

MALINCA
Institut Henri Poincaré
Paris, France

Language and Mathematics in the Time of AI
Philosophical and Theoretical Perspectives

Juan Luis Gastaldi
www.giannigastaldi.com

ETH zürich

October 1, 2025

Outline

Bring-Ins

Philosophical Perspectives

~~Historical Perspectives~~

Epistemological Perspectives

Conceptual Perspectives

Theoretical Perspectives

Takeaways

Bring-Ins

Philosophical Perspectives

Historical Perspectives

Epistemological Perspectives

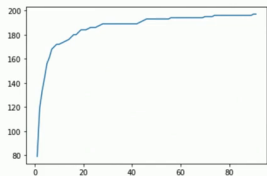
Conceptual Perspectives

Theoretical Perspectives

Takeaways

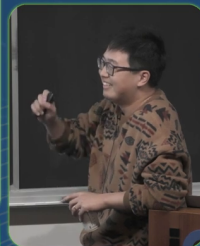
Main results – human prover

- For each problem, given a ground truth human solution, we sample up to 100 proofs from Codex.
- Codex solves 200 out of 488 problems with 100 samples:



Autoformalization with Large Language Models

MACHINE ASSISTED PROOFS, FEBRUARY 13 - 17, 2023, WWW.IPAM-UCLA.EDU



Tony Wu
Google

Tony Wu, *Autoformalization with Large Language Models* (IPAM (UCLA), Feb 15, 2023)

Melanie Mitchell on PaLM2



Melanie Mitchell

@MelMitchell1

...

Weird statement from Google's Palm 2 announcement.

(from [blog.google/technology/ai/...](https://blog.google/technology/ai/))

- **Reasoning:** PaLM 2's wide-ranging dataset includes scientific papers and web pages that contain mathematical expressions. As a result, it demonstrates improved capabilities in logic, common sense reasoning, and mathematics.

6:23 PM · May 12, 2023 · **52.9K** Views

<https://blog.google/technology/ai/google-palm-2-ai-large-language-model/>

Outline

Bring-Ins

Philosophical Perspectives

Historical Perspectives

Epistemological Perspectives

Conceptual Perspectives

Theoretical Perspectives

Takeaways

What's So Funny?

- ◇ Proof assistants and automated theorem provers were supposed to prevent the frailty of human proving practices instead of reproducing them and relying on them as a gold standard.

What's So Funny?

- ◇ **Proof assistants and automated theorem provers** were supposed to **prevent** the frailty of human proving practices instead of **reproducing** them and **relying** on them as a gold standard.
- ◇ **Mathematical literacy** was supposed to be **a condition** to write and read scientific papers and mathematical expressions **instead of their miraculous effect**.

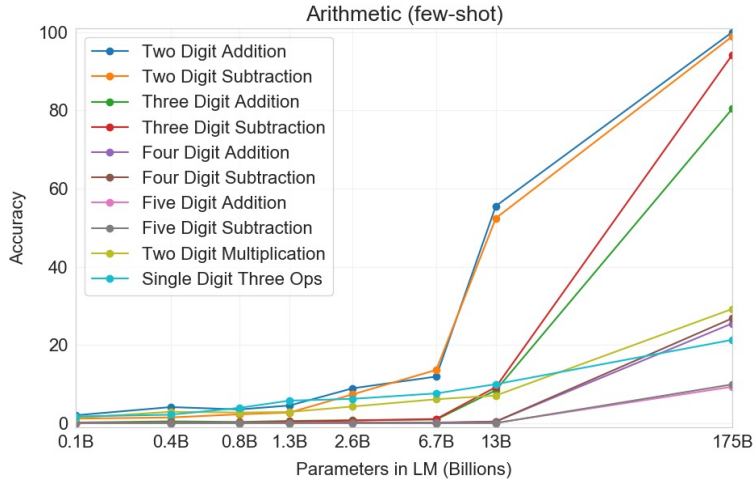
What's So Funny?

- ◇ **Proof assistants and automated theorem provers** were supposed to **prevent** the frailty of human proving practices instead of **reproducing** them and **relying** on them as a gold standard.
- ◇ **Mathematical literacy** was supposed to be **a condition** to write and read scientific papers and mathematical expressions **instead of their miraculous effect**.
- ◇ **Natural language** was considered **the cause of** rather than **the solution to** the multiple problems preventing mathematics from achieving higher degrees of precision.

What's So Funny?

- ◇ **Proof assistants and automated theorem provers** were supposed to **prevent** the frailty of human proving practices instead of **reproducing** them and **relying** on them as a gold standard.
- ◇ **Mathematical literacy** was supposed to be **a condition** to write and read scientific papers and mathematical expressions **instead of their miraculous effect**.
- ◇ **Natural language** was considered **the cause of** rather than **the solution to** the multiple problems preventing mathematics from achieving higher degrees of precision.
- ◇ More generally, **the *formal* nature of mathematics** was believed to make it **impassive to the strong empirical position** assumed by connectionist approaches guiding the application of DNNs.

Arithmetic in GPT-3



(Brown et al., 2020)

Reasoning in ChatGPT o3

If p_n is a sequence of probability distributions on a countably infinite set X that converge pointwise to a probability distribution p on X , do they necessarily converge uniformly?

Reasoned about uniform convergence of probability distributions for 4m 41s >

We will show that if

$$p_n(x) \rightarrow p(x) \quad \text{for every } x \in X,$$

where X is a countably infinite set and each p_n and p is a probability distribution on X (so that

$$\sum_{x \in X} p_n(x) = \sum_{x \in X} p(x) = 1),$$

then in fact the convergence is uniform; that is,

$$\sup_{x \in X} |p_n(x) - p(x)| \rightarrow 0.$$

In other words, point-wise convergence of the probability mass functions on a countable set (with the limit being a probability distribution) forces uniform convergence.

In what follows we give a proof.

Proof

Let $\varepsilon > 0$ be arbitrary. Our goal is to show that for all sufficiently large n we have

$$\sup_{x \in X} |p_n(x) - p(x)| < \varepsilon.$$

Because p is a probability measure on the countable set X , we can “capture almost all of its mass” in a finite set. (This is sometimes called a truncation argument.) More precisely, there exists a finite subset $F \subset X$

such that

$$p(X \setminus F) = \sum_{x \notin F} p(x) < \varepsilon/3.$$

Now, since F is finite and we know that for every fixed x we have $p_n(x) \rightarrow p(x)$, it follows that for each $x \in F$ there is an index N_x such that for all $n \geq N_x$

$$|p_n(x) - p(x)| < \varepsilon/(3|F|).$$

If we let

$$N_1 = \max\{N_x : x \in F\},$$

then for all $n \geq N_1$ and every $x \in F$ we have

$$|p_n(x) - p(x)| < \frac{\varepsilon}{3|F|}.$$

Thus, summing over $x \in F$ we obtain

Conclusion

Yes. If a sequence $\{p_n\}$ of probability distributions on a countably infinite set X converges point-wise to a probability distribution p , then in fact

$$\sup_{x \in X} |p_n(x) - p(x)| \rightarrow 0,$$

so the convergence is uniform.

A Note on the Issue

It is very important that the limit p is itself a probability distribution. (For example, the sequence of "Dirac measures" δ_n on $\mathbb{N}$ defined by

$$\delta_n(x) = \begin{cases} 1 & \text{if } x = n, \\ 0 & \text{otherwise,} \end{cases}$$

converges point-wise to the zero function; but the zero function is not a probability distribution, and indeed the convergence is not uniform.) The fact that no mass "escapes" ensures that the tiny probabilities in the tail force the errors there to be small uniformly.

Final Answer

Yes. If p_n is a sequence of probability distributions on a countably infinite set X that converges point-wise to a probability distribution p (so that no mass is lost), then in fact the convergence is uniform.

Reasoning in ChatGPT o3

Reasoned about uniform convergence of probability distributions for 4m 41s ✓

I'm curious if pointwise convergence of probability distributions on a countably infinite set implies uniform convergence. This involves examining the supremum of the absolute difference between distributions as it approaches zero.

OK, let me see: pointwise convergence on a countably infinite set doesn't guarantee uniform convergence without additional constraints, like finite sets or specific continuity conditions.

I'm curious about sequences of probability measures where each point's probability trends to zero. This hints at pointwise convergence but raises questions about uniform convergence due to potential "spikes."

OK, let's think it through: the maximum difference of $1/n$ between distributions p_n and p as n increases. This difference converges to zero, suggesting uniform convergence in a bounded setup.

I'm examining a counterexample of pointwise convergence without uniform convergence, where individual probabilities decline to zero but the overall distribution doesn't converge uniformly due to continuous fluctuation.

I'm considering pointwise convergence with a shifting spike in a probability distribution on natural numbers, highlighting the gap between pointwise and uniform convergence due to inconsistent distribution updates.

Main Orientations in DNN Applications to Maths

◇ Proof-Oriented

- Bansal et al., 2019; Polu and Sutskever, 2020; Wu et al., 2022.

◇ Object-Oriented

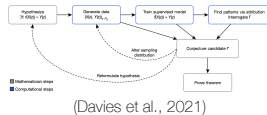
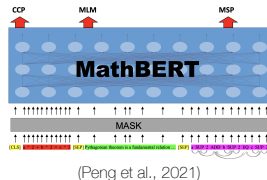
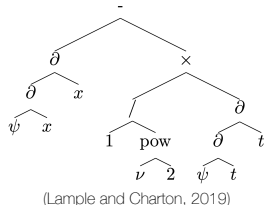
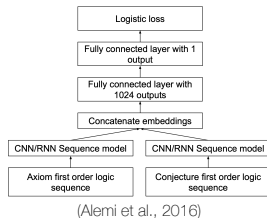
- Blechschmidt and Ernst, 2021; Charton, 2021; d'Ascoli et al., 2022; Lample and Charton, 2019; Li et al., 2021

◇ Skill-Oriented (e.g., Reasoning)

- Brown et al., 2020; Lewkowycz et al., 2022; Shen et al., 2021

◇ Heuristic-Oriented

- Davies et al., 2021; Wagner, 2021



Philosophical Significance: The Return of Language

- ◇ Research orientations tend to be spontaneously organized according to the **AI researchers' implicit assumptions** as to what characterizes **mathematical practice** (i.e. what it is that we do when we do mathematics).

Philosophical Significance: The Return of Language

- ◇ Research orientations tend to be spontaneously organized according to the **AI researchers' implicit assumptions** as to what characterizes **mathematical practice** (i.e. what it is that we do when we do mathematics).
- ◇ However, practically all applications share a common philosophical assumption: ***Written natural language plays a critical role in the processing mathematical knowledge.***

Philosophical Significance: The Return of Language

- ◇ Research orientations tend to be spontaneously organized according to the **AI researchers' implicit assumptions** as to what characterizes **mathematical practice** (i.e. what it is that we do when we do mathematics).
- ◇ However, practically all applications share a common philosophical assumption: ***Written natural language plays a critical role in the processing mathematical knowledge.***
- ◇ The potential success of DNN methods in mathematics is inseparable from a **reorientation of the epistemology of mathematics** from logic and formal systems to **natural language** and **vernacular writing practices**.

Bring-Ins

Philosophical Perspectives

Historical Perspectives

Epistemological Perspectives

Conceptual Perspectives

Theoretical Perspectives

Takeaways

Outline

Bring-Ins

Philosophical Perspectives

Historical Perspectives

Epistemological Perspectives

Conceptual Perspectives

Theoretical Perspectives

Takeaways

Stochastic parrots vs. AGI

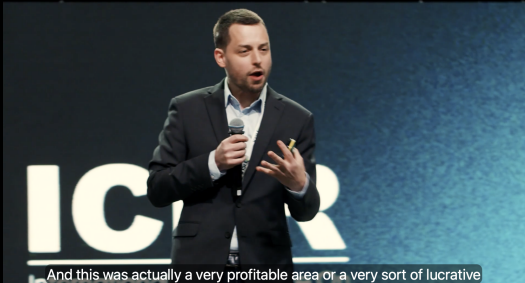


LLMs are not like us,
therefore they do not and can not have any relation to natural language.



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

Empirical Saturnalia



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

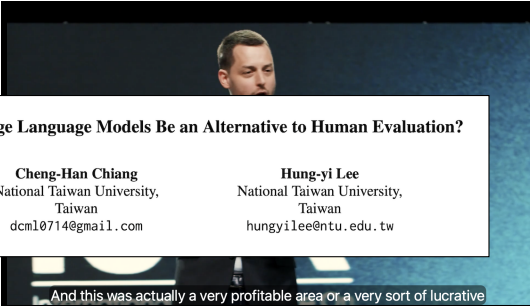
The empirics of deep learning

(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 Bezalet

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 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 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 Reasoning 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 Han^{1*}, Chaitin Hui², Dongyue Hu², Kevin H. Lee^{3*}, Shihang Ren²,
Udhay Narayanan¹,
¹University of Wisconsin-Madison, ²Google Research, ³Google Research
jh2324@wisc.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

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

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

Do LLMs “KNOW” INTERNALLY WHEN THEY FOLLOW INSTRUCTIONS?

Juyeon Hwang^{1*}, Christian Hutter², David Rosenberg², Kevin H. Poon^{3*}, Shih-Feng Lee²,
Udhay N. Singh¹, and
¹University of
jh2324@

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 Han^{1*}, Chaitin Hui², Danqi Chen², Gaurav Erubas², Kuan H. Lee^{3*}, Shih-Fu Lee³,
Udhay Narayanan¹,
¹University of Washington, ²Google Research, 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**

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*, **Christine Hui**, **Danqi**, **Qiyuan**, **Elad**, **Karen H. Lee**, **Chao**, **Shih**, **Bar**
¹University of California, Berkeley
jh2324@berkeley.edu

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

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

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

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

Language Models are Few-Shot Learners

Scaling Mathematical Topic Tree Evaluation of LLMs

M. Davoudi, Pouya Pezeshkpour²
Megagon Labs

K. Ryder*

Melanie Subbiah*

on

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 COMMON KNOWLEDGE?

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

Christian H. Hoerl^{1,2} Owen Etzioni² Kevin H. Poon^{3*} Shih-Feng Lee²

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

Language Models are Few-Shot Learners

Long Mathematical Topic Tree Evaluation of LLMs

Mehdi Davoudi, Pouya Pezeshkpour²
Megagon Labs

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

Kyle 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

Scaling Mathematical Topic Tree
Evaluation of LLMs

Davoudi, Pouya Pezeshkpour²
Megagon 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 Language Models Invent Algorithms to Improve Themselves?: Algorithms Discovered for Improving 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

n[†]

Ryder* Melanie Subbiah*

Pranav Shyam Girish Sastry

Gretchen Krueger Tom Henighan

Empirical Saturnalia

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

PROCEDURAL REASONING IN

Laura Ruis*
AI Centre, UCL

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

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

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

Empirical Saturnalia

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

PROCEDURAL REASONING IN

Laura Ruis*
AI Centre, UCL

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

TEACHING ARITHMETIC TO SMALL TRANSFORMERS

Nayoung Lee*
University of Wisconsin-Madison
nayoung.lee@wisc.edu

Kartik Sreenivasan*
University of Wisconsin-Madison
ksreenivasa2@wisc.edu

Jason D. Lee
Princeton University
jasonlee@princeton.edu

Kangwook Lee
University of Wisconsin-Madison
kangwook.lee@wisc.edu

ing Mathematical Topic Tree valuation of LLMs

Davoudi, Pouya Pezeshkpour²
Megagon Labs

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

DO LLMs HAVE CON

Sparks E

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
The Univer

Ryo Kamoi¹ Yusen Zhang¹ Nan Zhang¹ Jiawei Han² Rui Zhang¹

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

{ryokamoi, rmz5227}@psu.edu

Empirical Saturnalia

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

DO LLMs HAVE CON

Sparks

PROCEDURAL
REASONING IN

Laura Ruis*
AI Centre, UCL

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

TEACHING ARITHMETIC TO SMALL TRANSFORMERS

Nayoung Lee*
University of Wisconsin-Madison

Kartik Sreenivasan*
University of Wisconsin-Madison

IS YOUR MODEL REALLY A GOOD MATH REASONER?
EVALUATING MATHEMATICAL REASONING WITH
CHECKLIST

Zihao Zhou^{12*} Shudong Liu^{3*} Maizhen Ning¹²⁶ Wei Liu⁴ Jindong Wang⁵
Derek F. Wong³ Xiaowei Huang² Qiufeng Wang^{1†} Kaizhu Huang⁶
¹Xi'an Jiaotong-liverpool University ²University of Liverpool ³University of Macau
⁴HKUST ⁵William & Mary ⁶Duke Kunshan University

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

Machel Reid
Google Research*

Yutaka Matsuo
The University of Tokyo

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

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

Gretchen Krueger Tom Henighan

Empirical Saturnalia

Can Large Language Models Invent Algorithms to Improve Themselves?:

Algorithm Discovery for Recursive Self-Improvement through

DO LLMs HAVE CON

Sparks

E

Naam
Tel. A

Sébastien Bubeck
Eric Horvitz Ece Kar
Harsha Nori

**PROCEDURAL
REASONING IN**

Laura Ruis*
AI Centre, UCL

Hamid Palangi Ma

Microsoft Rese

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

Xue-

TEACHING ARITHMETIC TO SMALL TRANSFORMERS

Nayoung Lee*
University of Wisconsin-Madison

Kartik Sreenivasan*
University of Wisconsin-Madison

REASONER?
REASONING WITH

Wang⁵
ng⁶
University of Macau

Gretchen Krueger Tom Henighan

**Learning Formal Mathematics
From Intrinsic Motivation**

Gabriel Poesia¹ David Broman⁴ Nick Haber^{1,3} Noah D. Goodman^{1,2}
{poesia,nhaber,ngoodman}@stanford.edu dbro@kth.se
Departments of Computer Science¹, Psychology² and Education³, Stanford University, USA
EECS and Digital Futures⁴, KTH Royal Institute of Technology, Sweden

Machel Reid
Google Research* The

Ryo Kamoi¹ Yusen Zhan
¹Penn State University, USA
{ryokan

Takeshi Kojima
The University of To
t.kojima@weblab.t.u-to

Empirical Saturnalia

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

DO LLMs HAVE CON

PROCEDURAL
REASONING IN

Laura Ruis*
AI Centre, UCL

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

TEACHING ARITHMETIC TO SMALL TRANSFORMERS

Nayoung Lee*
University of Wisconsin-Madison

Kartik Sreenivasan*
University of Wisconsin-Madison

Takeshi Kojima

(How) Can Transformers Predict Pseudo-Random Numbers?

Tao Tao^{*1} Darshil Doshi^{*1} Dayal Singh Kalra^{*2} Tianyu He^{*1} Maissam Barkeshli¹³
{tao2021, ddoshi, dayal, tianyuhe, maissam}@umd.edu

{ryokan

Gabriel Poesia¹

David Broman⁴

Nick Haber^{1,3}

Noah D. Goodman^{1,2}

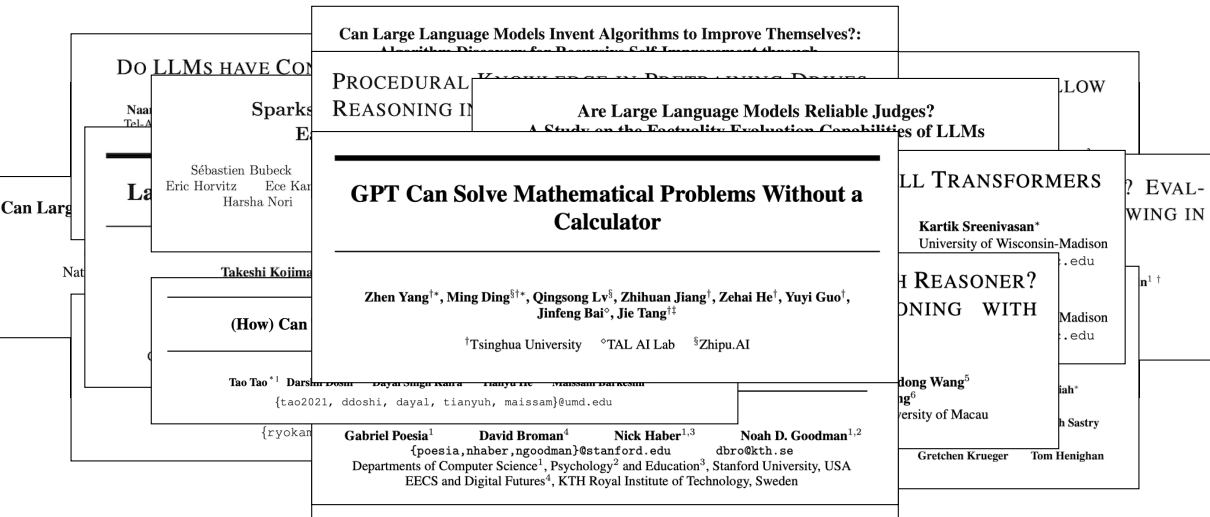
{poesia, nhaber, ngoodman}@stanford.edu dbro@kth.se

Departments of Computer Science¹, Psychology² and Education³, Stanford University, USA
EECS and Digital Futures⁴, KTH Royal Institute of Technology, Sweden

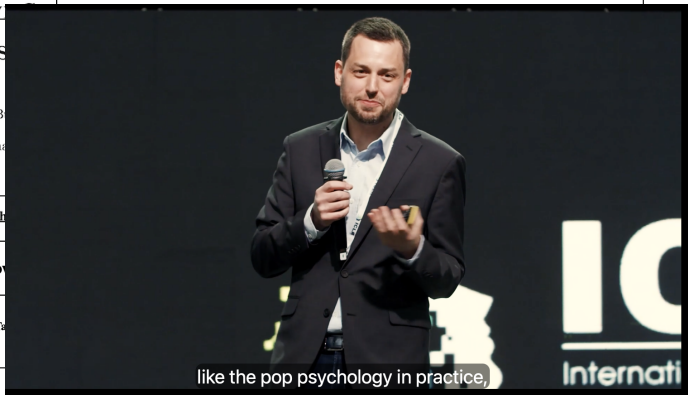
REASONER?
ONING WITH

dong Wang⁵
ng⁶
iversity of Macau

Gretchen Krueger Tom Henighan

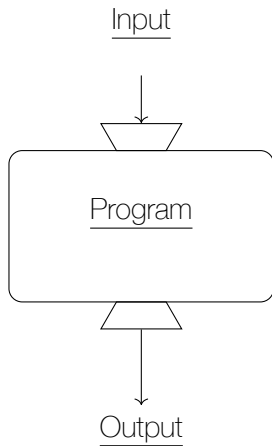


Can Large Language Models Invent Algorithms to Improve Themselves?:
Algorithmic Discovery for Recursive Self-Improvement through

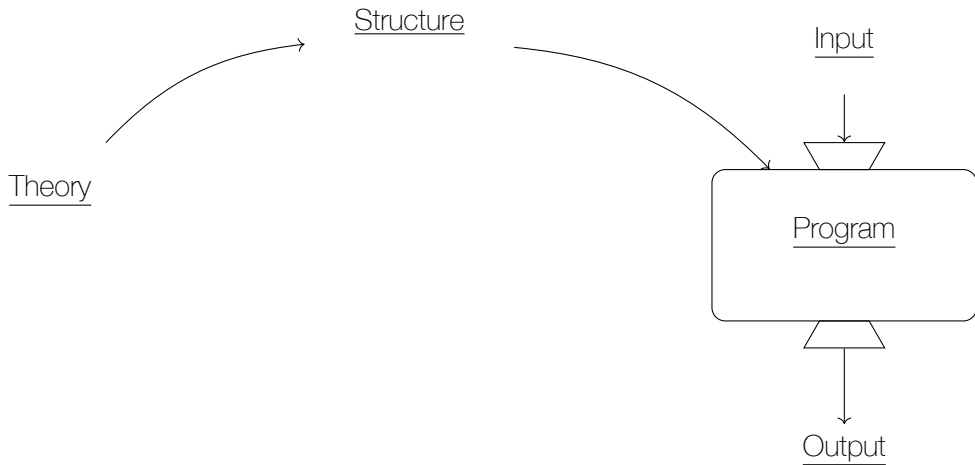


Departments of Computer Science¹, Psychology² and Education³, Stanford University, USA
EECS and Digital Futures⁴, KTH Royal Institute of Technology, Sweden

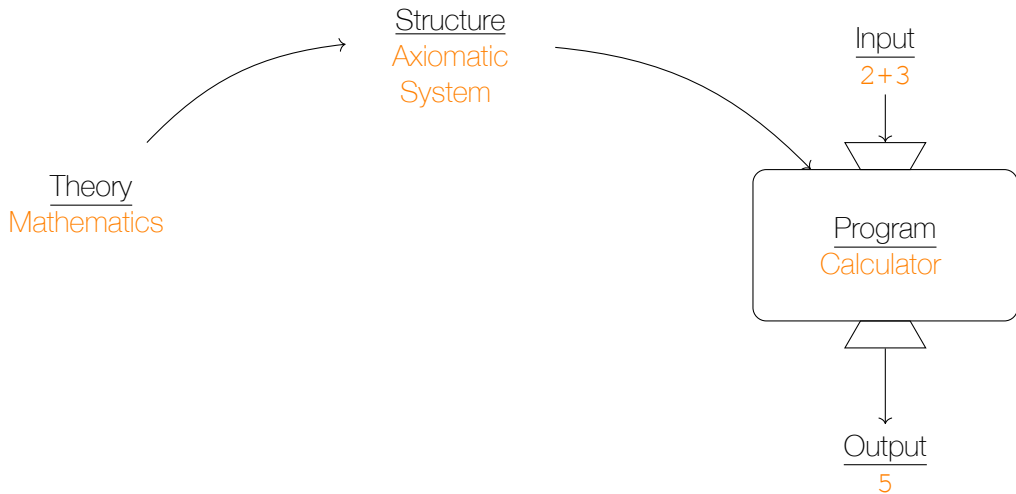
Programs As Formal Objects



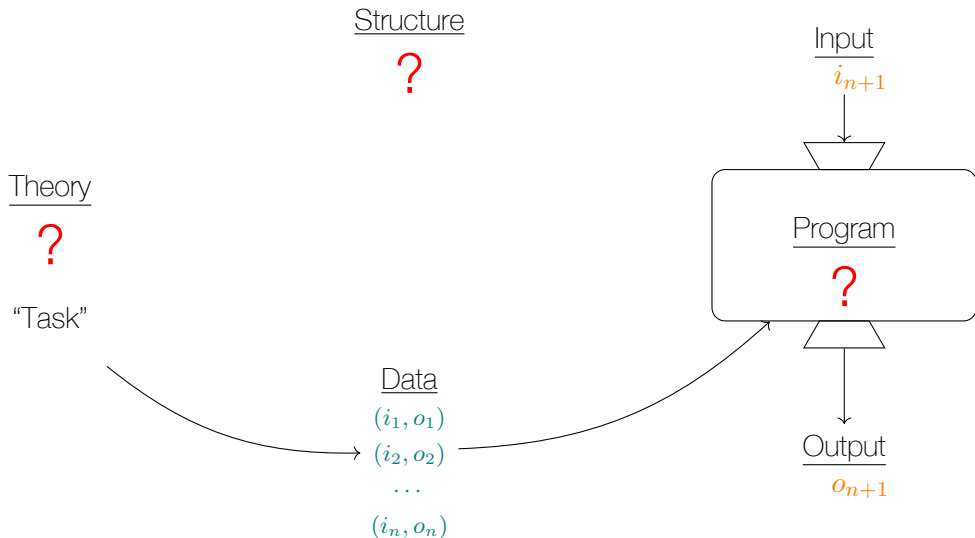
Programs As Formal Objects



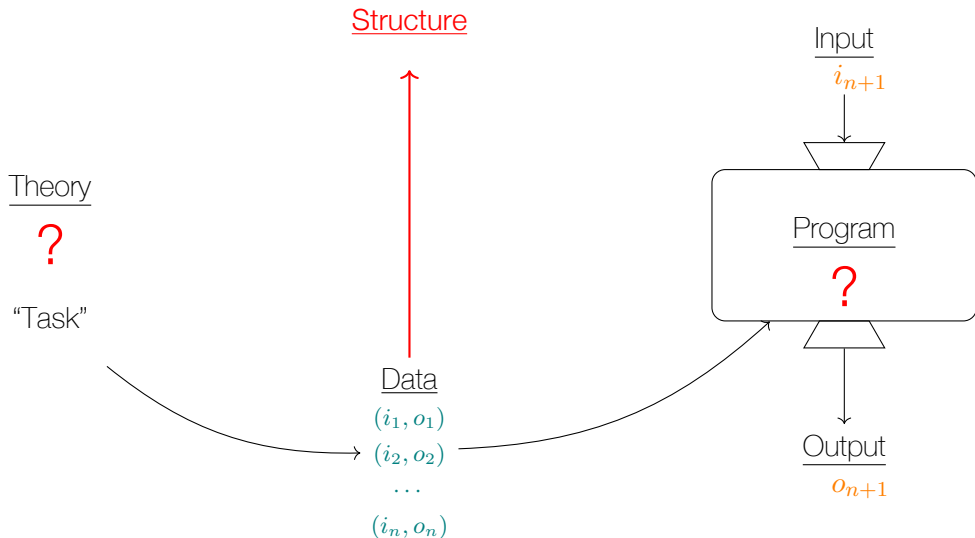
Programs As Formal Objects



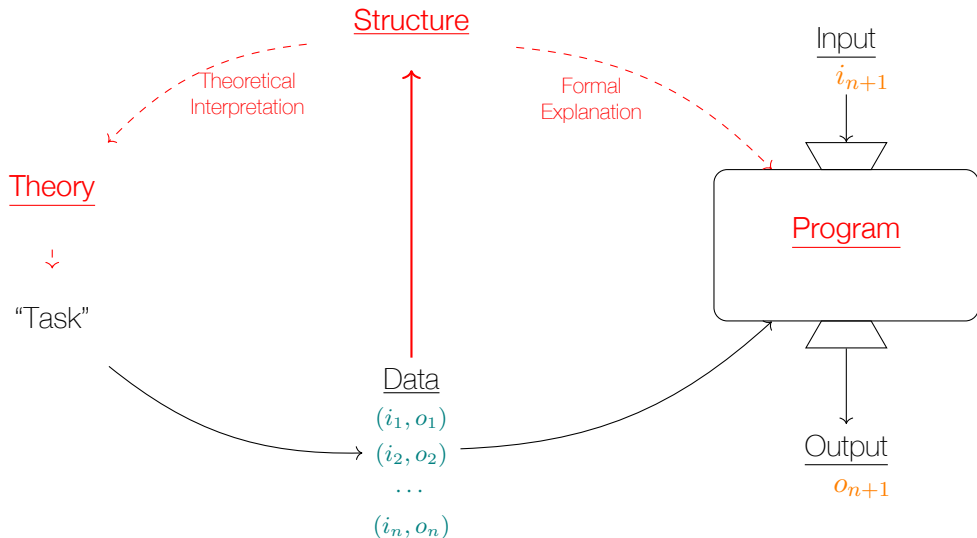
Programs As Formal Objects



Programs As Formal Objects



Programs As Formal Objects



New Research Orientation

- ◇ **Why and how**, *irrespective of the model*, a **phenomenon** under study can be captured through the **distributional analysis of data** at all?
- ◇ What are, *irrespective of the model*, the **general principles** for capturing **distributional phenomena**?

Outline

Bring-Ins

Philosophical Perspectives

Historical Perspectives

Epistemological Perspectives

Conceptual Perspectives

Theoretical Perspectives

Takeaways

Distributionalism and Word Embeddings

◇ Distributional Hypothesis

(Harris, 1960; Saussure, 1959)

- “You shall know a word by the company it keeps!”
(Firth, 1935)
- “Words which are similar in meaning occur in similar contexts” (Rubenstein & Goodenough, 1965)
- The content of a linguistic unit is determined by its **distribution** over a corpus (i.e., the other units appearing in its context)

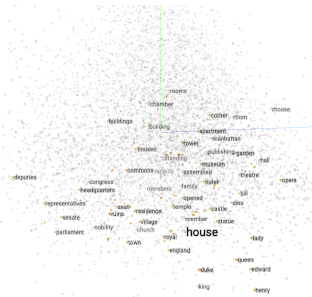
Distributionalism and Word Embeddings

◇ Distributional Hypothesis

(Harris, 1960; Saussure, 1959)

- “You shall know a word by the company it keeps!” (Firth, 1935)
- “Words which are similar in meaning occur in similar contexts” (Rubenstein & Goodenough, 1965)
- The content of a linguistic unit is determined by its **distribution** over a corpus (i.e., the other units appearing in its context)

◇ Computational version: **Word Embeddings**



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

Distributionalism and Word Embeddings

◇ Distributional Hypothesis

(Harris, 1960; Saussure, 1959)

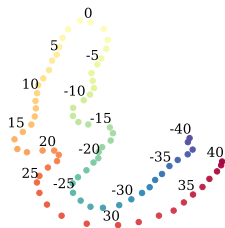
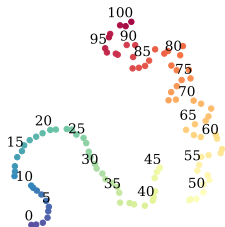
- “You shall know a word by the company it keeps!”
(Firth, 1935)
- “Words which are similar in meaning occur in similar contexts” (Rubenstein & Goodenough, 1965)
- The content of a linguistic unit is determined by its **distribution** over a corpus (i.e., the other units appearing in its context)

◇ Computational version:

Word Embeddings

◇ Interpretability efforts on vector representations of mathematical expressions

- (e.g. d’Ascoli et al., 2022; Gao et al., 2017; Krstovski & Blei, 2018; Purgał et al., 2021; Ryskina & Knight, 2021)



(d’Ascoli et al., 2022)

1. 0 is a number.

$$0 \in \mathbb{N}$$

2. If n is a number, the successor of n is a number.

$$n \in \mathbb{N} \implies \text{succ}(n) \in \mathbb{N}$$

3. 0 is not the successor of a number.

$$\forall n \in \mathbb{N}, 0 \neq \text{succ}(n)$$

4. Two numbers of which the successors are equal are themselves equal.

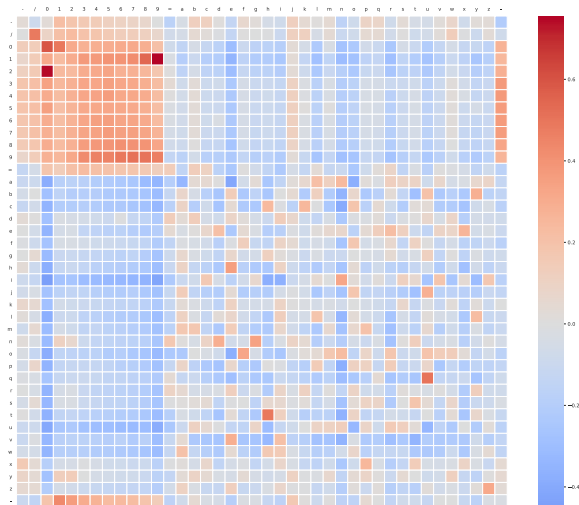
$$\forall n, m \in \mathbb{N}, \text{succ}(n) = \text{succ}(m) \implies n = m$$

5. If a set $\mathbf{S}$ of numbers contains 0 and also the successor of every number in $\mathbf{S}$, then every number is in $\mathbf{S}$ (induction axiom).

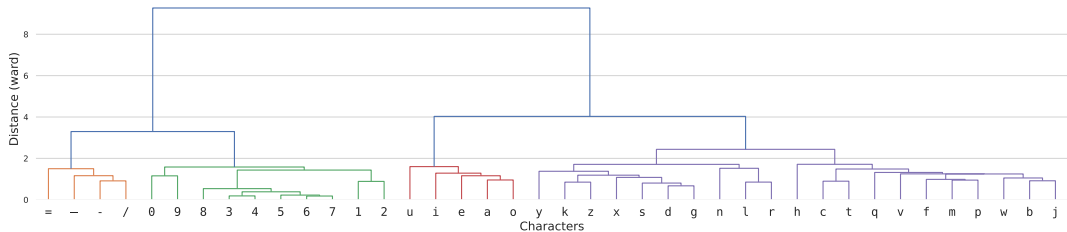
$$0 \in \mathbf{S} \wedge (\forall n, n \in \mathbf{S} \implies \text{succ}(n) \in \mathbf{S}) \implies \forall n \in \mathbb{N}, n \in \mathbf{S}$$

The Distributional Properties of Characters

$$A_{i,j} = pmi(c_i; c_j) = \log \frac{p(c_i, c_j)}{p(c_i)p(c_j)}$$



The Characteristic Content of Digits



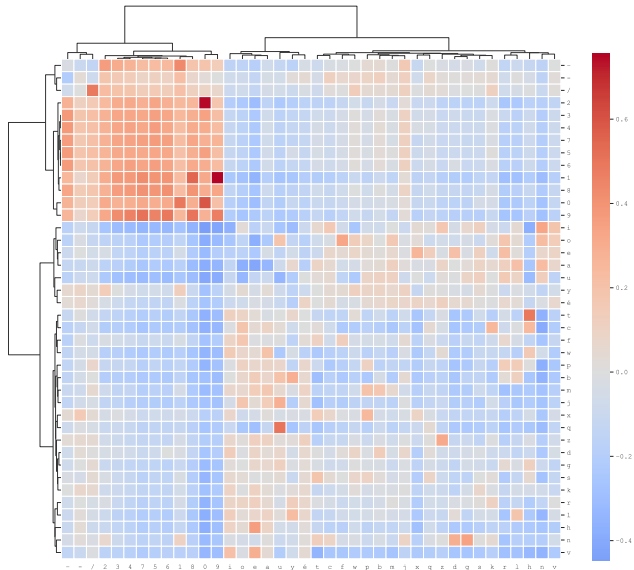
$O := \{=, -, -, /\}$

$D := \{0, 9, 8, 3, 4, 5, 6, 7, 1, 2\}$

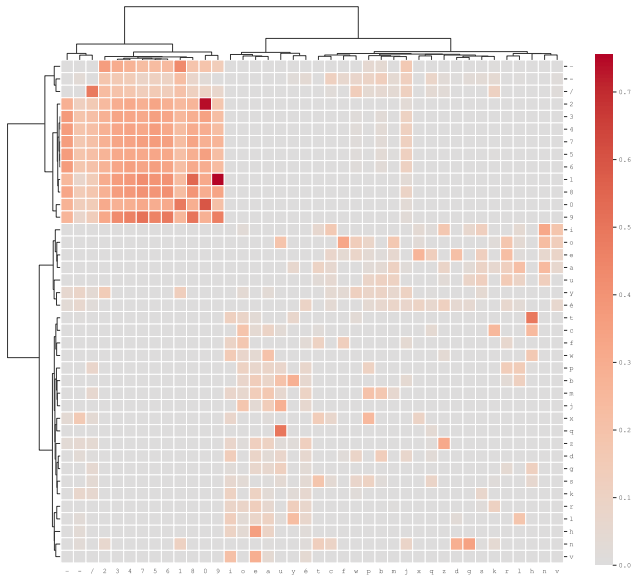
$V := \{u, i, e, a, o\}$

$C := \{y, k, z, x, s, d, g, n, l, r, h, c, t, q, v, f, m, p, w, b, j\}$

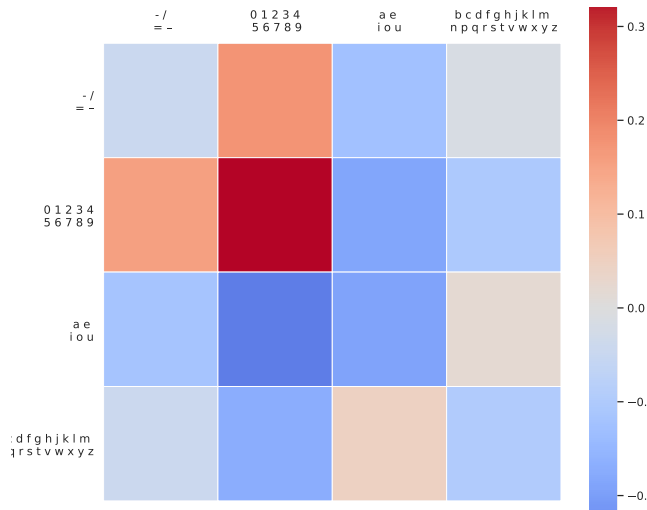
The Syntactic Content of Digits



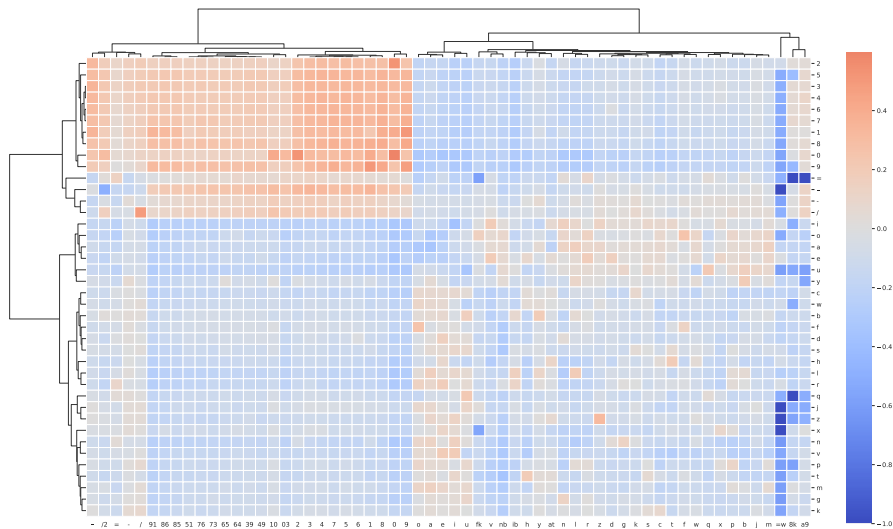
The Syntactic Content of Digits



The Syntactic Content of Digits



The Characteristic Content of Digit Pairs



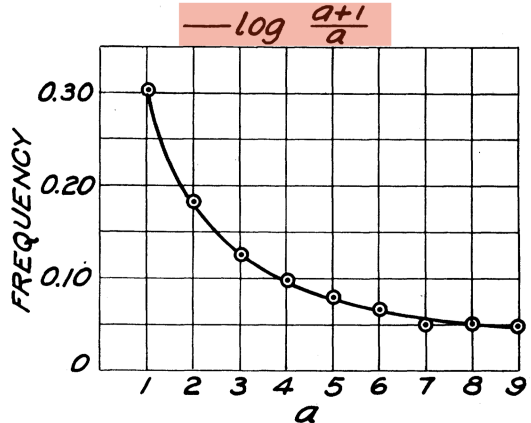
Benford's Law

TABLE I

PERCENTAGE OF TIMES THE NATURAL NUMBERS 1 TO 9 ARE USED AS FIRST DIGITS IN NUMBERS, AS DETERMINED BY 20,229 OBSERVATIONS

Group	Title	First Digit									Count
		1	2	3	4	5	6	7	8	9	
A	Rivers, Area	31.0	16.4	10.7	11.3	7.2	8.6	5.5	4.2	5.1	335
B	Population	33.9	20.4	14.2	8.1	7.2	6.2	4.1	3.7	2.2	3259
C	Constants	41.3	14.4	4.8	8.6	10.6	5.8	1.0	2.9	10.6	104
D	Newspapers	30.0	18.0	12.0	10.0	8.0	6.0	6.0	5.0	5.0	100
E	Spec. Heat	24.0	18.4	16.2	14.6	10.6	4.1	3.2	4.8	4.1	1389
F	Pressure	29.6	18.3	12.8	9.8	8.3	6.4	5.7	4.4	4.7	703
G	H.P. Lost	30.0	18.4	11.9	10.8	8.1	7.0	5.1	5.1	3.6	690
H	Mol. Wgt.	26.7	25.2	15.4	10.8	6.7	5.1	4.1	2.8	3.2	1800
I	Drainage	27.1	23.9	13.8	12.6	8.2	5.0	5.0	2.5	1.9	159
J	Atomic Wgt.	47.2	18.7	5.5	4.4	6.6	4.4	3.3	4.4	5.5	91
K	$n^{-1}, \sqrt{n}, \dots$	25.7	20.3	9.7	6.8	6.6	6.8	7.2	8.0	8.9	5000
L	Design	26.8	14.8	14.3	7.5	8.3	8.4	7.0	7.3	5.6	560
M	<i>Digest</i>	33.4	18.5	12.4	7.5	7.1	6.5	5.5	4.9	4.2	308
N	Cost Data	32.4	18.8	10.1	10.1	9.8	5.5	4.7	5.5	3.1	741
O	X-Ray Volts	27.9	17.5	14.4	9.0	8.1	7.4	5.1	5.8	4.8	707
P	Am. League	32.7	17.6	12.6	9.8	7.4	6.4	4.9	5.6	3.0	1458
Q	Black Body	31.0	17.3	14.1	8.7	6.6	7.0	5.2	4.7	5.4	1165
R	Addresses	28.9	19.2	12.6	8.8	8.5	6.4	5.6	5.0	5.0	342
S	$n^1, n^2 \dots n!$	25.3	16.0	12.0	10.0	8.5	8.8	6.8	7.1	5.5	900
T	Death Rate	27.0	18.6	15.7	9.4	6.7	6.5	7.2	4.8	4.1	418
Average		30.6	18.5	12.4	9.4	8.0	6.4	5.1	4.9	4.7	1011
Probable Error		± 0.8	± 0.4	± 0.4	± 0.3	± 0.2	± 0.2	± 0.2	± 0.2	± 0.3	—

**FREQUENCY
OF
FIRST PLACE DIGITS
OBSERVED**



Informational Content and Order

δ	$p(\delta)$	$n_{10}(\delta)$
1	0.301	1
2	0.176	2
3	0.125	3
4	0.097	4
5	0.079	5
6	0.067	6
7	0.058	7
8	0.051	8
9	0.046	9

$$D := \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 0\}$$

$$N := D^*$$

$$\delta \in D$$

$$\nu_1, \nu'_1, \nu_2, \nu'_2 \in N$$

$$n_b : N \rightarrow \mathbb{N}$$

$\nu \mapsto$ valeur numérique de ν en base b

$$p : D \rightarrow [0, 1]$$

$$\delta \mapsto \log \left(1 + \frac{1}{n_b(\delta)} \right)$$

Informational Content and Order

δ	$p(\delta)$	$n_{10}(\delta)$
1	0.301	1
2	0.176	2
3	0.125	3
4	0.097	4
5	0.079	5
6	0.067	6
7	0.058	7
8	0.051	8
9	0.046	9

$$D := \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 0\}$$

$$N := D^*$$

$$\delta \in D$$

$$\nu_1, \nu'_1, \nu_2, \nu'_2 \in N$$

$$n_b : N \rightarrow \mathbb{N}$$

$$\nu \mapsto \text{valeur numérique de } \nu \text{ en base } b$$

$$p : D \rightarrow [0, 1]$$

$$\delta \mapsto \log \left(1 + \frac{1}{n_b(\delta)} \right)$$

$$p(\nu\nu') \stackrel{?}{=} \mathcal{C}(p(\nu), p(\nu'))$$

$$n_b(\nu_1\nu'_1) \leq n_b(\nu_2\nu'_2) \iff \mathcal{C}(p(\nu_1), p(\nu'_1)) \leq \mathcal{C}(p(\nu_2), p(\nu'_2))$$

$$p(\delta) = \mathcal{B}(\delta) = \log_b \left(1 + \frac{1}{n_b(\delta)} \right)$$

$$p(\delta) = \mathcal{B}(\delta) = \log_b \left(1 + \frac{1}{n_b(\delta)} \right)$$
$$n_b(\delta) = \mathcal{B}^{-1}(\delta) = \frac{1}{b^{p(\delta)} - 1}$$

$$p(\delta) = \mathcal{B}(\delta) = \log_b \left(1 + \frac{1}{n_b(\delta)} \right)$$
$$n_b(\delta) = \mathcal{B}^{-1}(\delta) = \frac{1}{b^{p(\delta)} - 1}$$

$$\mathcal{B}(\nu\nu') = \log_b \left(1 + \frac{1}{n_b(\nu\nu')} \right)$$

$$p(\delta) = \mathcal{B}(\delta) = \log_b \left(1 + \frac{1}{n_b(\delta)} \right)$$

$$n_b(\delta) = \mathcal{B}^{-1}(\delta) = \frac{1}{b^{p(\delta)} - 1}$$

$$\begin{aligned} \mathcal{B}(\nu\nu') &= \log_b \left(1 + \frac{1}{n_b(\nu\nu')} \right) \\ &= \log_b \left(1 + \frac{1}{n_b(\nu) \cdot b^{|\nu'|} + n_b(\nu')} \right) \end{aligned}$$

$$p(\delta) = \mathcal{B}(\delta) = \log_b \left(1 + \frac{1}{n_b(\delta)} \right)$$

$$n_b(\delta) = \mathcal{B}^{-1}(\delta) = \frac{1}{b^{p(\delta)} - 1}$$

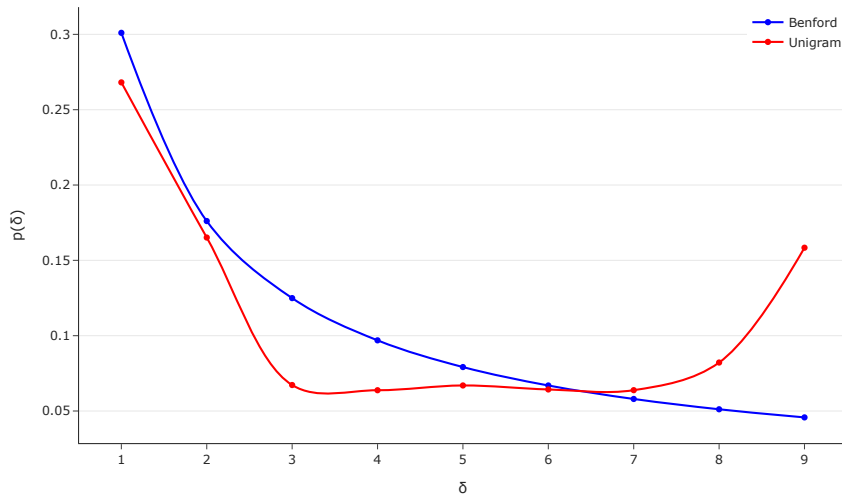
$$\begin{aligned} \mathcal{B}(\nu\nu') &= \log_b \left(1 + \frac{1}{n_b(\nu\nu')} \right) \\ &= \log_b \left(1 + \frac{1}{n_b(\nu) \cdot b^{|\nu'|} + n_b(\nu')} \right) \\ &= \log_b \left(1 + \frac{1}{\mathcal{B}^{-1}(\nu) \cdot b^{|\nu'|} + \mathcal{B}^{-1}(\nu')} \right) \\ &= \log_b \left(1 + \frac{1}{\frac{1}{b^{p(\nu)} - 1} \cdot b^{|\nu'|} + \frac{1}{b^{p(\nu')} - 1}} \right) \end{aligned} \quad \left(\text{with } |\nu| = \left\lfloor \log_b \left(\frac{1}{b^{p(\nu)} + 1} \right) \right\rfloor + 1 \right)$$

$$p(\delta) = \mathcal{B}(\delta) = \log_b \left(1 + \frac{1}{n_b(\delta)} \right)$$

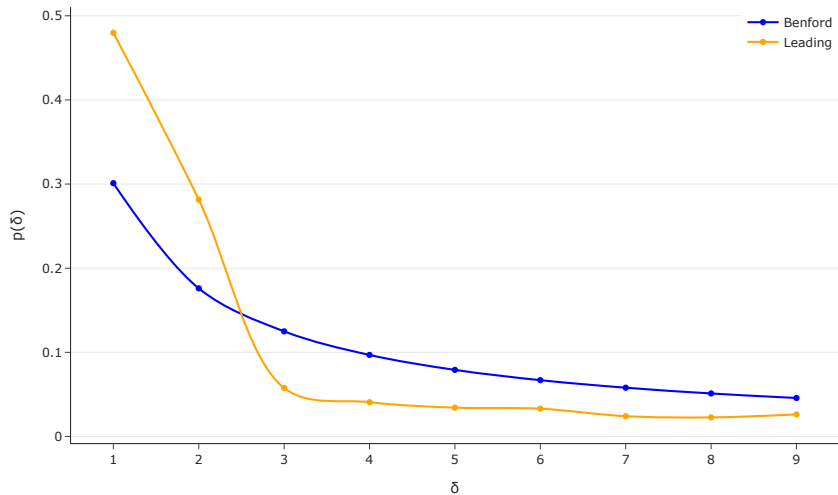
$$n_b(\delta) = \mathcal{B}^{-1}(\delta) = \frac{1}{b^{p(\delta)} - 1}$$

$$\begin{aligned}
 p(\nu\nu') &= \mathcal{B}(\nu\nu') = \log_b \left(1 + \frac{1}{n_b(\nu\nu')} \right) \\
 &= \log_b \left(1 + \frac{1}{n_b(\nu) \cdot b^{|\nu'|} + n_b(\nu')} \right) \\
 &= \log_b \left(1 + \frac{1}{\mathcal{B}^{-1}(\nu) \cdot b^{|\nu'|} + \mathcal{B}^{-1}(\nu')} \right) \\
 &= \log_b \left(1 + \frac{1}{\frac{1}{b^{p(\nu)} - 1} \cdot b^{|\nu'|} + \frac{1}{b^{p(\nu')} - 1}} \right) \quad \left(\text{with } |\nu| = \left\lfloor \log_b \left(\frac{1}{b^{p(\nu)} + 1} \right) \right\rfloor + 1 \right) \\
 &=: \mathcal{C}(p(\nu), p(\nu'))
 \end{aligned}$$

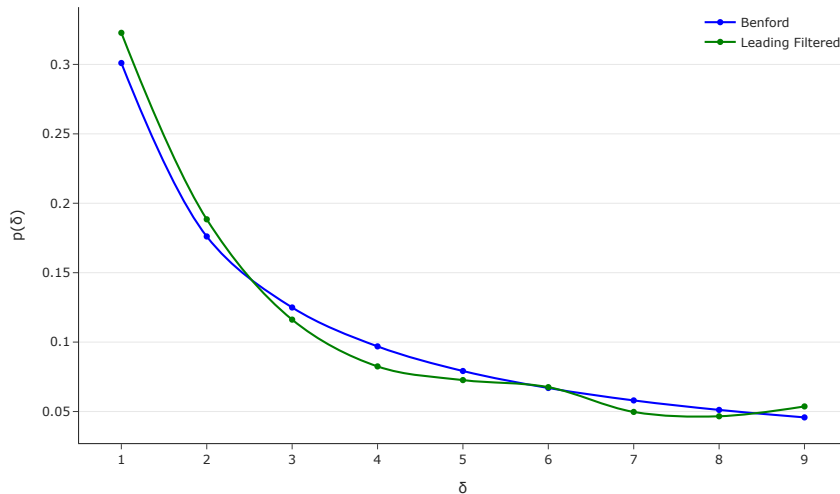
Benford in Data (δ)



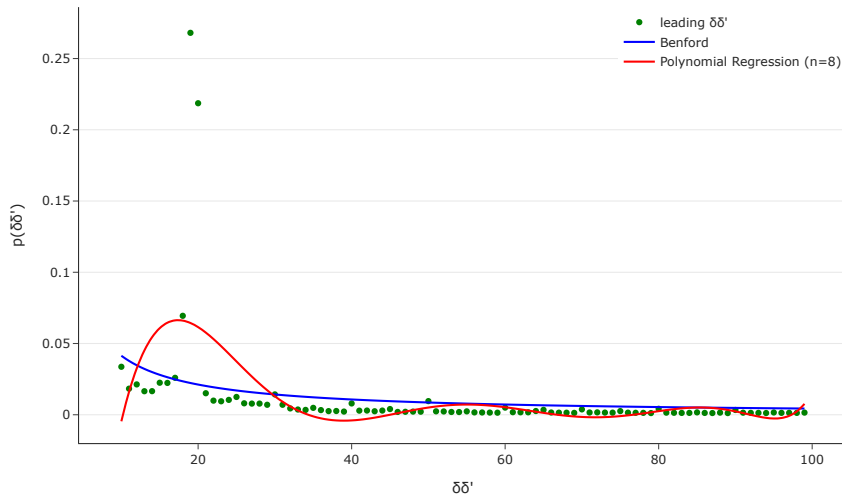
Benford in Data (δ)



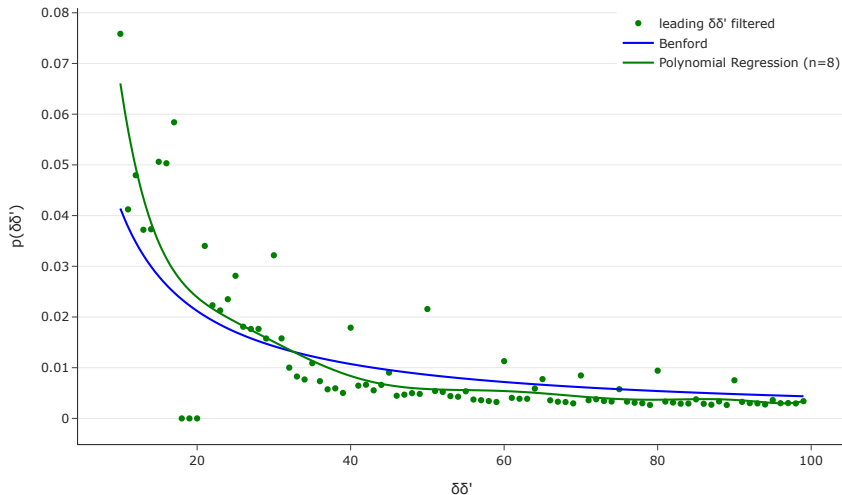
Benford in Data (δ)



Benford in Data ($\delta\delta'$)



Benford in Data ($\delta\delta'$)



Outline

Bring-Ins

Philosophical Perspectives

Historical Perspectives

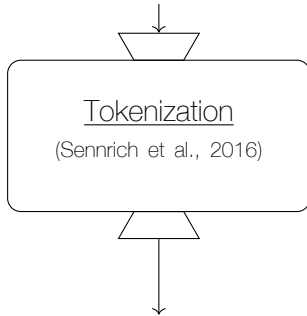
Epistemological Perspectives

Conceptual Perspectives

Theoretical Perspectives

Takeaways

Epistemology of Machine Learning
Distributional Language Models

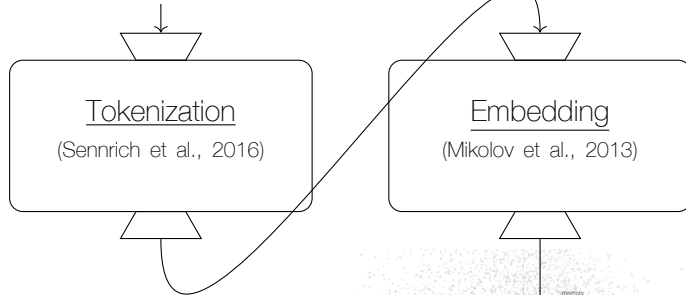


Epistemology of Machine Learning
Distributional Language Models

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

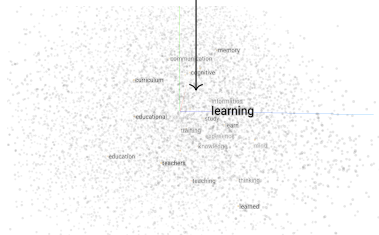
Embeddings in LLMs

Epistemology of Machine Learning
Distributional Language Models



Epistemology of Machine Learning
Distributional Language Models

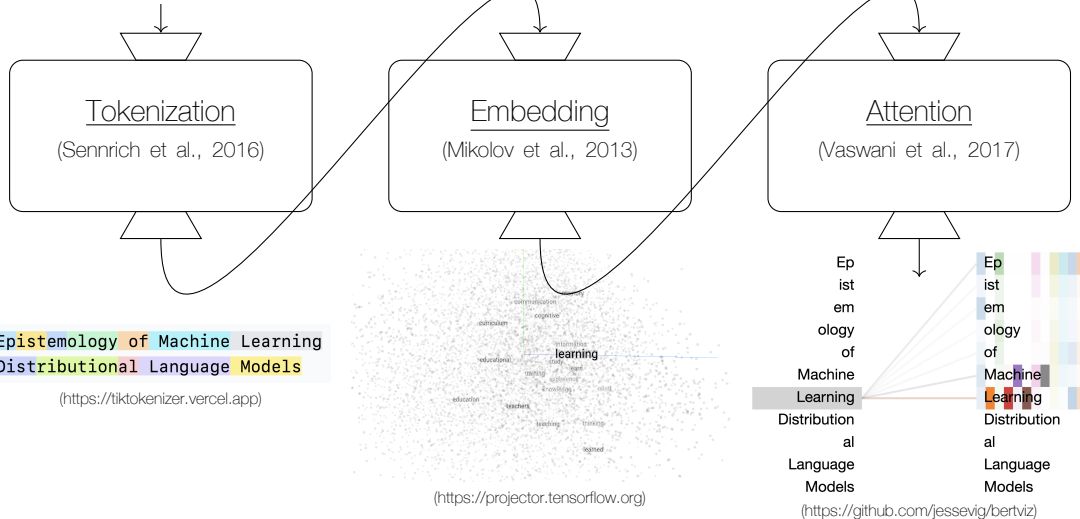
(<https://tiktokenizer.vercel.app>)



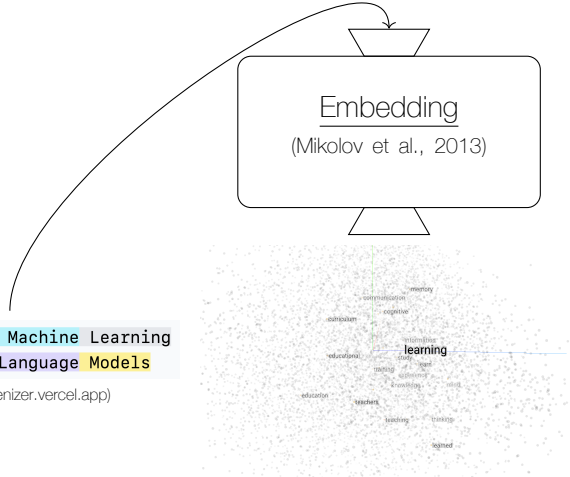
(<https://projector.tensorflow.org>)

Embeddings in LLMs

Epistemology of Machine Learning
Distributional Language Models



Embeddings in LLMs



Embedding
(Mikolov et al., 2013)

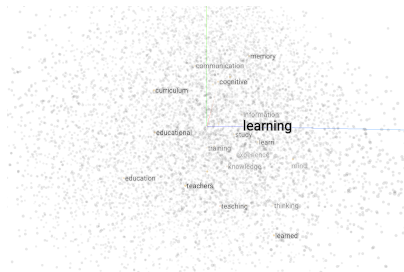
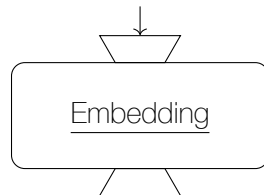
Epistemology of Machine Learning
Distributional Language Models

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

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

Formal Explainability

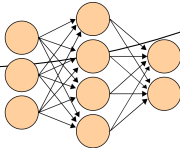
Epistemology of Machine Learning
Distributional Language Models



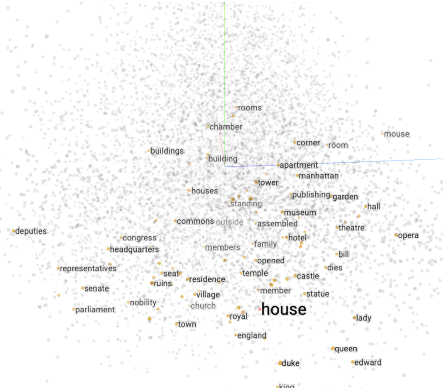
Epistemology of Machine Learning

Distributional Language Models

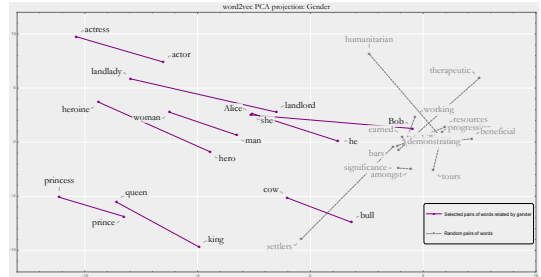
?

[illegible]

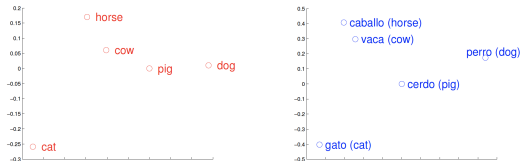
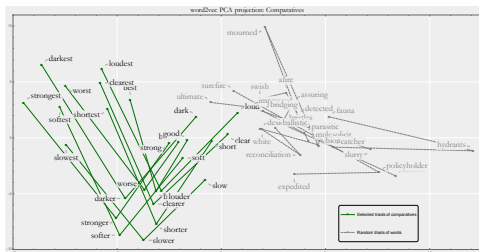
Embeddings: Similarity and Analogy



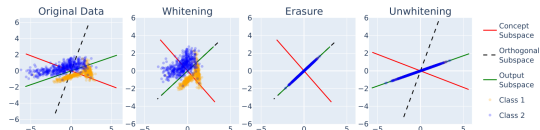
(<https://projector.tensorflow.org>)



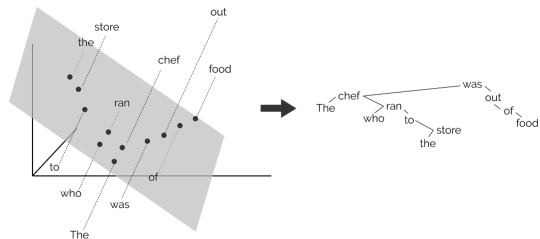
Embeddings: Other Applications



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



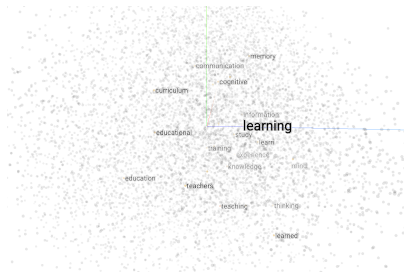
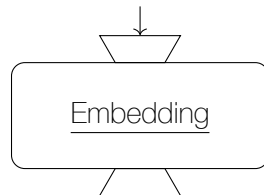
(Belrose et al., 2024)



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

Formal Explainability

Epistemology of Machine Learning
Distributional Language Models



Formal Explainability

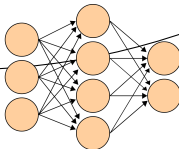
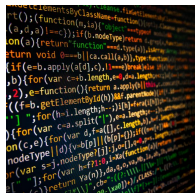
Structure

?

Epistemology of Machine Learning
Distributional Language Models

Embedding

Data



learning

education, teachers, teaching, thinking, knowledge, study, learn, curriculum, cognitive, memory, communication, educational, starting, information, mind, named

Formal Explainability

Structure



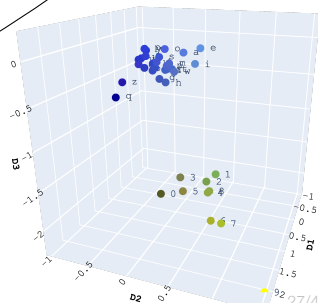
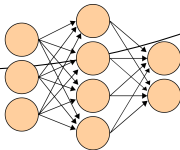
$\{-, /, 0, 1, 2, \dots, 8, 9, =,$
 $a, b, c, \dots, w, x, y, z, \acute{e}\}$



Embedding



Data



word2vec Explained

(Levy & 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 & 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.

Embedding structure

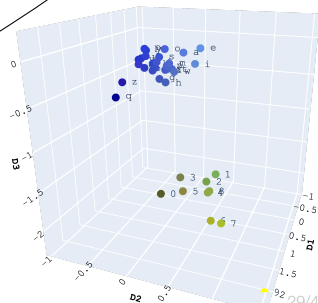
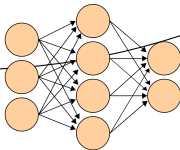
Structure

?

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

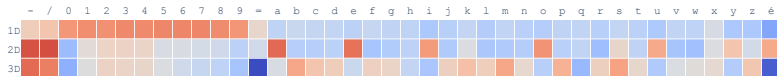
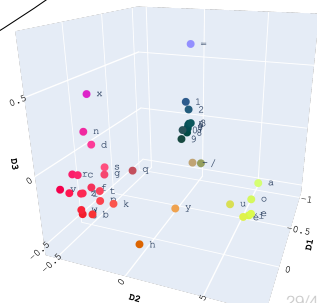
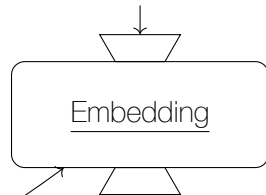
Embedding

Data

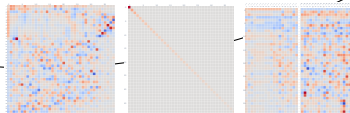


Embedding structure

Structure


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


Data



SVD

We need something more formal

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)

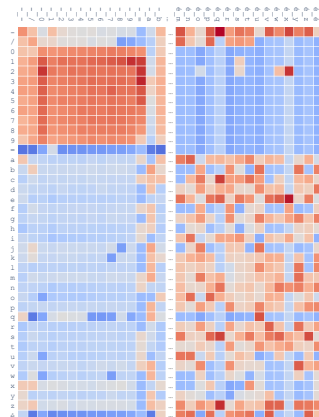
From matrices to distributional operators

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

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

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

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



From matrices to distributional operators

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

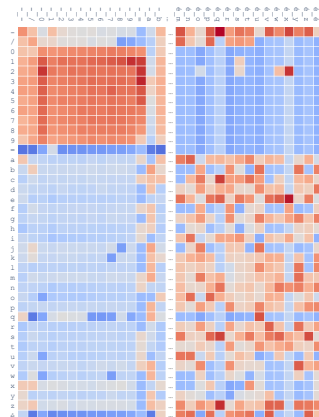
$$Y = X \times X = \{(-, -), (-, /), (-, 0), \dots, (\acute{e}, z), (\acute{e}, \acute{e})\}$$

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

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

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

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



From matrices to distributional operators

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

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

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

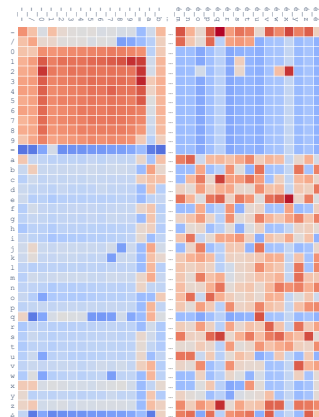
$$(\textcolor{red}{x}, \textcolor{blue}{y}) \mapsto \text{pmi}(\textcolor{red}{x}, \textcolor{blue}{y})$$

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

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

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

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



From matrices to distributional operators

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

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

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

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

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

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

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

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

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

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

From matrices to distributional operators

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

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

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

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

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

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

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

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

A commutative diagram with four nodes arranged in a square. The top-left node is $\textcolor{red}{X}$, the top-right node is $\mathbb{R}^{\textcolor{blue}{Y}}$, the bottom-left node is $\mathbb{R}^{\textcolor{red}{X}}$, and the bottom-right node is $\textcolor{blue}{Y}$. A horizontal arrow labeled M_x points from $\textcolor{red}{X}$ to $\mathbb{R}^{\textcolor{blue}{Y}}$. A horizontal arrow labeled M_y points from $\textcolor{blue}{Y}$ to $\mathbb{R}^{\textcolor{red}{X}}$. A vertical arrow points from $\textcolor{red}{X}$ down to $\mathbb{R}^{\textcolor{red}{X}}$. A vertical arrow points from $\textcolor{blue}{Y}$ up to $\mathbb{R}^{\textcolor{blue}{Y}}$.

From matrices to distributional operators

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

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

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

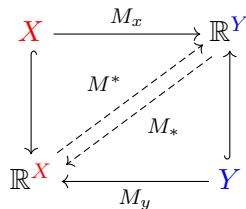
$$(\textcolor{red}{x}, \textcolor{blue}{y}) \mapsto \text{pmi}(\textcolor{red}{x}, \textcolor{blue}{y})$$

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

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

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

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

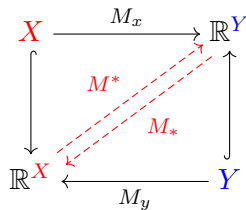


$$M^*: \mathbb{R}^{\textcolor{red}{X}} \rightarrow \mathbb{R}^{\textcolor{blue}{Y}}$$

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

From matrices to distributional operators

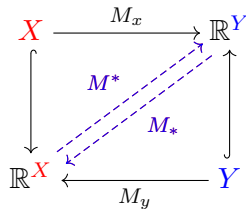
$$M_* M^* : \mathbb{R}^X \rightarrow \mathbb{R}^X$$



From matrices to distributional operators

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

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



From matrices to distributional operators

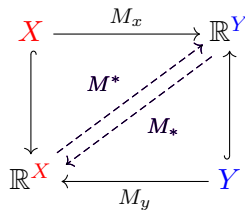
$$M_* M^* : \mathbb{R}^X \rightarrow \mathbb{R}^X$$

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

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

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

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



From matrices to distributional operators

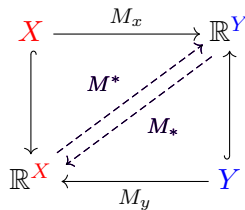
$$M_* M^* : \mathbb{R}^X \rightarrow \mathbb{R}^X$$

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

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

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

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



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

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

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

From matrices to distributional operators

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

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

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

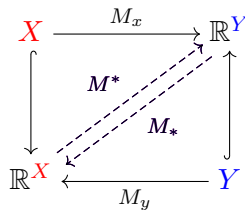
$$\{v_1, \dots, v_n\} \subset \mathbb{R}^Y$$

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

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

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

The u_i and v_i are (linear)
fixed points!



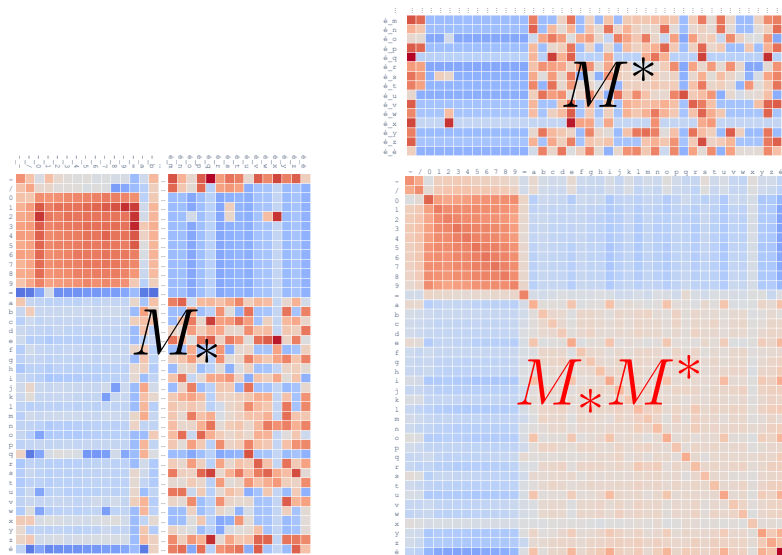
$$M = U \Sigma V^T$$

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

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

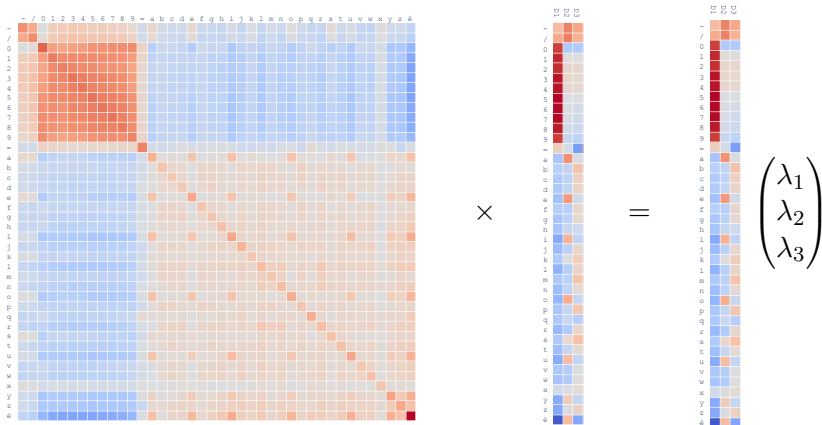
$$\Sigma := \begin{bmatrix} \sqrt{\lambda_1} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sqrt{\lambda_r} \end{bmatrix}$$

The operator M_*M^* is a covariance matrix

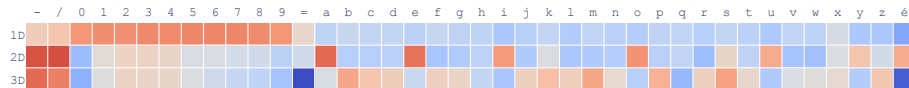


Eigenvectors as Fixed Points

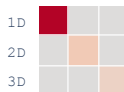
$$M_* M^* u = \lambda u$$



Eigenvectors of M_*M^* :



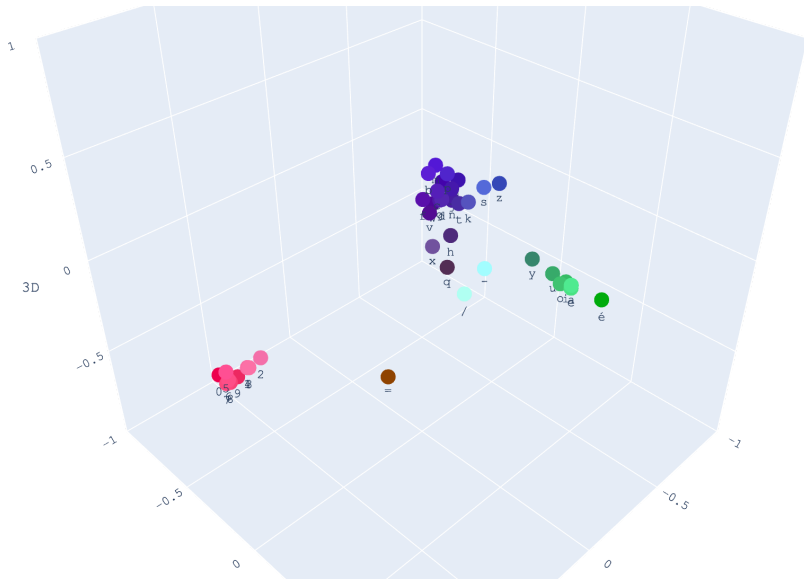
Eigenvalues of M_*M^* and M^*M_* :



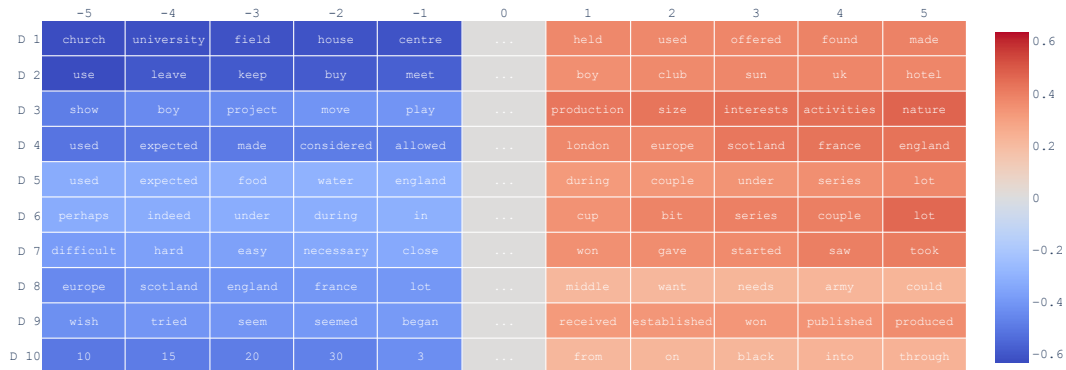
Eigenvectors of M^*M_* :



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



Words



Embeddings as Functors Over Categories

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

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

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

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

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

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

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

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

$$\begin{array}{ccc}
 \textcolor{red}{X} & \xrightarrow{M_x} & \mathbb{R}^{\textcolor{blue}{Y}} \\
 \downarrow & \nearrow M^* & \uparrow \\
 \mathbb{R}^{\textcolor{red}{X}} & \xleftarrow{M_y} & \textcolor{blue}{Y} \\
 & \nwarrow M_* &
 \end{array}$$

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

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

Embeddings as Functors Over Categories

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

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

Profunctor

$$\mathcal{M}: \mathbf{C}^{\text{op}} \times \mathbf{D} \rightarrow \mathbf{Set}$$

$$(c, d) \mapsto \mathcal{M}(c, d)$$

$$\mathcal{M}_c: \mathbf{C} \rightarrow (\mathbf{Set}^{\mathbf{D}})^{\text{op}}$$

$$c \mapsto \mathcal{M}(c, -)$$

$$\mathcal{M}_d: \mathbf{D} \rightarrow \mathbf{Set}^{\mathbf{C}^{\text{op}}}$$

$$d \mapsto \mathcal{M}(-, d)$$

$$\begin{array}{ccc}
 \mathbf{C} & \xrightarrow{\mathcal{M}_c} & (\mathbf{Set}^{\mathbf{D}})^{\text{op}} \\
 \text{Yoneda} \downarrow & \nearrow \mathcal{M}^* & \uparrow \text{Yoneda} \\
 \mathbf{Set}^{\mathbf{C}^{\text{op}}} & \xleftarrow{\mathcal{M}_d} & \mathbf{D}
 \end{array}$$

$\mathcal{M}_*$ (dashed arrow from $\mathbf{D}$ to $\mathbf{Set}^{\mathbf{C}^{\text{op}}}$)

$$\mathcal{M}^*: \mathbf{Set}^{\mathbf{C}^{\text{op}}} \rightarrow (\mathbf{Set}^{\mathbf{D}})^{\text{op}}$$

$$\mathcal{M}_*: (\mathbf{Set}^{\mathbf{D}})^{\text{op}} \rightarrow \mathbf{Set}^{\mathbf{C}^{\text{op}}}$$

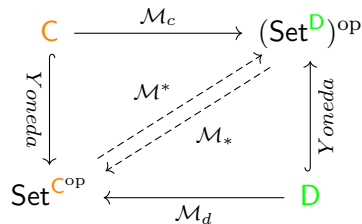
Embeddings as Functors Over Categories

Isbell Adjunction

$$\mathcal{M}^*: \mathbf{Set}^{\mathbf{C}^{\text{op}}} \rightleftarrows (\mathbf{Set}^{\mathbf{D}})^{\text{op}} : \mathcal{M}_*$$

$$\mathcal{M}_* \mathcal{M}^*: \mathbf{Set}^{\mathbf{C}^{\text{op}}} \rightarrow \mathbf{Set}^{\mathbf{C}^{\text{op}}}$$

$$\mathcal{M}^* \mathcal{M}_*: (\mathbf{Set}^{\mathbf{D}})^{\text{op}} \rightarrow (\mathbf{Set}^{\mathbf{D}})^{\text{op}}$$



Embeddings as Functors Over Categories

Isbell Adjunction

$$\mathcal{M}^*: \mathbf{Set}^{\mathbf{C}^{\text{op}}} \rightleftarrows (\mathbf{Set}^{\mathbf{D}})^{\text{op}} : \mathcal{M}_*$$

$$\mathcal{M}_* \mathcal{M}^*: \mathbf{Set}^{\mathbf{C}^{\text{op}}} \rightarrow \mathbf{Set}^{\mathbf{C}^{\text{op}}}$$

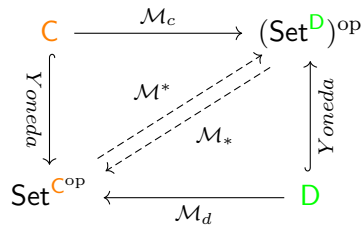
$$\mathcal{M}^* \mathcal{M}_*: (\mathbf{Set}^{\mathbf{D}})^{\text{op}} \rightarrow (\mathbf{Set}^{\mathbf{D}})^{\text{op}}$$

$$\text{Fix}(\mathcal{M}_* \mathcal{M}^*) := \{f \in \mathbf{Set}^{\mathbf{C}^{\text{op}}} \mid \mathcal{M}_* \mathcal{M}^*(f) \cong f\}$$

$$\text{Fix}(\mathcal{M}^* \mathcal{M}_*) := \{g \in (\mathbf{Set}^{\mathbf{D}})^{\text{op}} \mid \mathcal{M}^* \mathcal{M}_*(g) \cong g\}$$

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

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



Embeddings as Functors Over Categories

Isbell Adjunction

$$\mathcal{M}^*: \mathbf{Set}^{\mathbf{C}^{\text{op}}} \rightleftarrows (\mathbf{Set}^{\mathbf{D}})^{\text{op}} : \mathcal{M}_*$$

$$\mathcal{M}_* \mathcal{M}^*: \mathbf{Set}^{\mathbf{C}^{\text{op}}} \rightarrow \mathbf{Set}^{\mathbf{C}^{\text{op}}}$$

$$\mathcal{M}^* \mathcal{M}_*: (\mathbf{Set}^{\mathbf{D}})^{\text{op}} \rightarrow (\mathbf{Set}^{\mathbf{D}})^{\text{op}}$$

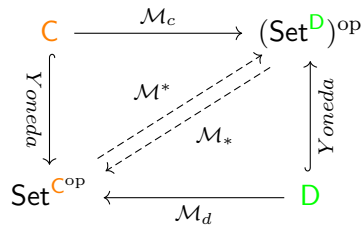
$$\text{Fix}(\mathcal{M}_* \mathcal{M}^*) := \{f \in \mathbf{Set}^{\mathbf{C}^{\text{op}}} \mid \mathcal{M}_* \mathcal{M}^*(f) \cong f\}$$

$$\text{Fix}(\mathcal{M}^* \mathcal{M}_*) := \{g \in (\mathbf{Set}^{\mathbf{D}})^{\text{op}} \mid \mathcal{M}^* \mathcal{M}_*(g) \cong g\}$$

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

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

The nucleus is a **category**
complete and **cocomplete**



Embeddings as Functors Over Categories

Isbell Adjunction

$$\mathcal{M}^*: \mathbf{Set}^{\mathbf{C}^{\text{op}}} \rightleftarrows (\mathbf{Set}^{\mathbf{D}})^{\text{op}} : \mathcal{M}_*$$

$$\mathcal{M}_* \mathcal{M}^*: \mathbf{Set}^{\mathbf{C}^{\text{op}}} \rightarrow \mathbf{Set}^{\mathbf{C}^{\text{op}}}$$

$$\mathcal{M}^* \mathcal{M}_*: (\mathbf{Set}^{\mathbf{D}})^{\text{op}} \rightarrow (\mathbf{Set}^{\mathbf{D}})^{\text{op}}$$

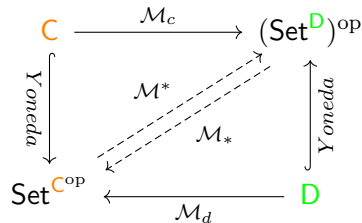
$$\text{Fix}(\mathcal{M}_* \mathcal{M}^*) := \{f \in \mathbf{Set}^{\mathbf{C}^{\text{op}}} \mid \mathcal{M}_* \mathcal{M}^*(f) \cong f\}$$

$$\text{Fix}(\mathcal{M}^* \mathcal{M}_*) := \{g \in (\mathbf{Set}^{\mathbf{D}})^{\text{op}} \mid \mathcal{M}^* \mathcal{M}_*(g) \cong g\}$$

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

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

The nucleus is a **category**
complete and **cocomplete**



Categories **C** and **D**
can be enriched!

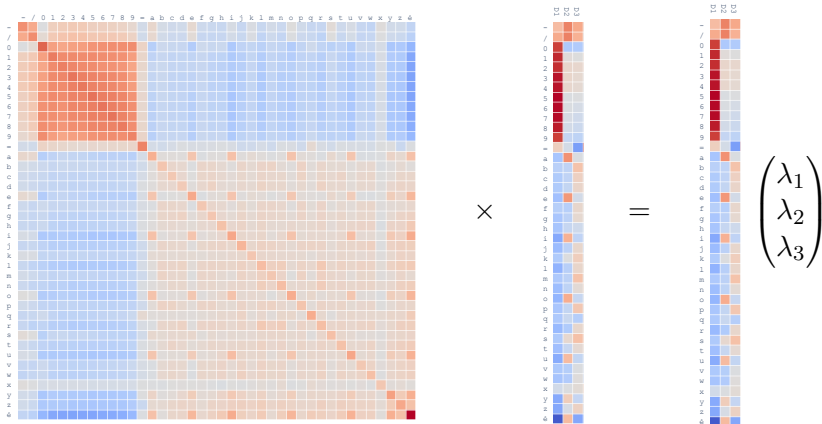
E.g.:

$$\mathcal{M}^*: \mathbf{2}^{\mathbf{C}^{\text{op}}} \rightleftarrows (\mathbf{2}^{\mathbf{D}})^{\text{op}} : \mathcal{M}_*$$

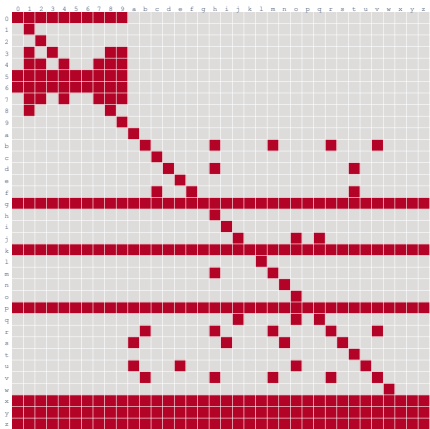
$$\mathcal{M}^*: \bar{\mathbf{R}}^{\mathbf{C}^{\text{op}}} \rightleftarrows (\bar{\mathbf{R}}^{\mathbf{D}})^{\text{op}} : \mathcal{M}_*$$

Binary Fixed Points

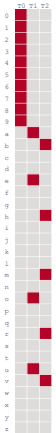
$$M_* M^* u = \lambda u$$



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



★



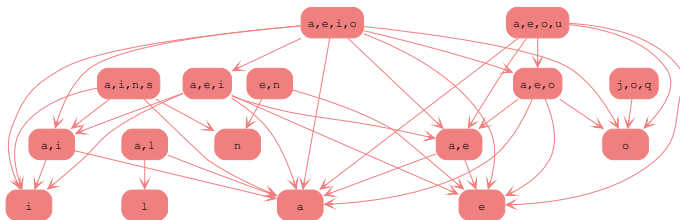
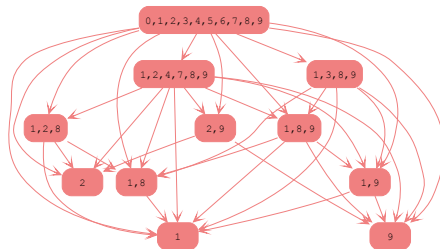
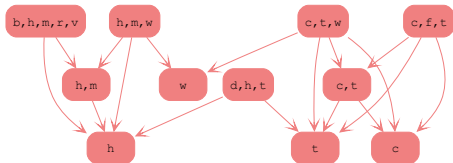
?



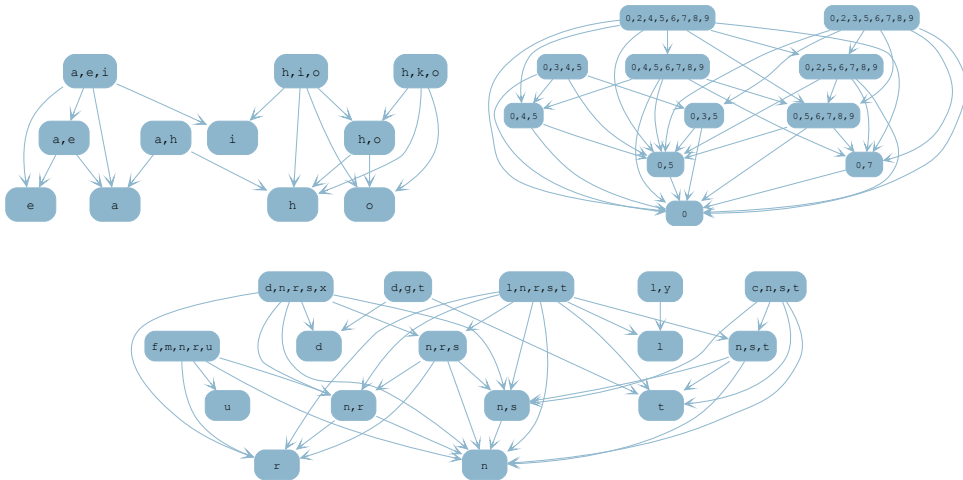
$$\mathcal{M}_* \mathcal{M}^* f = f$$

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				

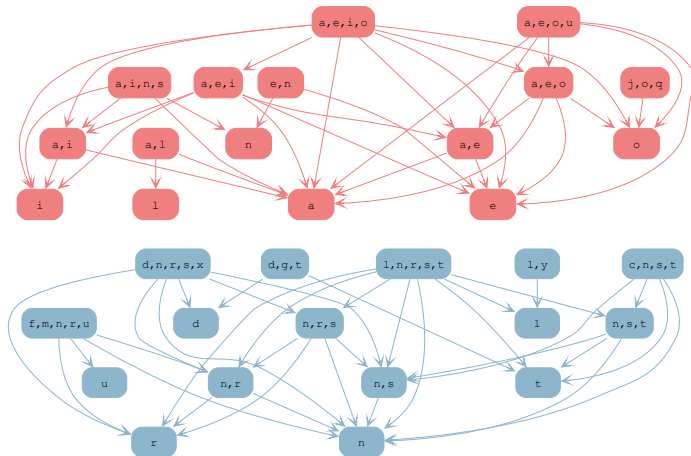
Partial Order Structure



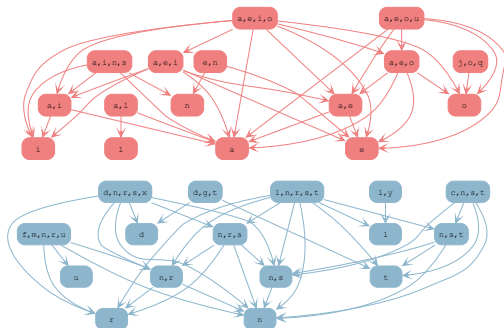
Dual Partial Order



Paring of Partial Ordered Fixed Points



Structure



$$\text{C}^{\text{op}} \times \text{D} \rightarrow 2$$

$$\Downarrow$$

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

Structure

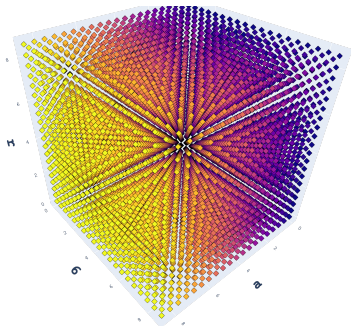
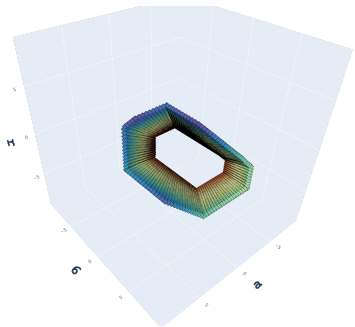
?

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

Structure

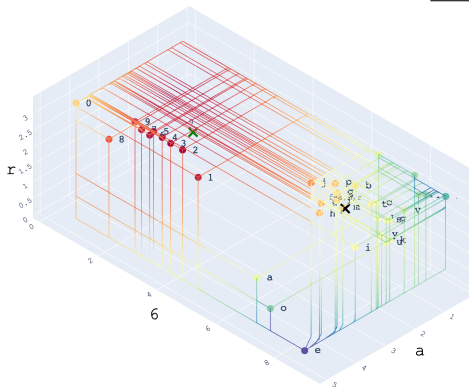
?

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



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

Structure

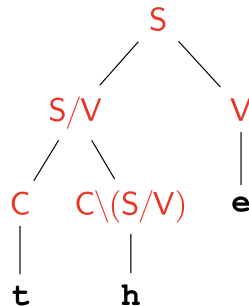
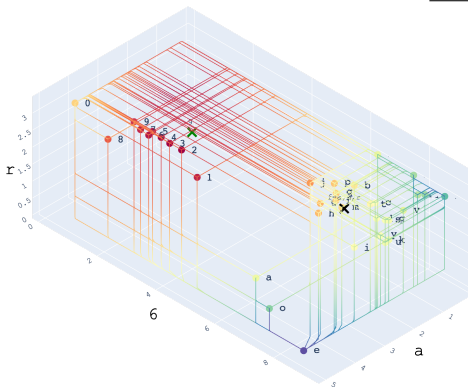


$$\mathbf{C}^{\text{op}} \times \mathbf{D} \rightarrow \bar{\mathbb{R}}$$

$$\Downarrow$$

$$\mathcal{M}^* : \bar{\mathbb{R}}^{\mathbf{C}^{\text{op}}} \rightleftarrows (\bar{\mathbb{R}}^{\mathbf{D}})^{\text{op}} : \mathcal{M}_*$$

Structure



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

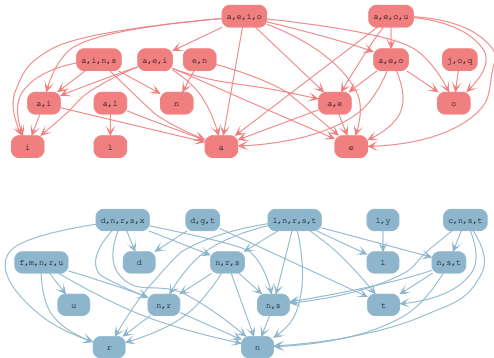
Definition (Polar/Orthogonal - Girard, 2011)

[G]iven a binary operation, noted $a, b \rightsquigarrow \langle a|b \rangle : A \times B \rightarrow C$ 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 $\text{Pol}(A) \subset \mathcal{P}(A)$ of *polar* sets, i.e., of the form $Y^\perp$, is closed under arbitrary intersections. In particular, A is polar and $X^{\perp\perp}$ is the smallest polar set containing X .
- ◇ As a consequence, $X^{\perp\perp\perp} = X^\perp$.



Theory of Computational Types

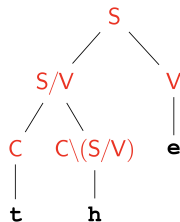
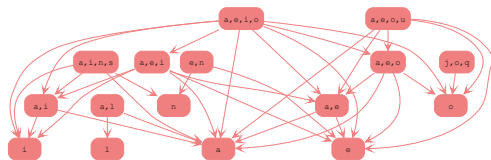
Definition (Polar/Orthogonal - Girard, 2011)

[G]iven a binary operation, noted $a, b \rightsquigarrow \langle a|b \rangle : A \times B \rightarrow C$ 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 $\text{Pol}(A) \subset \mathcal{P}(A)$ 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)

Outline

Bring-Ins

Philosophical Perspectives

Historical Perspectives

Epistemological Perspectives

Conceptual Perspectives

Theoretical Perspectives

Takeaways

- ◇ **Philosophical perspective**

The success of DNNs forces us to rethink the relationship between (natural) language and mathematics

- ◇ **Historical perspective**

After 150 years, the radical distinction between artificial and natural languages might become ineffective

- ◇ **Epistemological perspective**

Current empiricist approaches are ill-suited for this task

- ◇ **Conceptual perspective**

We need to develop distributional foundations for formal objects

- ◇ **Theoretical perspective**

“Geometric Structuralist Typing” may provide a solid framework to advance in this direction

Collaborations



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

Reference Papers

- ◇ Gastaldi, J. L. (2024). How to Do Maths with Words: Neural Machine Learning Applications to Mathematics and Their Philosophical Significance. In B. Sriraman (Ed.), *Handbook of the history and philosophy of mathematical practice* (pp. 3191–3226). Springer International Publishing. https://doi.org/10.1007/978-3-031-40846-5_142
- ◇ 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>
- ◇ 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>

References I

- Alemi, A. A., Chollet, F., Een, N., Irving, G., Szegedy, C., & Urban, J. (2016). Deepmath - deep sequence models for premise selection. *Proceedings of the 30th International Conference on Neural Information Processing Systems*, 2243–2251.
- Bansal, K., Loos, S. M., Rabe, M. N., Szegedy, C., & Wilcox, S. (2019). Holist: An environment for machine learning of higher-order theorem proving (extended version). *CoRR*, *abs/1904.03241*. <http://arxiv.org/abs/1904.03241>
- 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*.
- Blechs Schmidt, J., & Ernst, O. G. (2021). Three ways to solve partial differential equations with neural networks — a review. *GAMM-Mitteilungen*, 44(2), e202100006. <https://doi.org/https://doi.org/10.1002/gamm.202100006>
- 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>
- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., ... Amodei, D. (2020). Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, & H. Lin (Eds.), *Advances in neural information processing systems* (pp. 1877–1901, Vol. 33). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfc4967418bfb8ac142f64a-Paper.pdf
- Charton, F. (2021). Linear algebra with transformers. *CoRR*, *abs/2112.01898*. <https://arxiv.org/abs/2112.01898>
- d’Ascoli, S., Kamienny, P., Lample, G., & Charton, F. (2022). Deep symbolic regression for recurrent sequences. *CoRR*, *abs/2201.04600*.

References II

- Davies, A., Veličković, P., Buesing, L., Blackwell, S., Zheng, D., Tomašev, N., Tanburn, R., Battaglia, P., Blundell, C., Juhász, A., Lackenby, M., Williamson, G., Hassabis, D., & Kohli, P. (2021). Advancing mathematics by guiding human intuition with AI. *Nature*, 600(7887), 70–74. <https://doi.org/10.1038/s41586-021-04086-x>
- Firth, J. R. (1935). The technique of semantics. *Transactions of the Philological Society*, 34(1), 36–73. <https://doi.org/10.1111/j.1467-968X.1935.tb01254.x>
- Gao, L., Jiang, Z., Yin, Y., Yuan, K., Yan, Z., & Tang, Z. (2017). Preliminary exploration of formula embedding for mathematical information retrieval: Can mathematical formulae be embedded like a natural language? *CoRR*, *abs/1707.05154*. <http://arxiv.org/abs/1707.05154>
- Gastaldi, J. L. (2024). How to Do Maths with Words: Neural Machine Learning Applications to Mathematics and Their Philosophical Significance. In B. Sriraman (Ed.), *Handbook of the history and philosophy of mathematical practice* (pp. 3191–3226). Springer International Publishing. https://doi.org/10.1007/978-3-031-40846-5_142
- 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.
- Goldberg, Y., & Levy, O. (2014). Word2vec explained: Deriving mikolov et al.'s negative-sampling word-embedding method. *CoRR*, *abs/1402.3722*.
- Harris, Z. (1960). *Structural linguistics*. University of Chicago Press.
- Krstovski, K., & Blei, D. M. (2018). Equation embeddings.
- Lample, G., & Charton, F. (2019). Deep learning for symbolic mathematics.

References III

- 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.
- Lewkowycz, A., Andreassen, A., Dohan, D., Dyer, E., Michalewski, H., Ramasesh, V., Slone, A., Anil, C., Schlag, I., Gutman-Solo, T., Wu, Y., Neyshabur, B., Gur-Ari, G., & Misra, V. (2022). Solving quantitative reasoning problems with language models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, & A. Oh (Eds.), *Advances in neural information processing systems* (pp. 3843–3857, Vol. 35). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2022/file/18abbeef8cfe9203fdf9053c9c4fe191-Paper-Conference.pdf
- Li, Z., Kovachki, N. B., Azizzadenesheli, K., Liu, B., Bhattacharya, K., Stuart, A., & Anandkumar, A. (2021). Fourier neural operator for parametric partial differential equations. *International Conference on Learning Representations*. <https://openreview.net/forum?id=c8P9NQVtmnO>
- 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.
- Peng, S., Yuan, K., Gao, L., & Tang, Z. (2021). Mathbert: A pre-trained model for mathematical formula understanding. *CoRR*, abs/2105.00377. <https://arxiv.org/abs/2105.00377>
- Polu, S., & Sutskever, I. (2020). Generative language modeling for automated theorem proving. *CoRR*, abs/2009.03393. <https://arxiv.org/abs/2009.03393>
- Purgat, S., Parsert, J., & Kaliszyk, C. (2021). A study of continuous vector representations for theorem proving. *Journal of Logic and Computation*, 31(8), 2057–2083. <https://doi.org/10.1093/logcom/exab006>

References IV

- Ryskina, M., & Knight, K. (2021). Learning mathematical properties of integers. *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, 389–395. <https://doi.org/10.18653/v1/2021.blackboxnlp-1.30>
- Saussure, F. d. (1959). *Course in general linguistics* [Translated by Wade Baskin]. McGraw-Hill.
- 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.
- Shen, J. T., Yamashita, M., Prihar, E., Heffernan, N. T., Wu, X., & Lee, D. (2021). Mathbert: A pre-trained language model for general NLP tasks in mathematics education. *CoRR*, abs/2106.07340.
- 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
- Wagner, A. Z. (2021). Constructions in combinatorics via neural networks.
- Wu, Y., Jiang, A. Q., Li, W., Rabe, M. N., Staats, C. E., Jamnik, M., & Szegedy, C. (2022). Autoformalization with large language models. In A. H. Oh, A. Agarwal, D. Belgrave, & K. Cho (Eds.), *Advances in neural information processing systems*. <https://openreview.net/forum?id=IUikebJ1Bf0>

MALINCA
Institut Henri Poincaré
Paris, France

Language and Mathematics in the Time of AI
Philosophical and Theoretical Perspectives

Juan Luis Gastaldi
www.giannigastaldi.com

ETH zürich

October 1, 2025