Chat Token Vector
Università Ca' Foscari
Venice, Italy

Toward a Critical Formalism

Philosophical and Theoretical Effects of a Mathematical Critique of LLMs

Juan Luis Gastaldi

www.giannigastaldi.com

ETH zürich

June 12, 2025

Outline

Intro: Critique and Formalism

Epistemological Critique: LLMs as Formal Objects

Theoretical Critique: Formal Explainability
- The Algebra Behind the Embeddings
- The Structure Behind the Algebra
- The Categories Behind the Structure

Conclusion

Outline

Intro: Critique and Formalism

Epistemological Critique: LLMs as Formal Objects

Theoretical Critique: Formal Explainability
- The Algebra Behind the Embeddings
- The Structure Behind the Algebra
- The Categories Behind the Structure

Conclusion

Where Art Thou, Critique?

Where Art Thou, Critique?

- Good “externalist” critique

Where Art Thou, Critique?

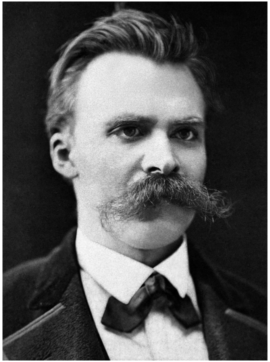
- Good “externalist” critique
- Poor “internalist” critique

- Good “externalist” critique
- Poor “internalist” critique
  - The main “critical” reference remains the “Stochastic Parrots” approach (Bender & Koller, 2020; Bender et al., 2021)

Where Art Thou, Critique?

- Good “externalist” critique
- Poor “internalist” critique
  - The main “critical” reference remains the “Stochastic Parrots” approach (Bender & Koller, 2020; Bender et al., 2021)
  - Kirschenbaum (2023):
    Bender et al.’s (2021) paper “offers a disarmingly linear account of how language, communication, intention, and meaning work, one that would seem to sidestep decades of scholarship around these same issues in literary theory [...] the passage would be red meat for a graduate critical-theory seminar.”
  - Underwood (2023):
    “The beautiful irony of this situation [...] is that a generation of humanists trained on Foucault have now rallied around “On the Dangers of Stochastic Parrots” to oppose a theory of language that their own disciplines invented, just at the moment when computer scientists are reluctantly beginning to accept it.”

The Birth of Contemporary Critique



"In some remote corner of the universe, flickering in the light of the countless solar systems into which it had been poured, there was once a planet on which clever animals invented cognition. It was the most arrogant and most mendacious minute in the 'history of the world'..."

"On Truth and Lying in a Non-Moral Sense"
(Nietzsche, 1873)

Knowledge depends on language

The Critical Argumentative Matrix

Knowledge depends on language

$$\downarrow$$

The relation between language and the world is essentially arbitrary

The Critical Argumentative Matrix

Knowledge depends on language

↓

The relation between language and the world is essentially arbitrary

↓

Any regularity in language/knowledge is not natural but cultural/social/political

The Critical Argumentative Matrix

Knowledge depends on language

↓

The relation between language and the world is essentially arbitrary

↓

Any regularity in language/knowledge is not natural but cultural/social/political

↓

We should resist existing regularities and create new ones

The Critical Argumentative Matrix

Knowledge depends on language
(Epistemological)

↓

The relation between language and the world is essentially arbitrary

↓

Any regularity in language/knowledge is not natural but cultural/social/political
(Political)

↓

We should resist existing regularities and create new ones
(Aesthetic)

The Critical Argumentative Matrix

Knowledge depends on language
(Epistemological)

[The relation between language and the world is essentially arbitrary?]

Any regularity in language/knowledge is not natural but cultural/social/political
(Political)

↓

We should resist existing regularities and create new ones
(Aesthetic)

- At the source of this situation is the new foundational role played by formal sciences in the 20th century
  - For a theory of language: Carnap, Gödel, Turing, Shannon, Harris, Chomsky...

- At the source of this situation is the new foundational role played by formal sciences in the 20th century
  - For a theory of language: Carnap, Gödel, Turing, Shannon, Harris, Chomsky...
- The critical tradition has either withdrawn from the areas conquered by formal approaches, or made formal approaches the target of criticism

Critique and Formalism

- At the source of this situation is the new foundational role played by formal sciences in the 20th century
    - For a theory of language: Carnap, Gödel, Turing, Shannon, Harris, Chomsky...
- The critical tradition has either withdrawn from the areas conquered by formal approaches, or made formal approaches the target of criticism
- We need a new strategy: Elaborate a critical formalism

- In the case of AI, a critical formalism can provide:
    - New epistemological tools countering dogmatic perspectives stemming from within the field
    - New theoretical tools contributing to the non-dogmatic positive production of knowledge

Outline

Intro: Critique and Formalism

Epistemological Critique: LLMs as Formal Objects

Theoretical Critique: Formal Explainability

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

Conclusion

Neural LMs as Computable Functions

Neural LM

↓

?

Neural LMs as Computable Functions

Neural LM

 ?

Neural LMs as Computable Functions

Neural LM

↓

🤖 ❓

Neural LMs as Computable Functions

Neural LM

Neural LMs as Computable Functions

Neural LM

Neural LMs as Computable Functions

Neural LM

↓

🗺️ ❓

Neural LMs as Computable Functions

$$\begin{array}{c}\text{Neural LM}\\ \downarrow \\ f\ \textcolor{red}{!}\end{array}$$

Neural LMs as Computable Functions

$$\begin{array}{c}\text{Neural LM}\\ \Big\downarrow\\ \text{Function}\end{array}$$

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

Neural LMs as Computable Functions

$$\text{Neural LM} \quad \big\downarrow \quad \textcolor{red}{\text{Computable}} \ \text{Function}$$

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

Neural LMs as Computable Functions

Neural LM

$\downarrow$

Computable

Function

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

$\downarrow$

Turing

Machine

(Turing, 1937)

Neural LMs as Computable Functions

Neural LM

Computable
Function

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

Recursive
Function
(Gödel, 1934)

Turing
Machine
(Turing, 1937)

Lambda
Term
(Church, 1936)

Neural LMs as Computable Functions

Neural LM

$\downarrow$

Computable
Function

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

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

Neural LMs as Computable Functions

Neural LM

Computable
Function

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

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

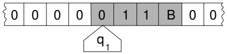
Neural LMs as Computable Functions

Neural LM

Computable Function

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

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



credit: Nynexman4464

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

$$yxz$$

λ-abstraction and β-reduction in λ-calculus

$$\lambda \color{red}{x} \color{black}{.y} \color{red}{x} \color{black}{z}$$

$$(\lambda {\color{red}x}.y{\color{red}x}z){\color{blue}t}$$

λ-abstraction and β-reduction in λ-calculus

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

$$\rightsquigarrow$$

$$ytz$$

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

Empirical Evaluation

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

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

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

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

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

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

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

$\ldots$

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

Empirical Evaluation

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

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

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

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

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

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

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

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

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

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

$\dots$

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

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

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

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

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

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

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

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

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

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

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

$\ldots$

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

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

Interpretability

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

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

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

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

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

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

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

$\dots$

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

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

$$\rightsquigarrow$$

$$\rightsquigarrow$$

$$\rightsquigarrow$$

$$\rightsquigarrow$$

$$\rightsquigarrow$$

$$\rightsquigarrow$$

$$\rightsquigarrow$$

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

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

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

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

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

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

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

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

$\dots$

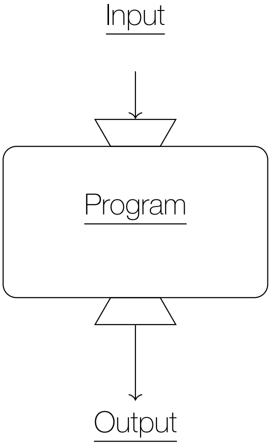
n: $\lambda f.\lambda x.\underbrace{f(\dots(f}_{n \text{ times}}\,x)\dots)$

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

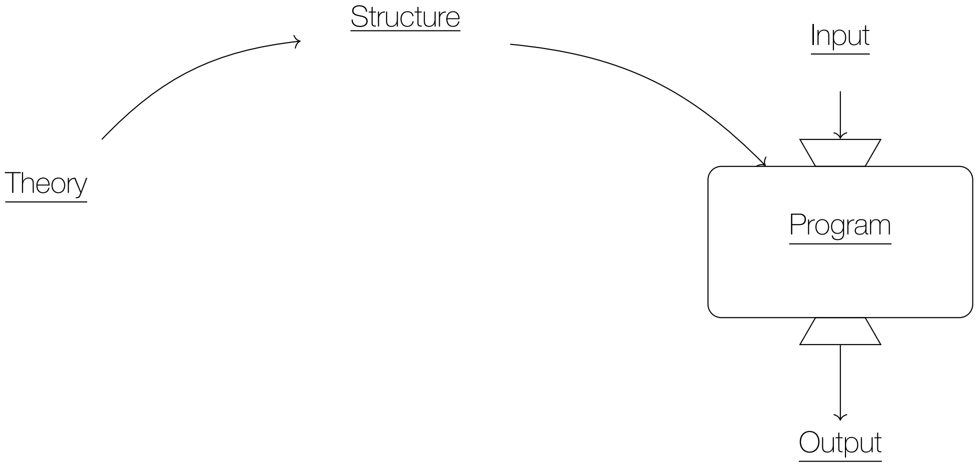
$P := \lambda m.\lambda n.\lambda f.\lambda x.mf(nfx)$

P'' :=λRófÄÒêÑ5È|Àxñ=∞ù ÿmWf286ëy'SÒú>v&ìÂ ¬2 óÉ7öç∞{ã>2flB°µG#À9ÇU
∞btYBô¥ Ù ë‰3;5 å[l–èuô¨Ü ˚ 7–Ù.λ:ˆ4mÓØ¥´è—+ÌsÖ,$+gï„B™÷o–#ïŸê Ûv
–gÓÿ/ëiijO‡Œfi • J1«€ø¸Ï hãêt‡æY$ˆ6 FïW»RÙKg¢¨.λ‡d¯… D2÷¸ò ˚ xêÈy. Ó”¢b
Bé£NÈ1Ê‡/Û9Ñµ–/JYÇõË9ÿÀÈ.λÁÍ Àˆö Ç,»fq∞±îˇB5Ì>O˜g™“6Ωe“æëC/ã… Ö
· fÓ Å]ÑåyÊN°Ê ›¨.λÆà€fUòfEÙ˙Í m#„¸4\r√¬÷Îpò»y*vtÄJÃF1ûÁóz«ñM”DjŒ
BËèÍT _ Êa‰ÁÇΩ @\Øˆ˜]Î›h‡.:ˆ4mÓØ¥´è—+ÌsÖ,$+gï„B™÷o–#ïŸê Ûv–gÓÿ
/ëiijO‡Œfi • J1«€ø¸Ï hãêt‡æY$ˆ6 FïW»RÙKg¢¨ÁÍ Àˆö Ç,»fq∞±îˇB5Ì>O˜g™“6
Ωe“æëC/ã… Ö · fÓ Å]ÑåyÊN°Ê ›¨(‡d¯… D2÷¸ò ˚ xêÈy. Ó”¢bBé£NÈ1Ê‡/Û9Ñµ–/
JYÇõË9ÿÀÈÁÍ Àˆö Ç,»fq∞±îˇB5Ì>O˜g™“6Ωe“æëC/ã… Ö · fÓ Å]ÑåyÊN°Ê ›¨Æ
à€fUòfEÙ˙Í m#„¸4\r√¬÷Îpò»y*vtÄJÃF1ûÁóz«ñM”DjŒBËèÍT _ Êa‰ÁÇΩ @\
Øˆ˜]Î›h‡)(ËÎÛeí4WµÍ }w„$Ω“K5èÃ¶%3[m ’ˇBÃfifÒ; óJçCËîõŸØcB,ñ$Ãå}ÓÅØ3¡
ˇ?ö¯oŒ@ fl8˘R CÆo˘* &<Ÿ ¬o12À‰ àÓÜ#ï ˝,ú¨ «ô¸„∞Îâä“øÃd|ˆÑ’Ê yØ;ˆW
›˝wò[}\»ÖÈúw’6<ù˘=ãÕ ¯Ì‘Dz?2±|é‘ 3Â/rxµ∞µ$ÀëÂ*lf˜îû’+ÍVìyªGæßägô/,uÑ)

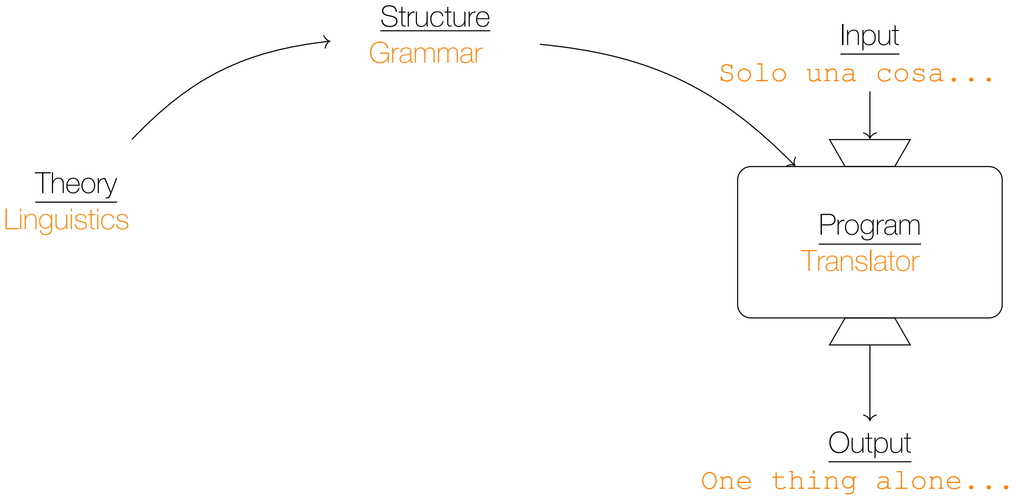
Making It Explicit



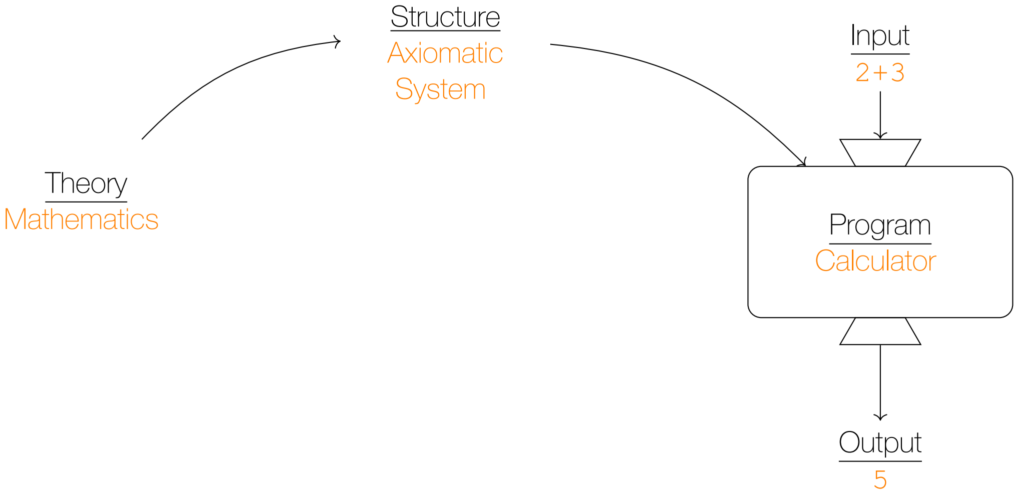
Making It Explicit



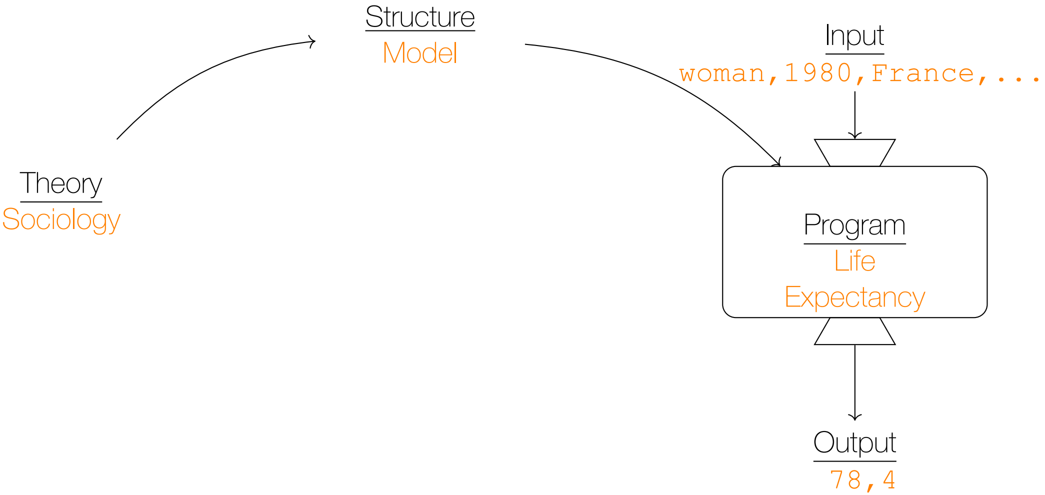
Making It Explicit



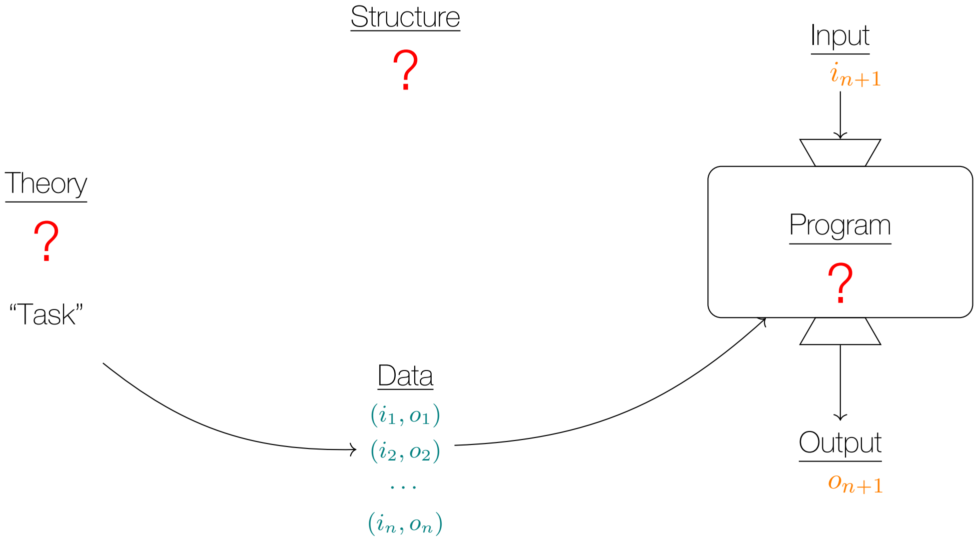
Making It Explicit



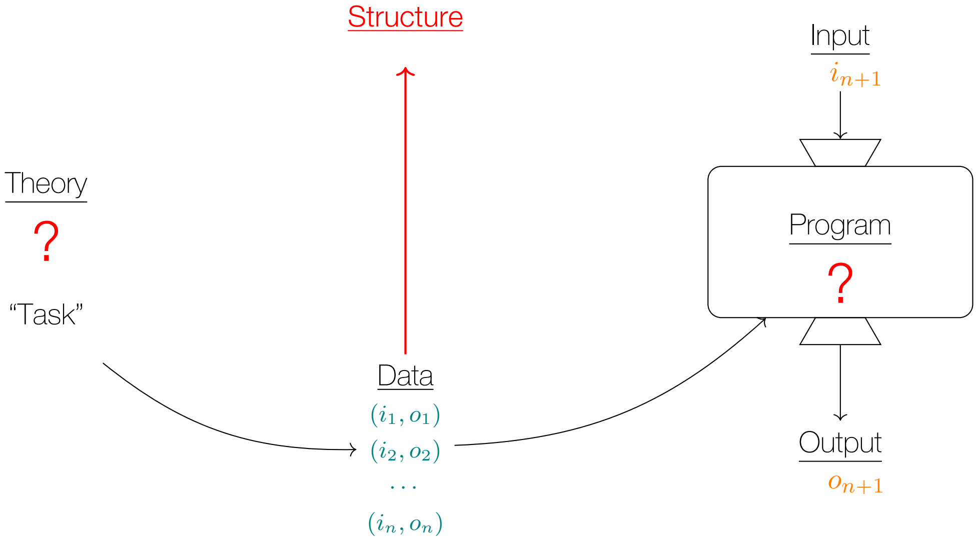
Making It Explicit



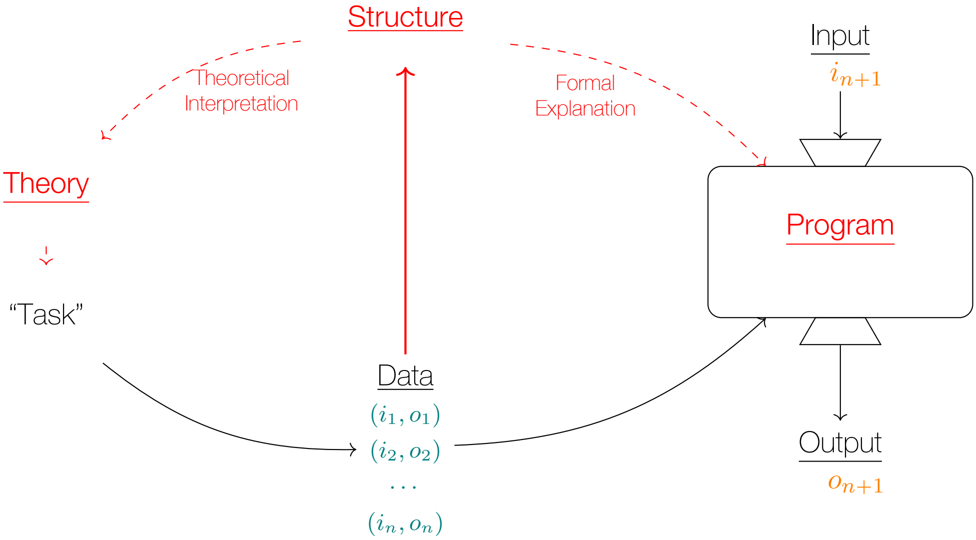
Making It Explicit



Making It Explicit



Making It Explicit



Outline

Intro: Critique and Formalism

Epistemological Critique: LLMs as Formal Objects

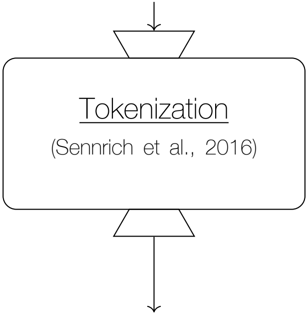
Theoretical Critique: Formal Explainability

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

Conclusion

Formal Explainability

Epistemology of Machine Learning
Distributional Language Models



`Epistemology of Machine Learning`
`Distributional Language Models`

(https://tiktokenizer.vercel.app)

Formal Explainability

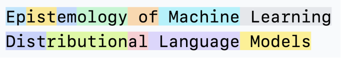
Epistemology of Machine Learning
Distributional Language Models

<u>Tokenization</u>

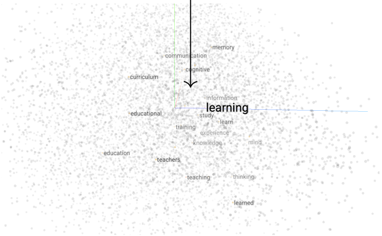
(Sennrich et al., 2016)

<u>Embedding</u>

(Mikolov et al., 2013)



(https://tiktokenizer.vercel.app)



(https://projector.tensorflow.org)

Formal Explainability

`Epistemology of Machine Learning Distributional Language Models`

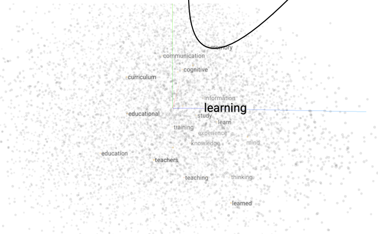
Tokenization
(Sennrich et al., 2016)

Embedding
(Mikolov et al., 2013)

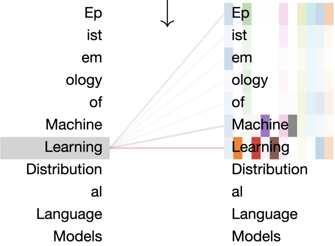
Attention
(Vaswani et al., 2017)

`Epistemology of Machine Learning Distributional Language Models`

(https://tiktokenizer.vercel.app)



(https://projector.tensorflow.org)



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

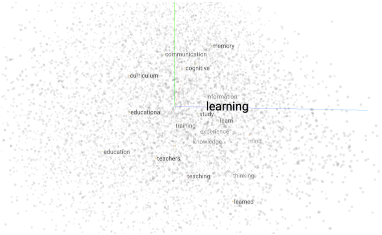
Formal Explainability

Embedding

(Mikolov et al., 2013)

Epistemology of Machine Learning
Distributional Language Models

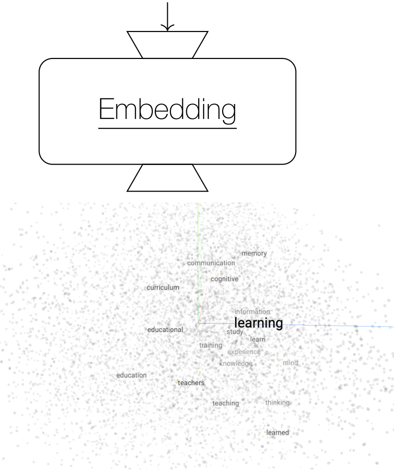
(https://tiktokenizer.vercel.app)



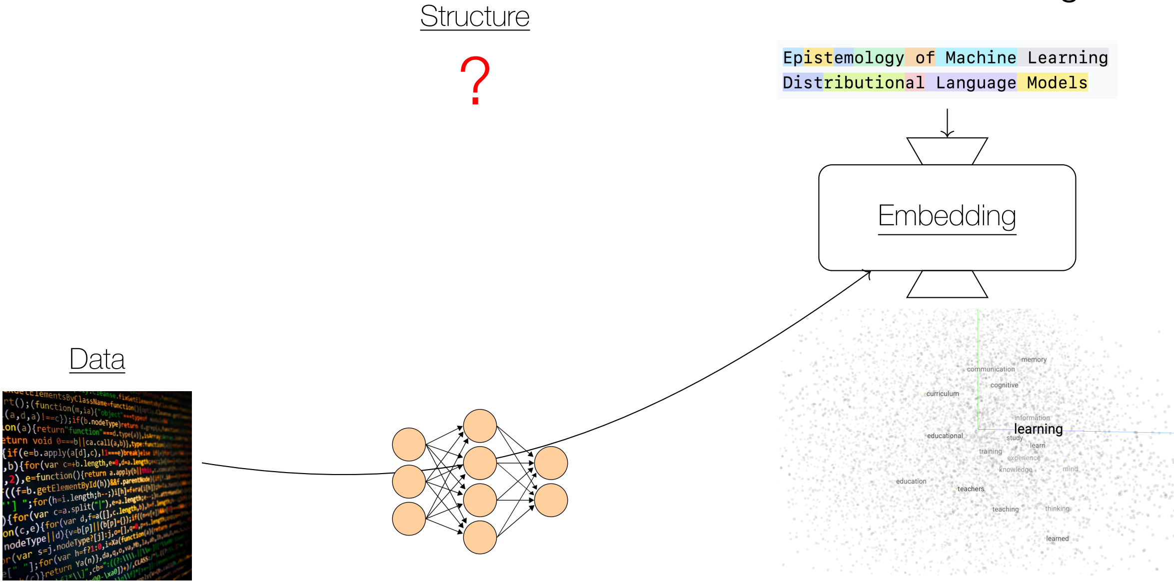
(https://projector.tensorflow.org)

The Structure of Embeddings

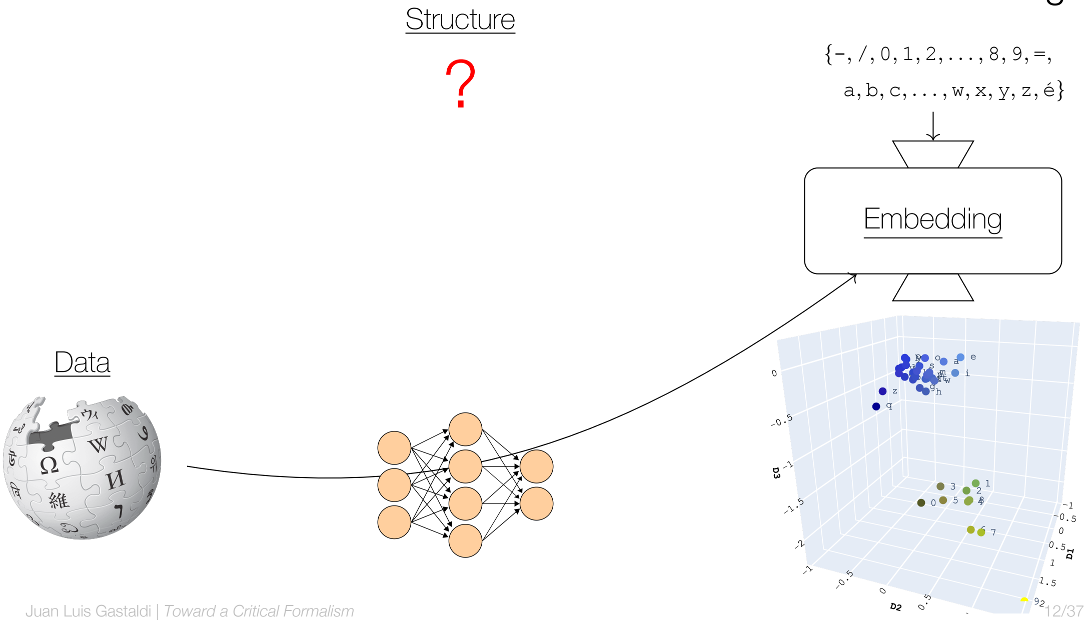
Epistemology of Machine Learning
Distributional Language Models



The Structure of Embeddings



The Structure of Embeddings



Outline

Intro: Critique and Formalism

Epistemological Critique: LLMs as Formal Objects

Theoretical Critique: Formal Explainability

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

Conclusion

word2vec Explained

(Levy and Goldberg, 2014)

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

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

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

word2vec Explained
(Levy and Goldberg, 2014)

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

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

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

word2vec Explained

(Levy and Goldberg, 2014)

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

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

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

word2vec Explained

(Levy and Goldberg, 2014)

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

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

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

word2vec Explained

(Levy and Goldberg, 2014)

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

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

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

word2vec Explained
(Levy and Goldberg, 2014)

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

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

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

word2vec Explained

(Levy and Goldberg, 2014)

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

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

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

word2vec Explained

(Levy and Goldberg, 2014)

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

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

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

- The Singular Value Decomposition (SVD) provides an exact solution to this problem.

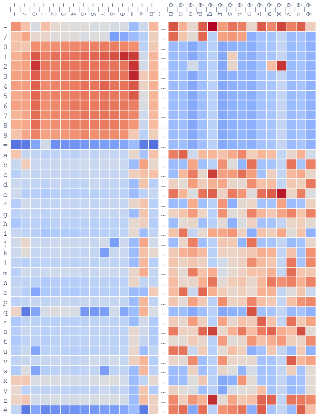
Example: Characters in Wikipedia

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

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

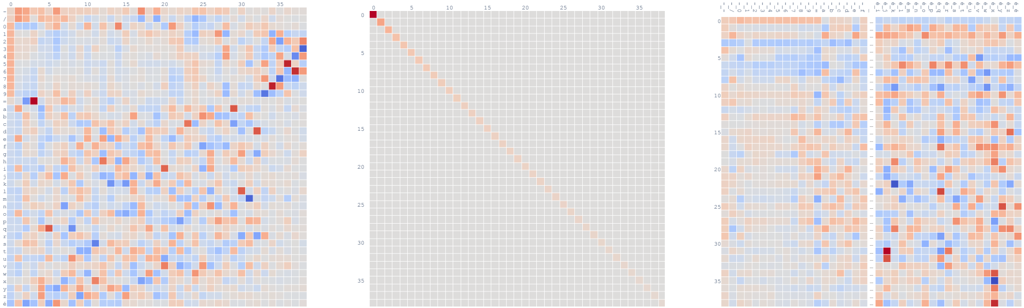
$$M_{\color{red}{w}\color{blue}{c}} = \mathrm{pmi}(\color{red}{w}, \color{blue}{c})$$

$$= \log \frac{p(\color{red}{w}, \color{blue}{c})}{p(\color{red}{w})p(\color{blue}{c})}$$



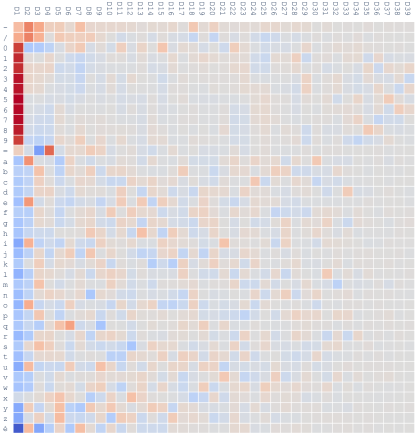
SVD of Wikipedia Character PMI Matrix

$U \qquad \Sigma \qquad V^T$



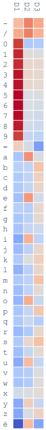
Truncate

$$U \times \Sigma$$



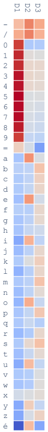
Truncate

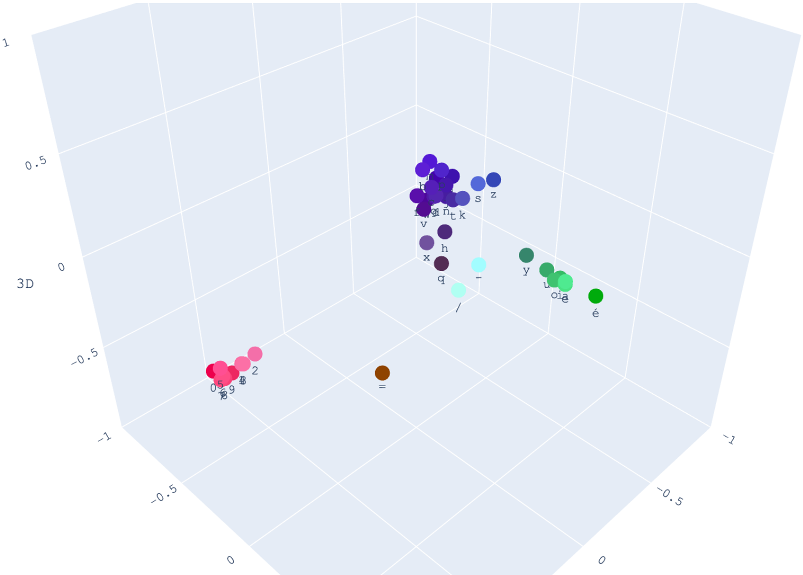
$$\hat{U} \times \hat{\Sigma}$$



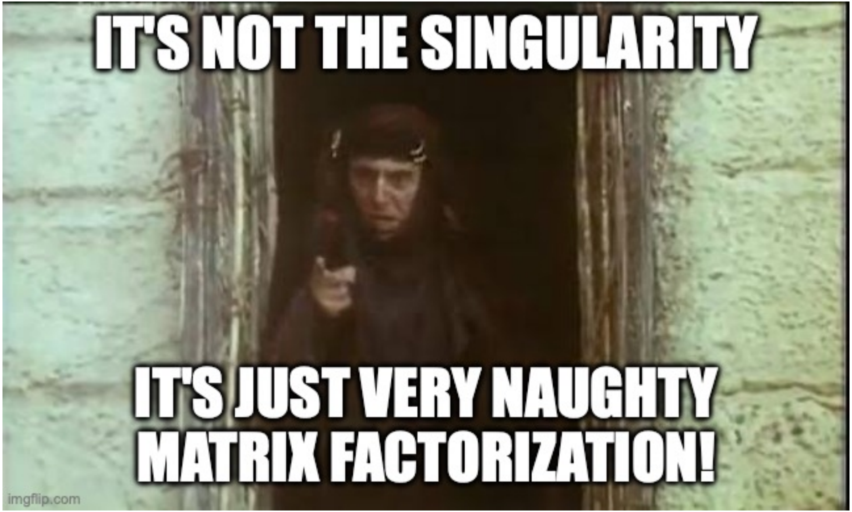
Plot

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

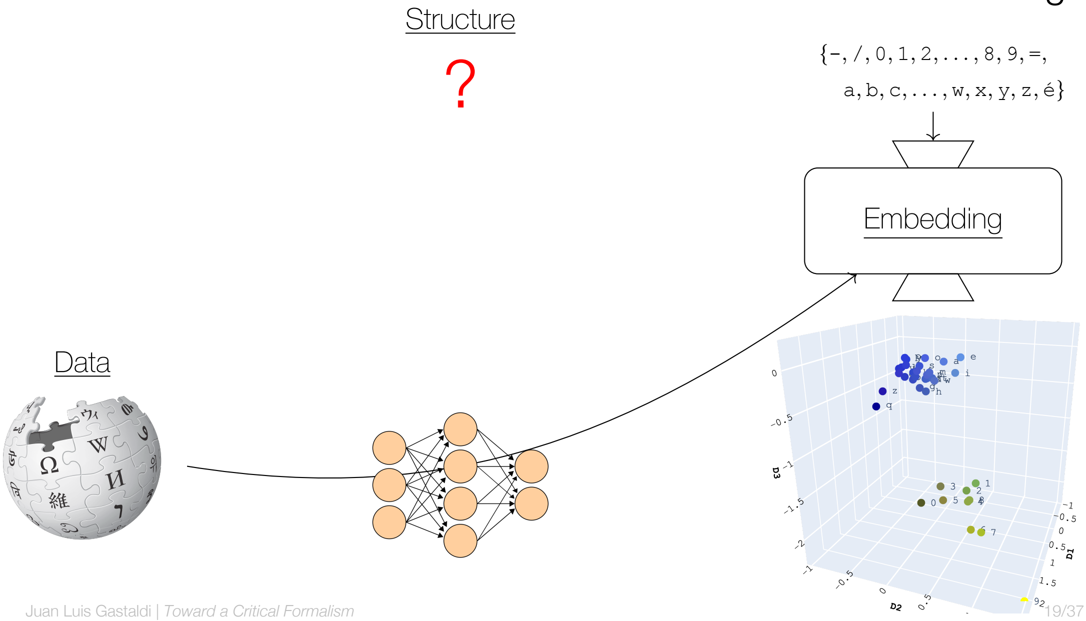




What to conclude?

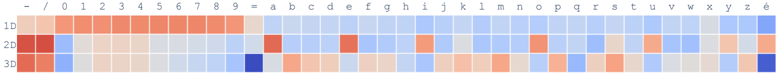


The Structure of Embeddings

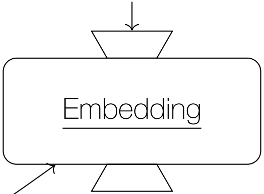


The Structure of Embeddings

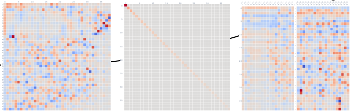
Structure



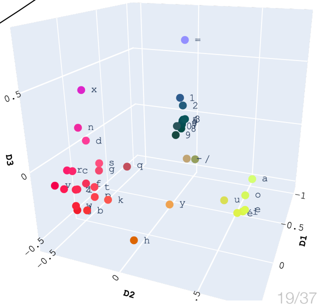
$\{-, /, 0, 1, 2, \ldots, 8, 9, =, a, b, c, \ldots, w, x, y, z, é\}$



Data



SVD



4 Why does this produce good word representations?

Good question. We don't really know.

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

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

(Goldberg and Levy, 2014)

Outline

Intro: Critique and Formalism

Epistemological Critique: LLMs as Formal Objects

Theoretical Critique: Formal Explainability

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

Conclusion

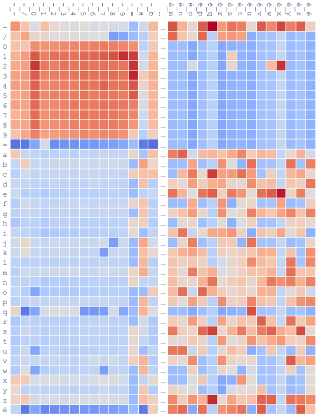
Embeddings as Functions Over Sets

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

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

$$M \colon \color{red}{X} \color{black}{} \times \color{blue}{Y} \color{black}{} \to \mathbb{R}$$

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



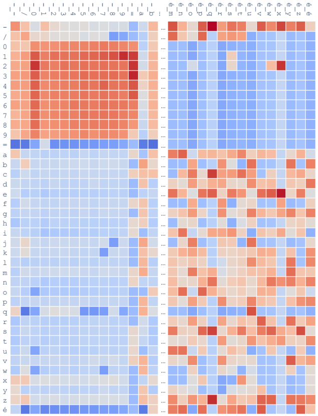
Embeddings as Functions Over Sets

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

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

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

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



Embeddings as Functions Over Sets

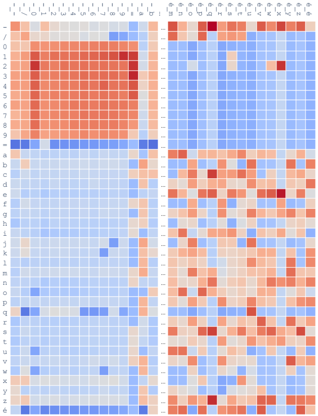
$$\color{red}{X} = \{\texttt{-}, \texttt{/}, \texttt{0}, \texttt{1}, \texttt{2}, \texttt{3}, \texttt{4}, \texttt{5}, \texttt{6}, \texttt{7}, \texttt{8}, \texttt{9}, \texttt{=}, \texttt{a}, \texttt{b}, \texttt{c}, \ldots, \texttt{w}, \texttt{x}, \texttt{y}, \texttt{z}, \texttt{é}\}$$

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

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

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

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



Embeddings as Functions Over Sets

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

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

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

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

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

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

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

Embeddings as Functions Over Sets

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

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

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

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

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

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

Embeddings as Functions Over Sets

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

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

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

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

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

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

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

Embeddings as Functions Over Sets

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

$$\begin{tikzcd}
X \arrow[r, "M_x"] \arrow[d, hook] & \mathbb{R}^Y \arrow[dl, dashed, bend left=10, "M_*"] \\
\mathbb{R}^X \arrow[ur, dashed, bend left=10, "M^*"] & Y \arrow[l, "M_y"] \arrow[u, hook]
\end{tikzcd}$$

Embeddings as Functions Over Sets

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

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

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

Embeddings as Functions Over Sets

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

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

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

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

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

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

Embeddings as Functions Over Sets

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

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

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

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

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

Embeddings as Functions Over Sets

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

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

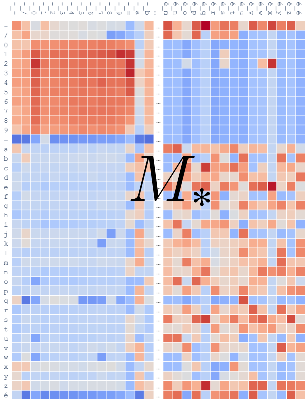
$$M_* M^* u_i = \lambda_i u_i$$
$$M^* M_* v_i = \lambda_i v_i$$

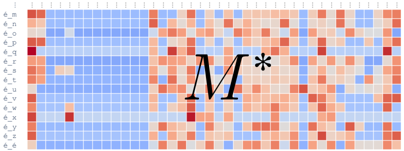
The u_i and v_i are (linear) fixed points!

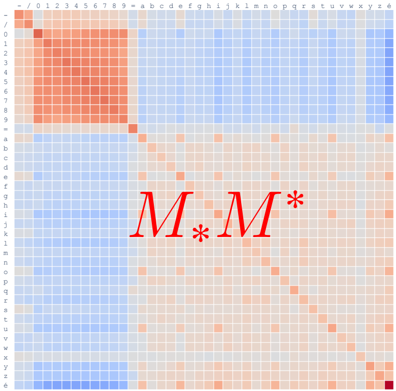
$$\begin{tikzcd}
X \arrow[r, "M_x"] \arrow[d, hook] & \mathbb{R}^Y \arrow[dl, dashed, shift left, "M_*"] \\
\mathbb{R}^X \arrow[ur, dashed, shift left, "M^*"] & Y \arrow[l, "M_y"] \arrow[u, hook]
\end{tikzcd}$$

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

$M_* M^*$ as a Covariance Matrix

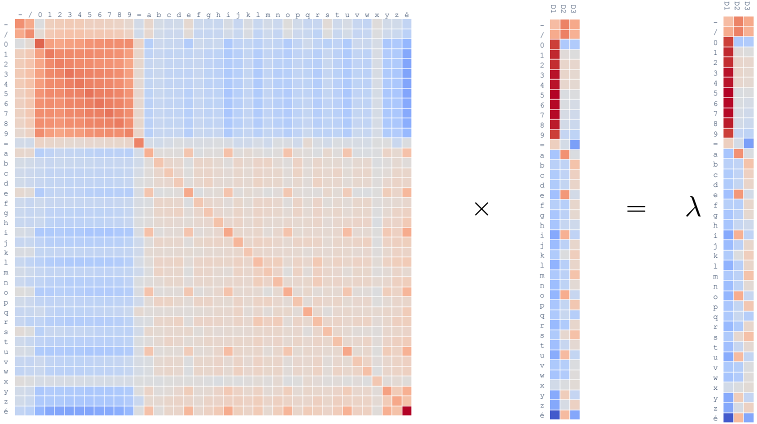


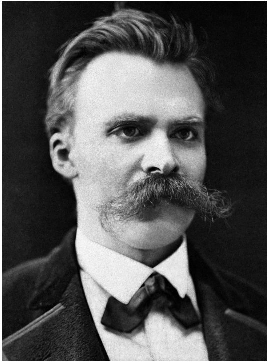




Eigenvectors as Fixed Points

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

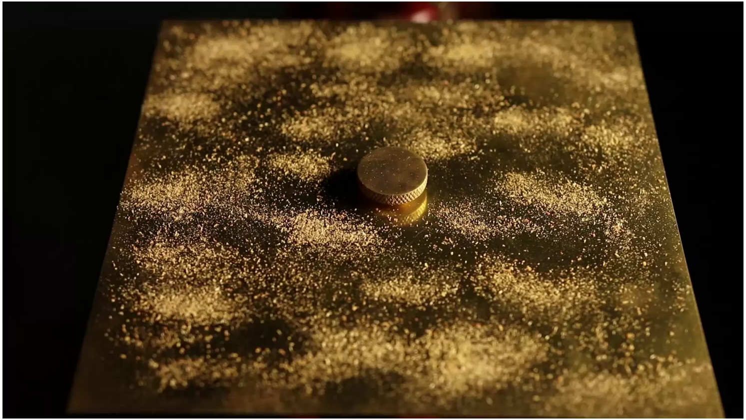




"One can conceive of a profoundly deaf human being who has never experienced sound or music; just as such a person will gaze in astonishment at the Chladnian sound-figures in sand, find their cause in the vibration of a string, and swear that he must now know what men call sound — this is precisely what happens to all of us with language."

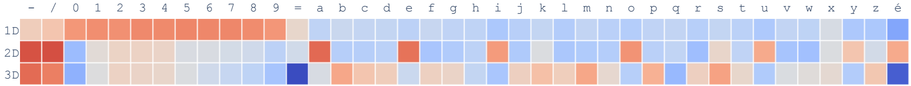
(Nietzsche, 1873)

Chladni Figures

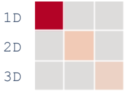


Structural Features

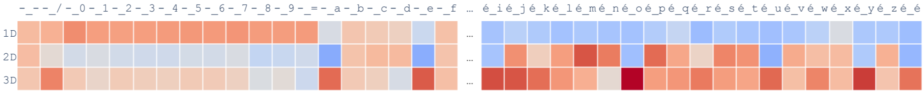
Eigenvectors of $M_* M^*$:



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



Eigenvectors of $M^* M_*$:



Words

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

Outline

Intro: Critique and Formalism

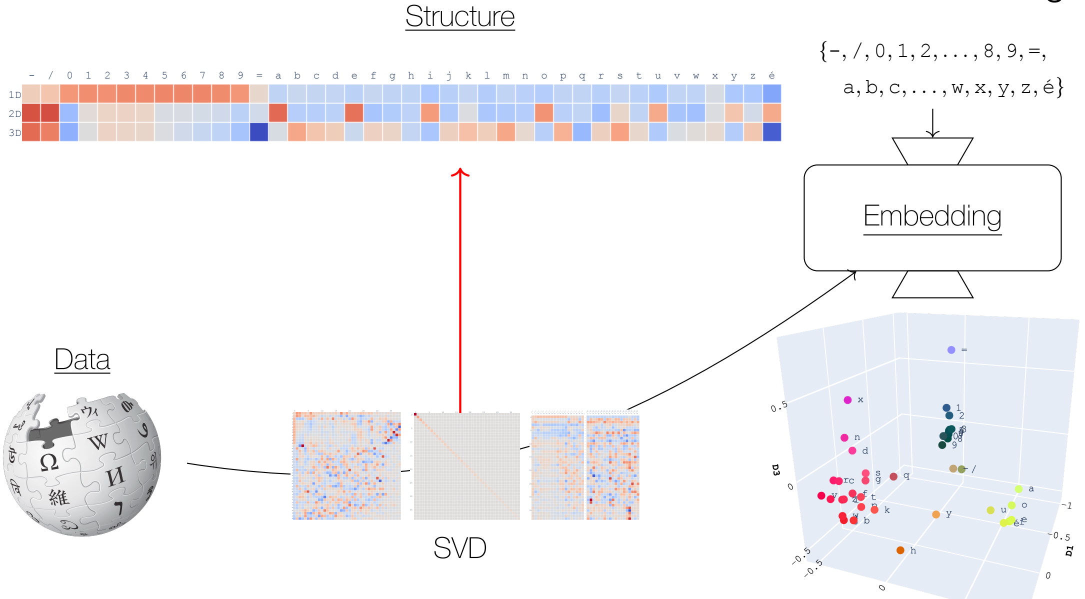
Epistemological Critique: LLMs as Formal Objects

Theoretical Critique: Formal Explainability

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

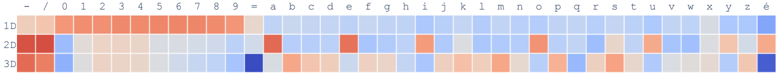
Conclusion

The Structure of Embeddings

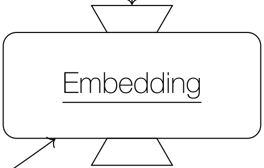


The Structure of Embeddings

Structure

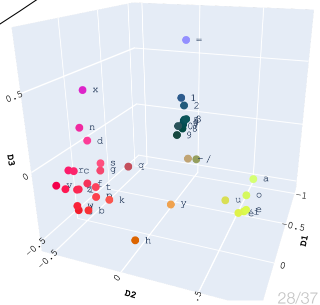


$\{-, /, 0, 1, 2, \ldots, 8, 9, =,$
$a, b, c, \ldots, w, x, y, z, é\}$



Data

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



The Structure of Embeddings

Structure

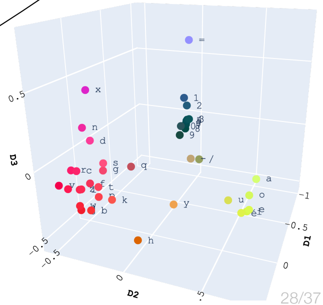
?

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

Embedding

Data

$$
\begin{tikzcd}
{\color{orange}\mathsf{C}} \arrow[r, "\mathcal{M}_c"] \arrow[d, hook, "Yoneda"'] & (\mathsf{Set}^{\color{green}\mathsf{D}})^{op} \arrow[dl, dashed, bend left=10, "\mathcal{M}_*"] \\
\mathsf{Set}^{{\color{orange}\mathsf{C}}^{op}} \arrow[ur, dashed, bend left=10, "\mathcal{M}^*"] & {\color{green}\mathsf{D}} \arrow[l, "\mathcal{M}_d"] \arrow[u, hook, "Yoneda"']
\end{tikzcd}
$$



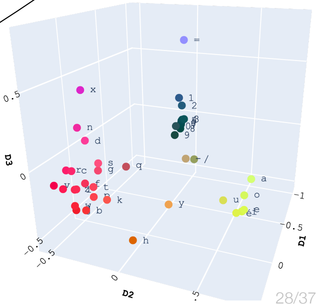
The Structure of Embeddings

Structure

?

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

Embedding



Data

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

Structure

?

$$\underbrace{\mathsf{C}^{\mathrm{op}}}_{\textit{term}_i} \times \underbrace{\mathsf{D}}_{\textit{context}_i} \longrightarrow \underbrace{\mathsf{Set}}_{\text{measure}}$$

Structure

?

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

$term_i$ → C^{op}

$context_i$ → D

measure → Set

Structure

?

$$\overset{term_i}{\downarrow}\ \ \overset{context_i}{\downarrow}\ \ \overset{\text{measure}}{\downarrow}$$

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

Structure

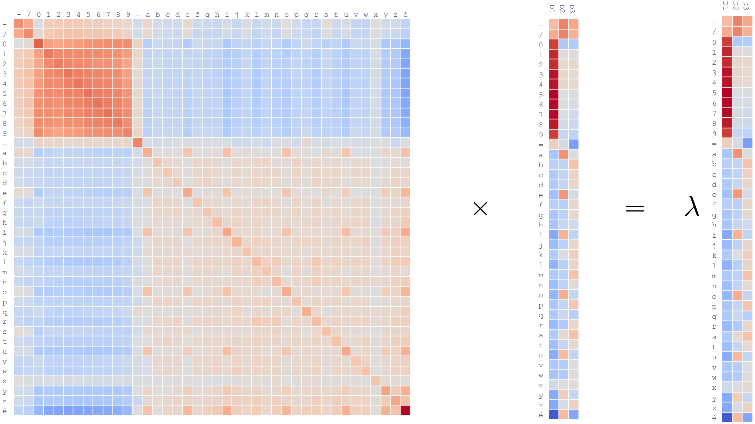
$$\mathsf{C}^{\mathrm{op}} \times \mathsf{D} \to 2$$

$$\Downarrow$$

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

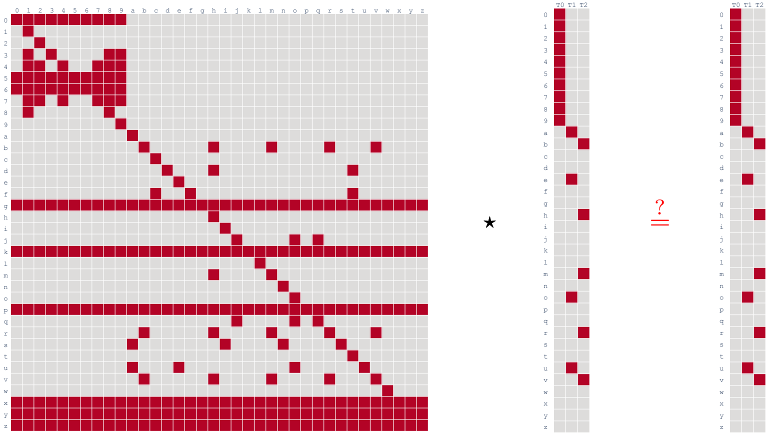
Binary Fixed Points

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



Binary Fixed Points

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

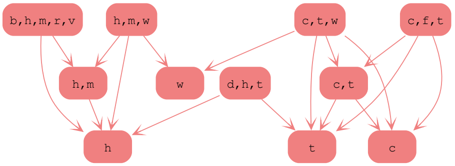


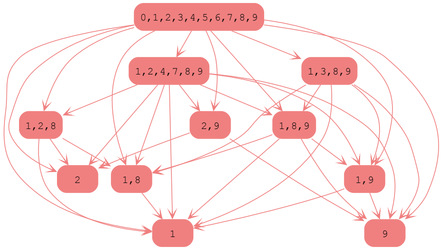
"Eigensets"

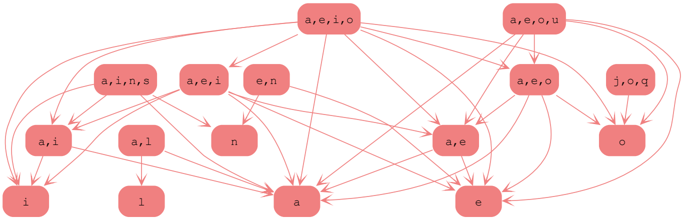
$$\mathcal{M}_*\mathcal{M}^* f = f$$

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

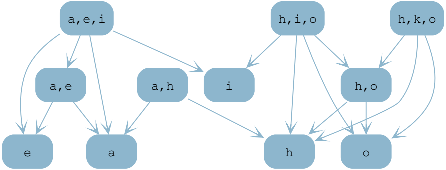
Partial Order Structure

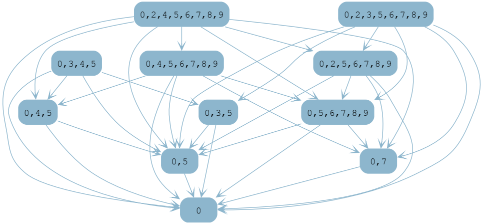


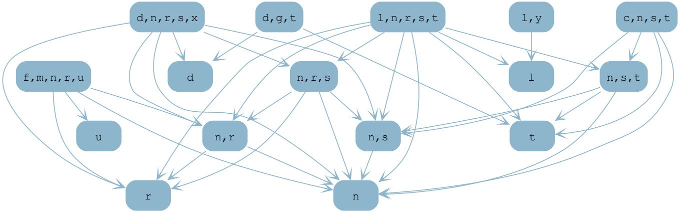




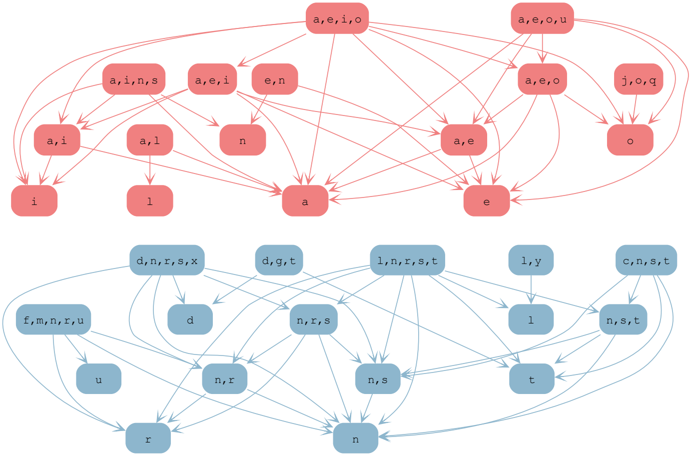
Dual Partial Order



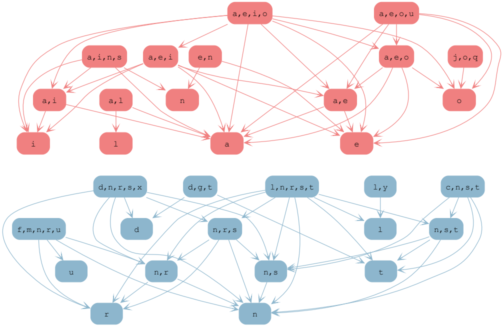




Paring of Partial Ordered Fixed Points



Structure



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

$$\Downarrow$$

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

Structure

?

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

$$\Downarrow$$

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

Structure

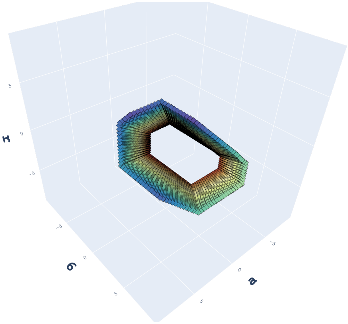
?

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

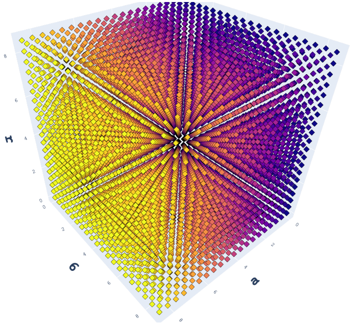
$$\Downarrow$$

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

Structure



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

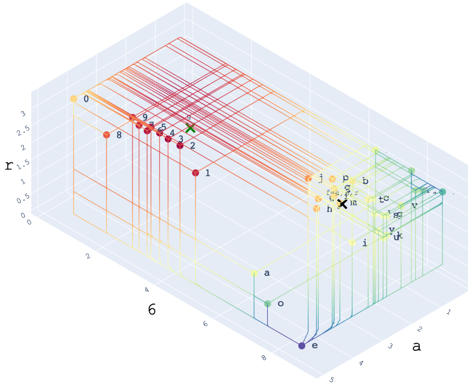


$$C^{\mathrm{op}} \times D \longrightarrow \bar{\mathbb{R}}$$

$$\Downarrow$$

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

Structure



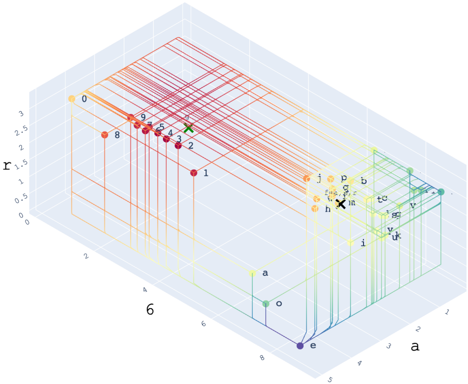
$$\mathsf{C}^{\mathrm{op}} \times \mathsf{D} \longrightarrow \bar{\mathbb{R}}$$

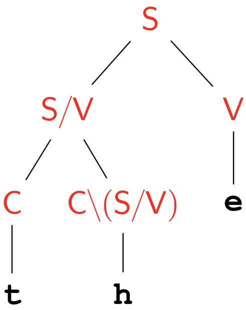
$$\Downarrow$$

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

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

Structure



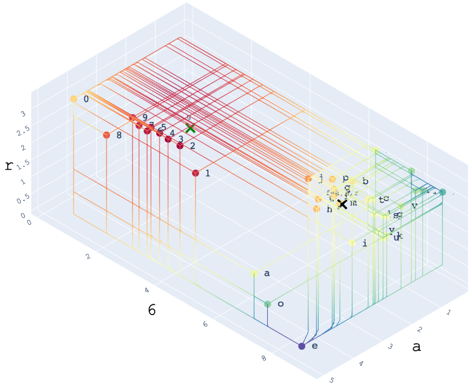


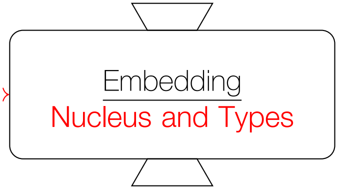
$$C^{\mathrm{op}} \times D \longrightarrow \bar{\mathbb{R}}$$

$$\Downarrow$$

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

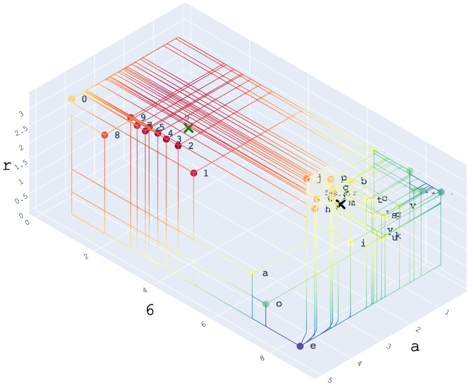
Structure

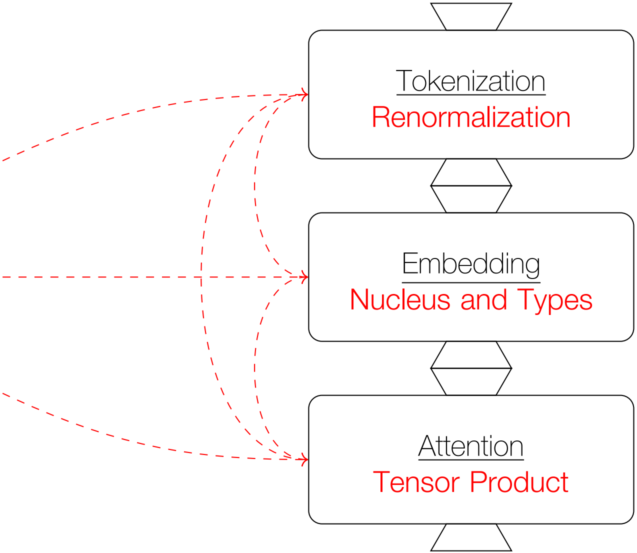




Formal Explainability

Structure





Outline

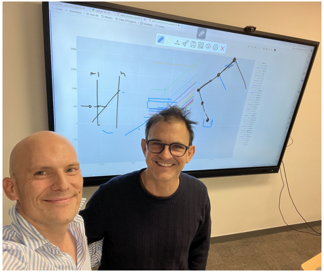
Conclusion: For a Critical Formalism

- It is urgent to address the epistemological dimension of the critical project in its own terms

- This requires to develop a critical approach within formal sciences where formalization is not assumed to lead to naturalization
    - The new role of data within formal sciences is crucial in this sense

- A critical formalism will be incomplete if it remains disconnected from the political, and even the artistic dimension of the critical program
    - Pure form of data articulated with a politics of the corpus
    - We need a new alliance between the formal sciences, the human sciences, and the arts.

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

Références I

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021).On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. https://doi.org/10.1145/3442188.3445922

Bender, E. M., & Koller, A. (2020).Climbing towards NLU: On meaning, form, and understanding in the age of data. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5185–5198. https://doi.org/10.18653/v1/2020.acl-main.463

Bourdieu, P. (1979). *La distinction: Critique sociale du jugement*. Éditions de Minuit.

Bourdieu, P. (1994). *Raisons pratiques: Sur la théorie de l'action*. Éditions du Seuil.

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

Church, A. (1936).An unsolvable problem of elementary number theory. *American Journal of Mathematics*, *58*(2), 345–363.

Foucault, M. (1966). *Les mots et les choses : Une archéologie des sciences humaines*. Gallimard.

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. (2006). *Le point aveugle: Cours de logique. vers la perfection*. Editions Hermann.

Gödel, K. (1934). On undecidable propositions of formal mathematical systems. In *Collected works* (pp. 346–371). Clarendon Press Oxford University Press.

Références II

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.

Hjelmslev, L. (1935). *La catégorie des cas*. Wilhelm Fink Verlag.

Hjelmslev, L. (1971). La structure fondamentale du langage. In *Prolégomènes à une théorie du langage* [Prolégomènes à une theorie du langage] (pp. 177–231). Éditions de Minuit.

Hjelmslev, L. (1975). *Résumé of a Theory of Language*. Nordisk Sprog-og Kulturforlag.

Jakobson, R., Fant, G. M., & Halle, M. (1952). *Preliminaries to speech analysis: The distinctive features and their correlates*. MIT Press.

Kirschenbaum, M. (2023). *Again theory: A forum on language, meaning, and intent in the time of stochastic parrot*. https://critinq.wordpress.com/2023/06/26/again-theory-a-forum-on-language-meaning-and-intent-in-the-time-of-stochastic-parrots/

Latour, B., Jensen, P., Venturini, T., Grauwin, S., & Boullier, D. (2012).'The whole is always smaller than its parts' - a digital test of Gabriel Tardes' monads. *The British Journal of Sociology*, *63*(4), 590–615.

Lévi-Strauss, C. (1949). *Les structures élémentaires de la parenté*. Presses Universitaires de France.

Lévi-Strauss, C. (1962). *La pensée sauvage*. Plon.

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. (2013).Distributed representations of words and phrases and their compositionality. *CoRR*, *abs/1310.4546*.

Références III

Nietzsche, F. (1873). On truth and lying in a non-moral sense [Originally unpublished; written in 1873.]. (R. Speirs, Trans.). In R. Geuss & R. Speirs (Eds.), *The birth of tragedy and other writings* (pp. 141–153). Cambridge University Press.

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.

Spang-Hanssen, H. (1959). *Probability and structural classification in language description*. Rosenkilde; Bagger.

Turing, A. (1937).On Computable Numbers, with an Application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society*, *s2-42*(1), 230–265. https://doi.org/10.1112/plms/s2-42.1.230

Underwood, T. (2023, October 15). *The empirical triumph of theory* [Accessed: 2023-10-15]. https://critinq.wordpress.com/2023/06/29/the-empirical-triumph-of-theory/

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

Chat Token Vector

Università Ca' Foscari

Venice, Italy

Toward a Critical Formalism

Philosophical and Theoretical Effects of a Mathematical Critique of LLMs

Juan Luis Gastaldi

`www.giannigastaldi.com`

ETH zürich

June 12, 2025