat{U} \times \hat{\Sigma}$$



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



What to conclude?



The Structure of Embeddings

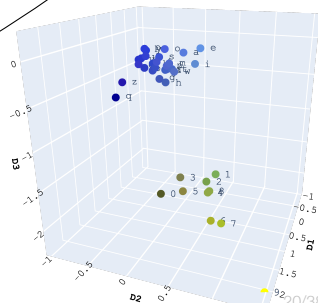
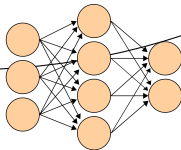
Structure

?

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

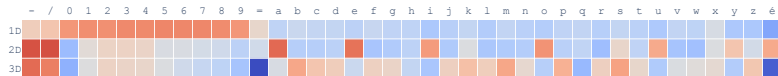
Embedding

Data

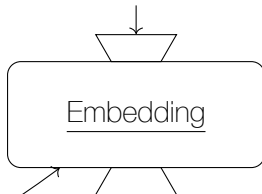


The Structure of Embeddings

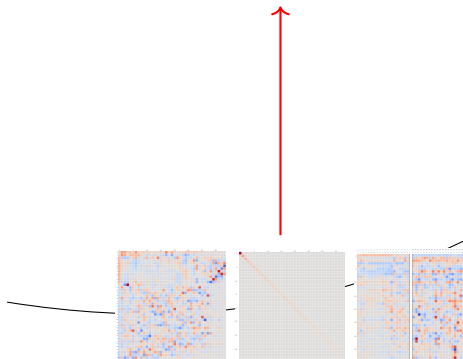
Structure



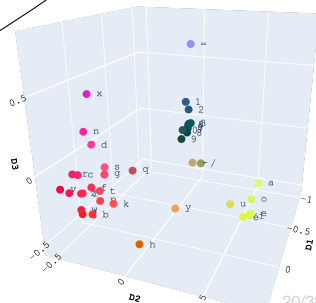
$\{-, /, 0, 1, 2, \dots, 8, 9, =, a, b, c, \dots, w, x, y, z, \acute{e}\}$



Data



SVD



4 Why does this produce good word representations?

Good question. We don't really know.

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

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

(Goldberg and Levy, 2014)

Introduction

Epistemological Perspectives

Theoretical Perspectives

The Algebra Behind the Embeddings

The Structure Behind the Algebra

The Categories Behind the Structure

Take Aways

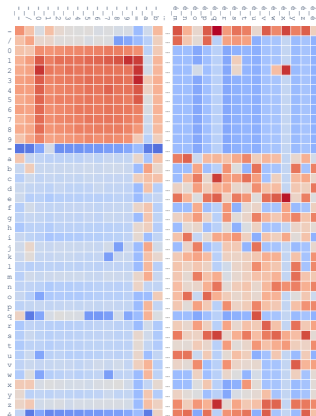
Embeddings as Functions Over Sets

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

$$\textcolor{blue}{Y} = X \times X = \{(-, -), (-, /), (-, 0), \dots, (\acute{e}, z), (\acute{e}, \acute{e})\}$$

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

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



Embeddings as Functions Over Sets

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

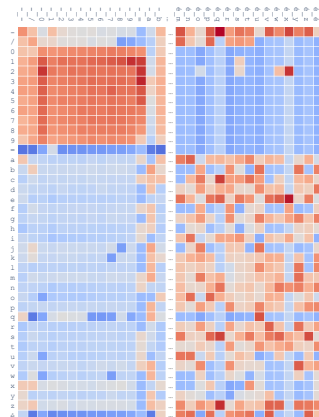
$Y = X \times X = \{(-, -), (-, /), (-, 0), \dots, (\acute{e}, z), (\acute{e}, \acute{e})\}$

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

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

$$M_x: X \rightarrow \mathbb{R}^Y$$

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



Embeddings as Functions Over Sets

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

$$\textcolor{blue}{Y} = X \times X = \{(-, -), (-, /), (-, 0), \dots, (\acute{e}, z), (\acute{e}, \acute{e})\}$$

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

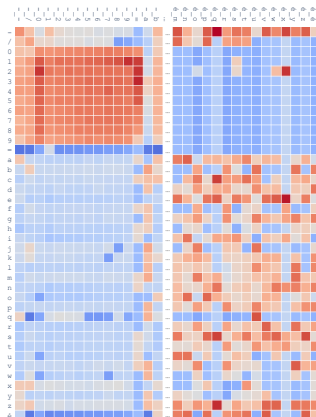
$$(\textcolor{red}{x}, \textcolor{blue}{y}) \mapsto \text{pmi}(\textcolor{red}{x}, \textcolor{blue}{y})$$

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

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

$$M_y: \textcolor{blue}{Y} \rightarrow \mathbb{R}^{\textcolor{red}{X}}$$

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



Embeddings as Functions Over Sets

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

$$\textcolor{blue}{Y} = X \times X = \{(-, -), (-, /), (-, 0), \dots, (\acute{\text{e}}, \text{z}), (\acute{\text{e}}, \acute{\text{e}})\}$$

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

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

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

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

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

$$\mathbb{R}^{\textcolor{red}{X}} \xleftarrow{M_y} \textcolor{blue}{Y}$$

$$M_y: \textcolor{blue}{Y} \rightarrow \mathbb{R}^{\textcolor{red}{X}}$$

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

Embeddings as Functions Over Sets

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

$$\textcolor{blue}{Y} = X \times X = \{(-, -), (-, /), (-, 0), \dots, (\acute{\text{e}}, \text{z}), (\acute{\text{e}}, \acute{\text{e}})\}$$

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

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

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

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

$$M_y: \textcolor{blue}{Y} \rightarrow \mathbb{R}^{\textcolor{red}{X}}$$

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

A commutative diagram illustrating the relationship between the sets $\textcolor{red}{X}$ and $\textcolor{blue}{Y}$ and their embeddings into vector spaces. The diagram consists of four nodes arranged in a square: $\textcolor{red}{X}$ at the top-left, $\mathbb{R}^{\textcolor{blue}{Y}}$ at the top-right, $\mathbb{R}^{\textcolor{red}{X}}$ at the bottom-left, and $\textcolor{blue}{Y}$ at the bottom-right. The horizontal arrows are labeled M_x (from $\textcolor{red}{X}$ to $\mathbb{R}^{\textcolor{blue}{Y}}$) and M_y (from $\textcolor{blue}{Y}$ to $\mathbb{R}^{\textcolor{red}{X}}$). The vertical arrows are unlabeled, with one pointing down from $\textcolor{red}{X}$ to $\mathbb{R}^{\textcolor{red}{X}}$ and another pointing up from $\textcolor{blue}{Y}$ to $\mathbb{R}^{\textcolor{blue}{Y}}$.

Embeddings as Functions Over Sets

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

$$\textcolor{blue}{Y} = X \times X = \{(-, -), (-, /), (-, 0), \dots, (\acute{\text{e}}, \text{z}), (\acute{\text{e}}, \acute{\text{e}})\}$$

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

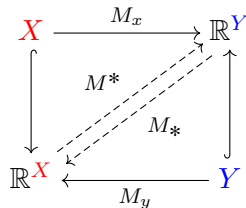
$$(\textcolor{red}{x}, \textcolor{blue}{y}) \mapsto \text{pmi}(\textcolor{red}{x}, \textcolor{blue}{y})$$

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

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

$$M_y: \textcolor{blue}{Y} \rightarrow \mathbb{R}^{\textcolor{red}{X}}$$

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

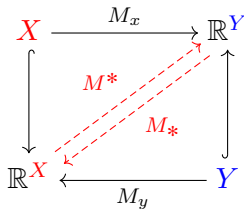


$$M^*: \mathbb{R}^{\textcolor{red}{X}} \rightarrow \mathbb{R}^{\textcolor{blue}{Y}}$$

$$M_*: \mathbb{R}^{\textcolor{blue}{Y}} \rightarrow \mathbb{R}^{\textcolor{red}{X}}$$

Embeddings as Functions Over Sets

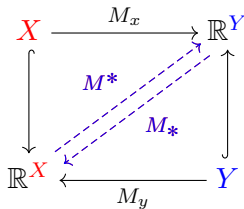
$$M_* M^* : \mathbb{R}^X \rightarrow \mathbb{R}^X$$



Embeddings as Functions Over Sets

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

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



Embeddings as Functions Over Sets

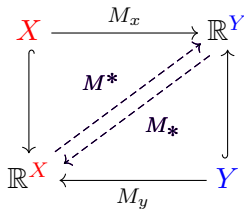
$$M_* M^* : \mathbb{R}^X \rightarrow \mathbb{R}^X$$

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

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

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

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



Embeddings as Functions Over Sets

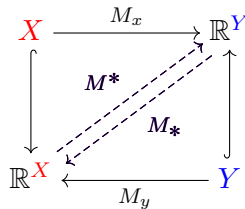
$$M_* M^* : \mathbb{R}^X \rightarrow \mathbb{R}^X$$

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

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

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

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



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

$$M = U \Sigma V^T \quad V := [v_1, \dots, v_n]$$

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

Embeddings as Functions Over Sets

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

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

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

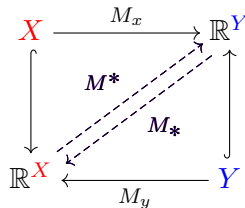
$$\{v_1, \dots, v_n\} \subset \mathbb{R}^Y$$

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

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

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

The u_i and v_i are (linear)
fixed points!



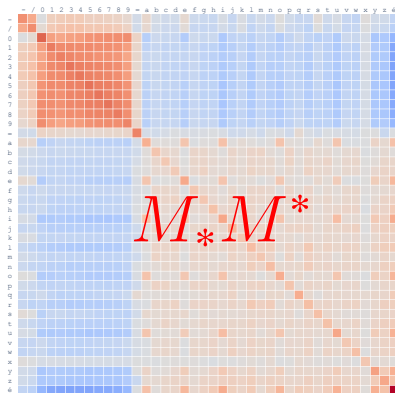
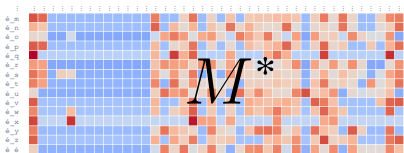
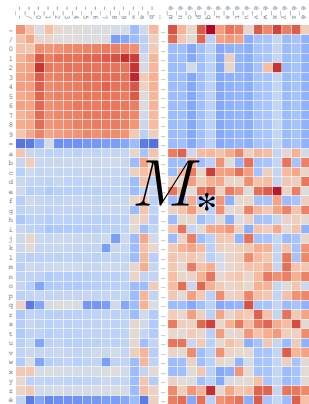
$$M = U \Sigma V^T$$

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

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

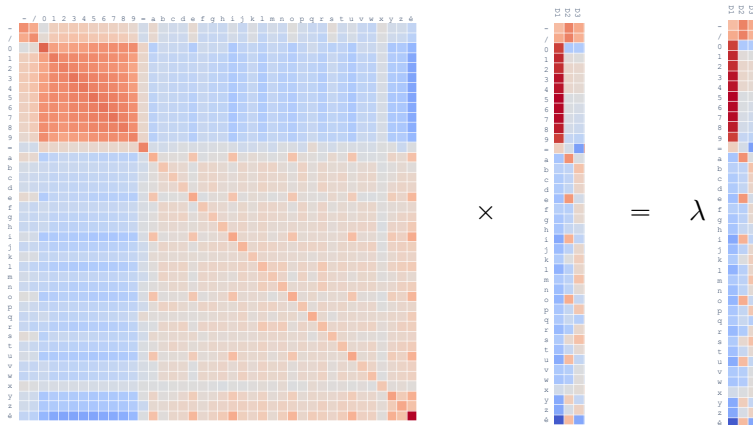
$$\Sigma := \begin{bmatrix} \sqrt{\lambda_1} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sqrt{\lambda_r} \end{bmatrix}$$

M_*M^* as a Covariance Matrix

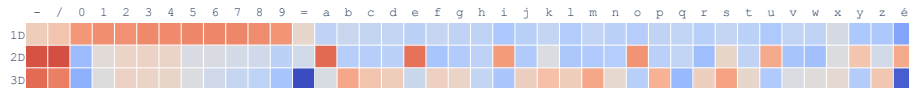


Eigenvectors as Fixed Points

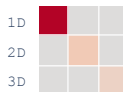
$$M_* M^* u = \lambda u$$



Eigenvectors of M_*M^* :



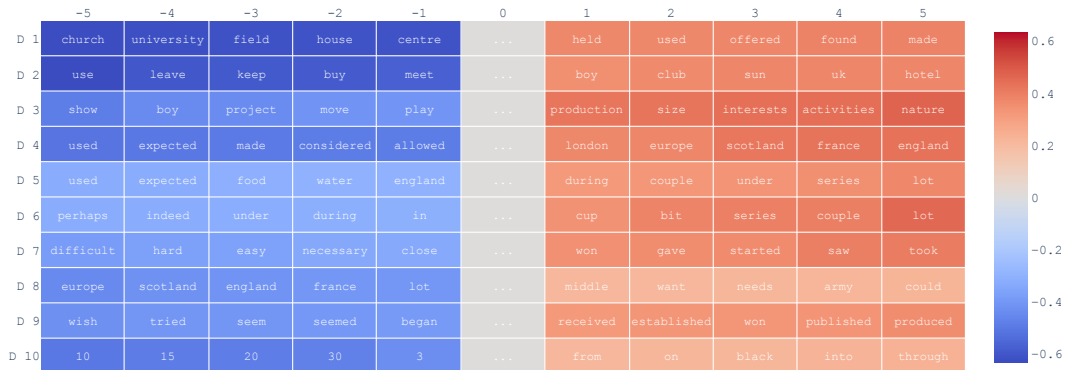
Eigenvalues of M_*M^* and M^*M_* :



Eigenvectors of M^*M_* :



Words



Introduction

Epistemological Perspectives

Theoretical Perspectives

The Algebra Behind the Embeddings

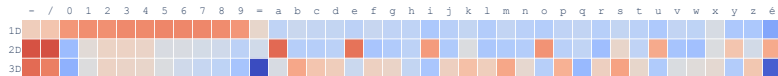
The Structure Behind the Algebra

The Categories Behind the Structure

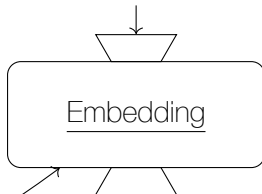
Take Aways

The Structure of Embeddings

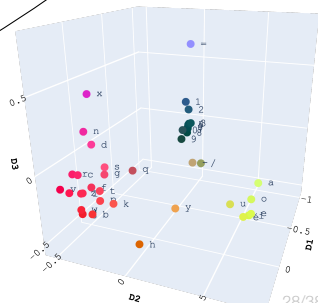
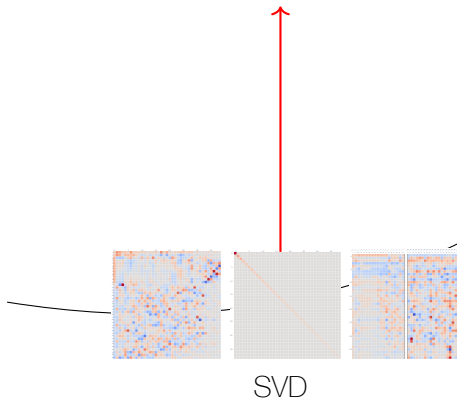
Structure



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

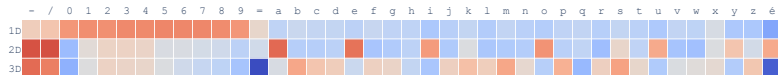


Data

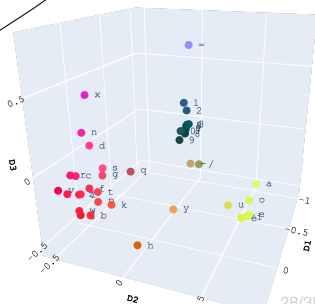
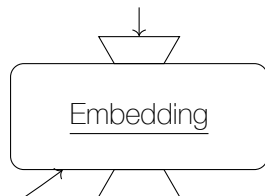
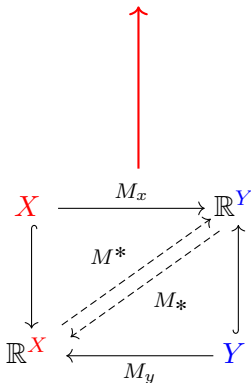


The Structure of Embeddings

Structure



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

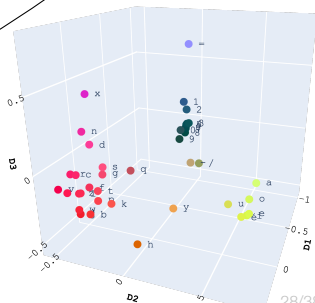
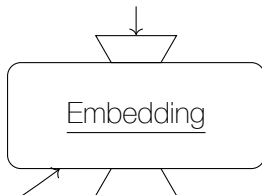


The Structure of Embeddings

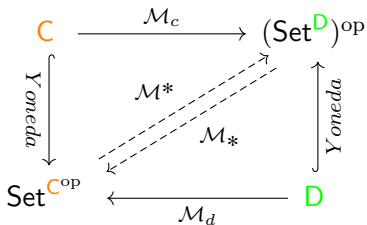
Structure

?

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



Data



Structure

?

$$\begin{array}{ccc} \textit{term}_i & \textit{context}_i & \textit{measure} \\ \downarrow & \downarrow & \swarrow \\ \text{C}^{\text{op}} & \times \text{D} & \rightarrow \text{Set} \end{array}$$

Structure

?

$$\begin{array}{ccc} \textit{term}_i & \textit{context}_i & \textit{measure} \\ \downarrow & \downarrow & \swarrow \\ \text{C}^{\text{op}} & \times \text{D} & \rightarrow \text{Set} \end{array}$$

Structure

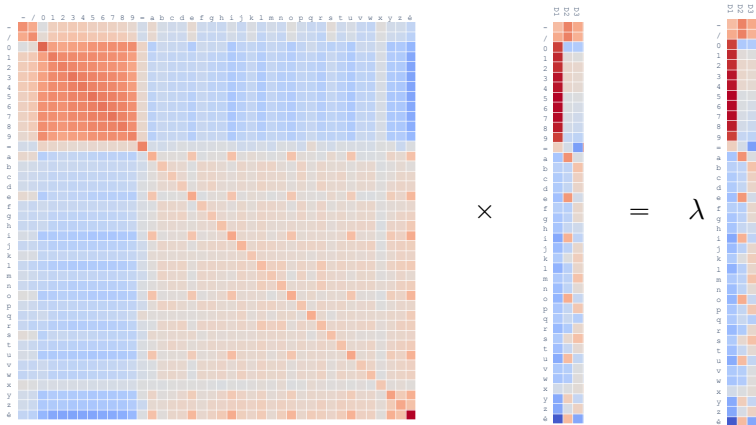
?

$$\begin{array}{ccc} \text{term}_i & \text{context}_i & \text{measure} \\ \downarrow & \downarrow & \downarrow \\ \text{C}^{\text{op}} \times \text{D} & \rightarrow & 2 \end{array}$$

Structure

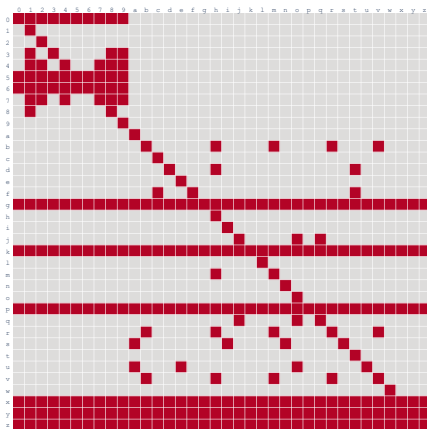
$$\begin{array}{c}
 \mathbf{C}^{\text{op}} \times \mathbf{D} \rightarrow 2 \\
 \Downarrow \\
 \mathcal{M}^* : 2^{\mathbf{C}^{\text{op}}} \rightleftarrows (2^{\mathbf{D}})^{\text{op}} : \mathcal{M}_*
 \end{array}$$

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



Binary Fixed Points

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



★



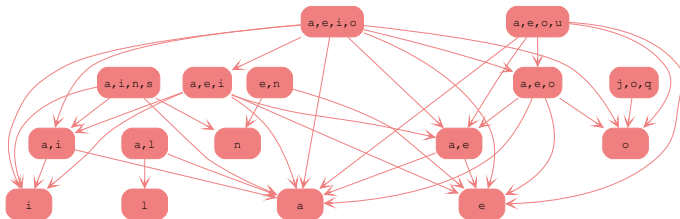
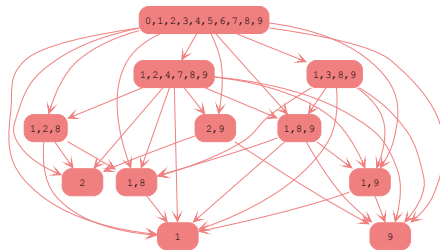
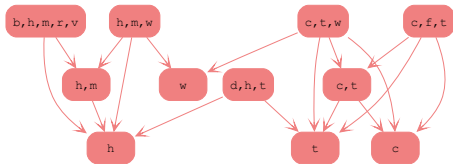
?



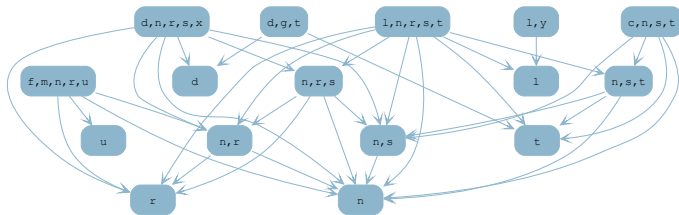
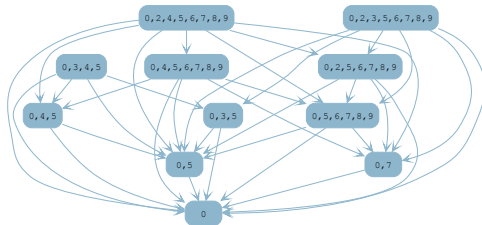
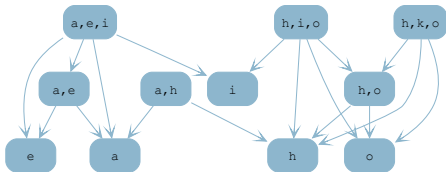
$$\mathcal{M}_* \mathcal{M}^* f = f$$

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	l	9
e,n	a,l	c,t				

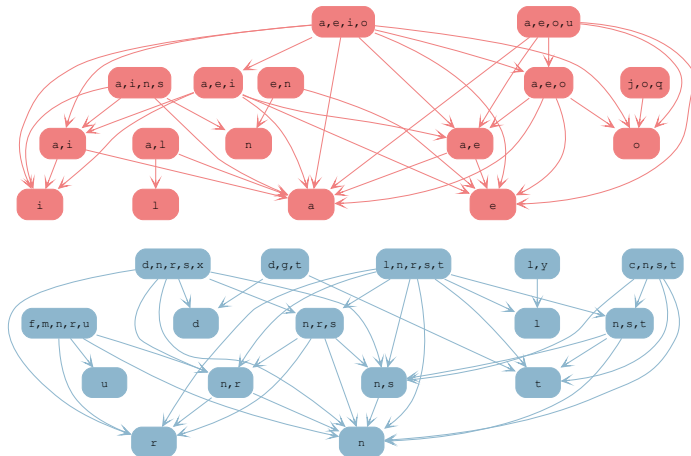
Partial Order Structure



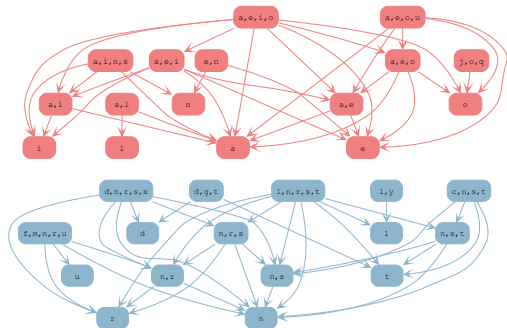
Dual Partial Order



Paring of Partial Ordered Fixed Points



Structure



$$\begin{aligned}
 & \text{C}^{\text{op}} \times \text{D} \rightarrow 2 \\
 & \Downarrow \\
 & \mathcal{M}^* : 2^{\text{C}^{\text{op}}} \rightleftharpoons (2^{\text{D}})^{\text{op}} : \mathcal{M}_*
 \end{aligned}$$

A red arrow points from the $2^{\text{C}^{\text{op}}}$ term in the bottom equation to the top graph.

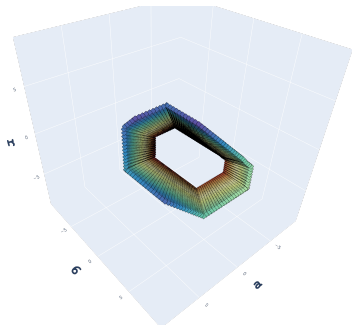
Structure

?

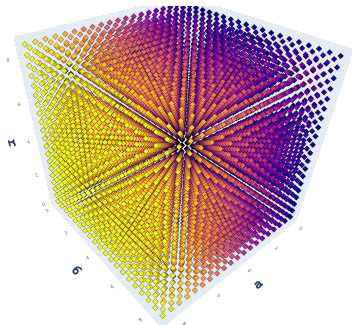
$$\begin{array}{ccc}
 \text{C}^{\text{op}} \times \text{D} & \rightarrow & \bar{\mathbb{R}} \\
 \Downarrow & & \\
 \mathcal{M}^* : \bar{\mathbb{R}}^{\text{C}^{\text{op}}} & \rightleftharpoons & (\bar{\mathbb{R}}^{\text{D}})^{\text{op}} : \mathcal{M}_*
 \end{array}$$

Structure

?

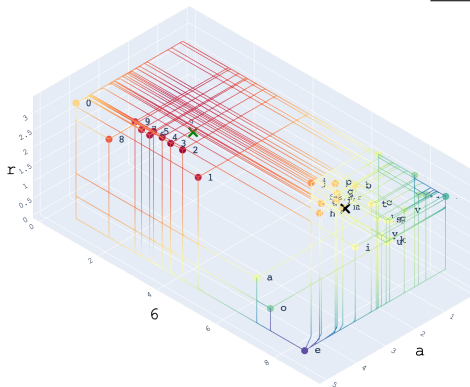


$$\leftarrow \mathcal{M}_* \mathcal{M}^*$$



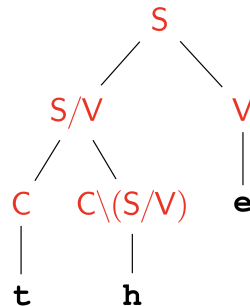
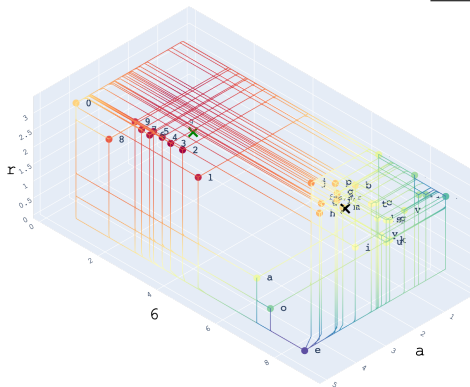
$$\begin{aligned} \mathcal{C}^{\text{op}} \times \mathcal{D} &\rightarrow \bar{\mathbb{R}} \\ &\Downarrow \\ \mathcal{M}^* : \bar{\mathbb{R}}^{\mathcal{C}^{\text{op}}} &\Leftrightarrow (\bar{\mathbb{R}}^{\mathcal{D}})^{\text{op}} : \mathcal{M}_* \end{aligned}$$

Structure



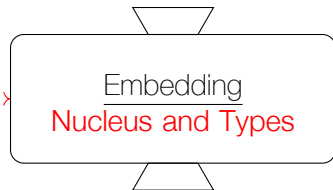
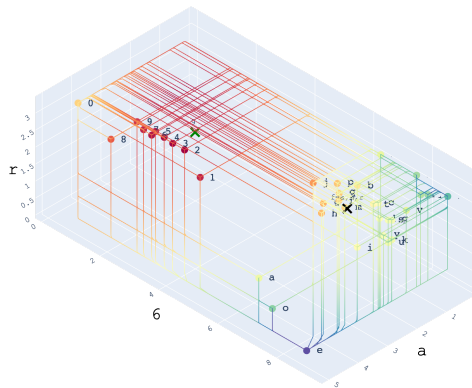
$$\begin{array}{c}
 \text{C}^{\text{op}} \times \text{D} \rightarrow \bar{\mathbb{R}} \\
 \Downarrow \\
 \mathcal{M}^* : \bar{\mathbb{R}}^{\text{C}^{\text{op}}} \rightleftharpoons (\bar{\mathbb{R}}^{\text{D}})^{\text{op}} : \mathcal{M}_*
 \end{array}$$

Structure



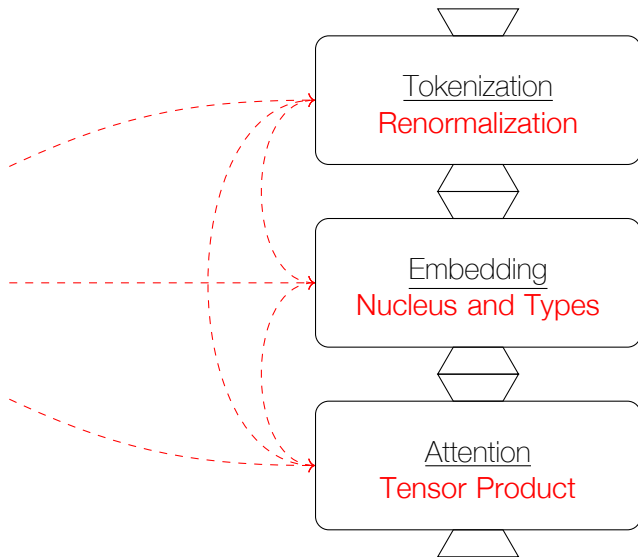
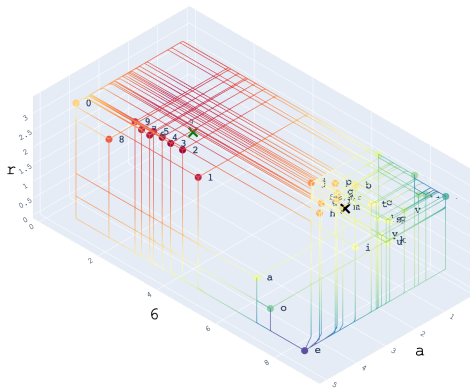
$$\begin{aligned}
 & \mathbf{C}^{\text{op}} \times \mathbf{D} \rightarrow \bar{\mathbb{R}} \\
 & \Downarrow \\
 & \mathcal{M}^* : \bar{\mathbb{R}}^{\mathbf{C}^{\text{op}}} \rightleftharpoons (\bar{\mathbb{R}}^{\mathbf{D}})^{\text{op}} : \mathcal{M}_*
 \end{aligned}$$

Structure



Formal Explainability

Structure



Introduction

Epistemological Perspectives

Theoretical Perspectives

- The Algebra Behind the Embeddings

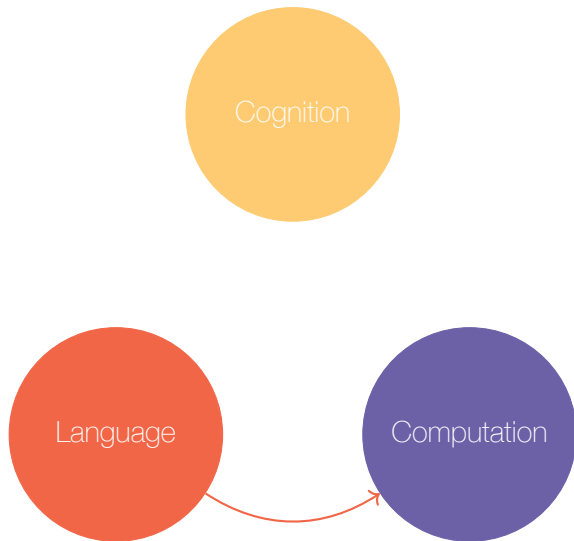
- The Structure Behind the Algebra

- The Categories Behind the Structure

Take Aways

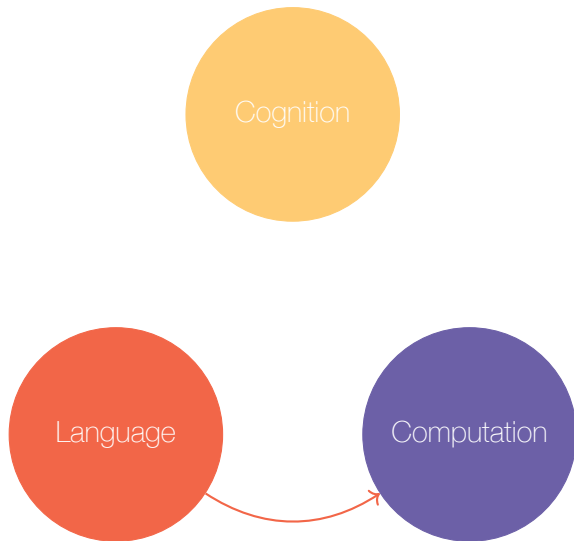
Take Aways

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



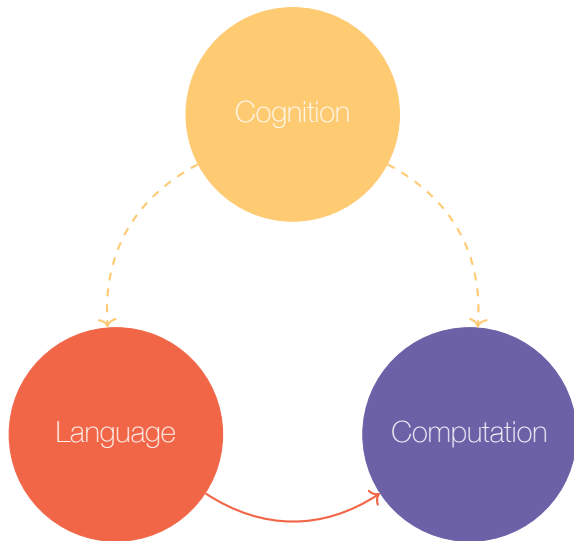
Take Aways

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



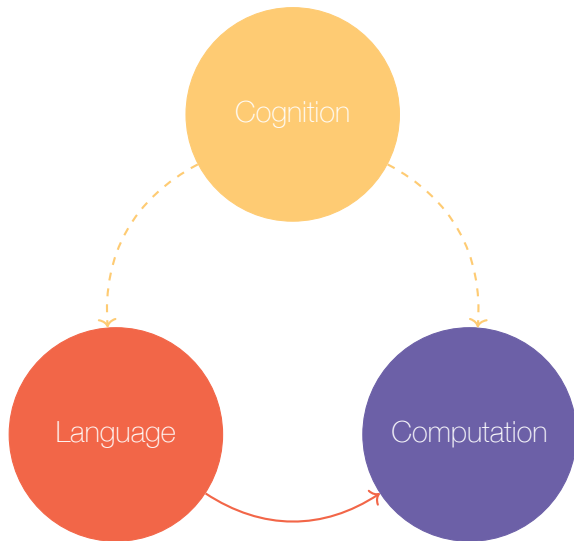
Take Aways

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



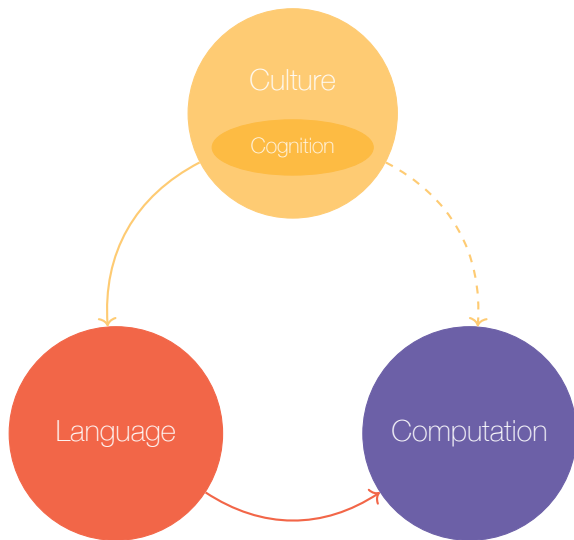
Take Aways

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



Take Aways

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



Collaborations



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

Reference Papers

- ◇ Gastaldi, J. L. (2021). Why Can Computers Understand Natural Language? *Philosophy & Technology*, 34(1), 149–214. <https://doi.org/10.1007/s13347-020-00393-9>
- ◇ Gastaldi, J. L., & Pellissier, L. (2021). The calculus of language: explicit representation of emergent linguistic structure through type-theoretical paradigms. *Interdisciplinary Science Reviews*, 46(4), 569–590. <https://doi.org/10.1080/03080188.2021.1890484>
- ◇ Bradley, T.-D., Gastaldi, J. L., & Terilla, J. (2024). The structure of meaning in language: Parallel narratives in linear algebra and category theory. *Notices of the American Mathematical Society*. <https://api.semanticscholar.org/CorpusID:263613625>

References I

- Belrose, N., Schneider-Joseph, D., Ravfogel, S., Cotterell, R., Raff, E., & Biderman, S. (2024). Leace: Perfect linear concept erasure in closed form. *Proceedings of the 37th International Conference on Neural Information Processing Systems*.
- Bradley, T.-D., Gastaldi, J. L., & Terilla, J. (2024). The structure of meaning in language: Parallel narratives in linear algebra and category theory. *Notices of the American Mathematical Society*.
<https://api.semanticscholar.org/CorpusID:263613625>
- Chomsky, N. (1953). Systems of syntactic analysis. *Journal of Symbolic Logic*, 18(3), 242–256.
<https://doi.org/10.2307/2267409>
- Chomsky, N. (1956). Three models for the description of language. *IRE Transactions on Information Theory*, 2(3), 113–124. <https://doi.org/10.1109/TIT.1956.1056813>
- Chomsky, N. (1957). *Syntactic structures*. Mouton; Co.
- Chomsky, N. (1959). *Language*, 35(1), 26–58. Retrieved July 7, 2025, from <http://www.jstor.org/stable/411334>
- Chomsky, N. (1992, November). Language and the “cognitive revolutions” [Delivered November 23–27, 1992].
- Delétang, G., Ruoss, A., Grau-Moya, J., Genewein, T., Wenliang, L. K., Catt, E., Cundy, C., Hutter, M., Legg, S., Veness, J., & Ortega, P. A. (2023). Neural networks and the chomsky hierarchy.
<https://arxiv.org/abs/2207.02098>
- Gastaldi, J. L. (2021). Why Can Computers Understand Natural Language? *Philosophy & Technology*, 34(1), 149–214.
<https://doi.org/10.1007/s13347-020-00393-9>
- Gastaldi, J. L., & Pellissier, L. (2021). The calculus of language: explicit representation of emergent linguistic structure through type-theoretical paradigms. *Interdisciplinary Science Reviews*, 46(4), 569–590.
<https://doi.org/10.1080/03080188.2021.1890484>
- Girard, J.-Y. (2011, September). *The blind spot*. European Mathematical Society.

References II

- Goldberg, Y., & Levy, O. (2014). Word2vec explained: Deriving mikolov et al.'s negative-sampling word-embedding method. *CoRR*, *abs/1402.3722*.
- Levy, O., & Goldberg, Y. (2014). Neural word embedding as implicit matrix factorization. *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, 2177–2185.
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J., Le, Q., & Strohmann, T. (2013). *Learning representations of text using neural networks. NIPS deep learning workshop 2013 slides*.
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *CoRR*, *abs/1310.4546*.
- Sennrich, R., Haddow, B., & Birch, A. (2016). Neural machine translation of rare words with subword units. *Proceedings of the 54th Annual Meeting of the ACL*, 1715–1725.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 30). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

NeuroMod Annual Meeting
Université Côte d'Azur
Antibes, France

What are Neural Language Models the Model of?
Epistemological and Theoretical Perspectives on LLMs

Juan Luis Gastaldi
www.giannigastaldi.com

ETH zürich

July 8, 2025