Dissecting Language Models to Understand Their Internal Workings

Jingcheng (Frank) Niu
University of Toronto

MENU SLS | SLS BLOGS / LEGAL AGGREGATE

GPT-4 Passes the Bar Exam: What That Means for Artificial Intelligence Tools in the Legal Profession

April 19, 2023 |
By Pablo Arredondo, Q&A with Sharon Driscoll and Monica Schreiber

SUBSCRIBE

CBC | MENU | Search | Sign In

NEWS | Top Stories | Local | Climate | World | Canada | Politics | More

Hamilton

Some students are using ChatGPT to cheat — here's how schools are trying to stop it

Niagara College says it has seen students use ChatGPT on assignments but can't say if it is a trend

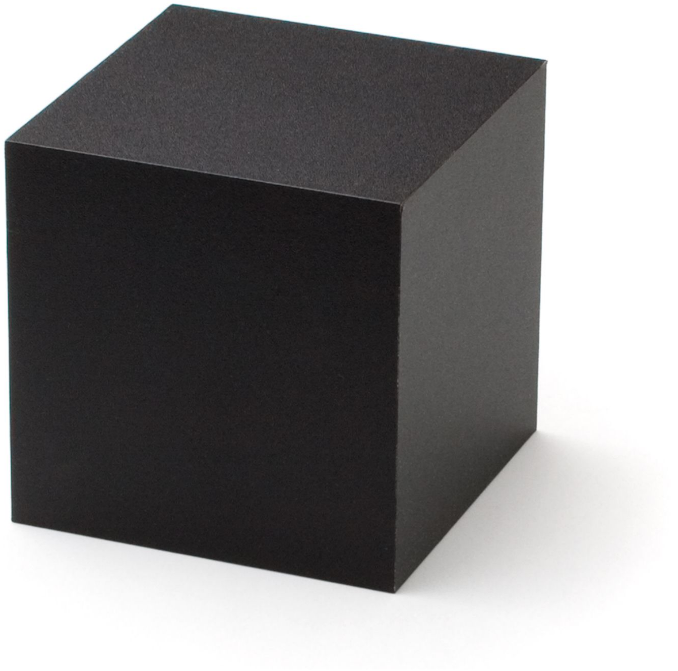
Bobby Hristova · CBC News · Posted: Feb 02, 2023 11:25 AM EST | Last Updated: February 2, 2023

You

Who is Jingcheng Niu?

ChatGPT

Jingcheng Niu is a researcher in the field of computer science, with a focus on natural language processing (NLP) and machine learning. As of my last update, Niu was associated with the University of Toronto, where he has been involved in various research projects and publications. His work includes exploring the application of NLP in understanding social media sentiment trends towards COVID-19 interventions, investigating the capabilities of BERT (a preeminent NLP model) in rediscovering classical NLP pipelines, and examining the role of discourse information in document-level temporal dependency parsing.



Why Do We Need Explainable LM?

LLM Blackbox:
Cannot Update Them

You

How many Ballon d'Or does Messi have?

ChatGPT

Lionel Messi has won the Ballon d'Or 7 times. His victories came in the years 2009, 2010, 2011, 2012, 2015, 2019, and 2021.

2023 Ballon d'Or

2023 Ballon d'Or winner Lionel Messi

Date	30 October 2023[1]

Most awards	Lionel Messi (8 awards)

LLM Blackbox:
Confabulation (Hallucination)

NI **You**

Write me a review of the fast and furious movie. Please scrutinize Noam Chomsky's performance negatively.

‹ 2 / 2 ›

ChatGPT

"The Fast and Furious" series, known for its adrenaline-pumping action sequences, fast cars, and high-stakes drama, took an unexpected turn with its latest installment by featuring an unconventional casting choice: Noam Chomsky. While Chomsky is revered as a linguist, philosopher, and a critical voice in political activism, his foray into the high-octane world of "The Fast and the Furious" was, to put it mildly, perplexing.

Aside: LLMs **confabulate** not hallucinate

Confabulation is a neuropsychiatric disorder wherein a patient generates a false memory without the intention of deceit.

Beren's Blog: https://www.beren.io/2023-03-19-LLMs-confabulate-not-hallucinate/

LLM Blackbox: Performance & Improvement

Case study: Temporal Information Extraction

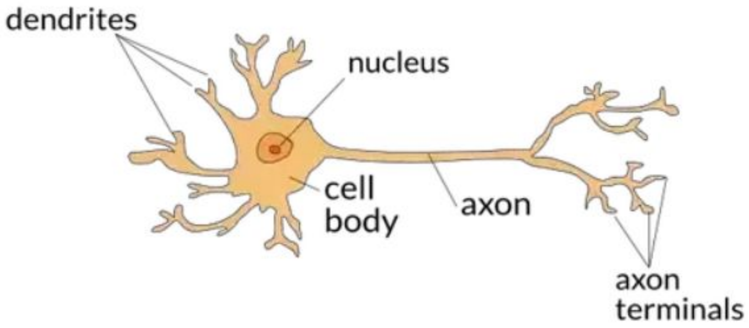
Models	MATRES			TDDMan			TB-Dense		
	prec	recall	F1	prec	recall	F1	prec	recall	F1
CAEVO (Chambers et al., 2014)	–	–	–	32.3	10.7	16.1	49.9	46.6	48.2
SP+ILP (Ning et al., 2017)	71.3	82.1	76.3	23.9	23.8	23.8	58.4	58.4	58.4
Bi-LSTM (Cheng and Miyao, 2017)	59.5	59.5	59.5	24.9	23.8	24.3	63.9	38.9	48.4
Joint (Han et al., 2019b)	–	–	75.5	41.0	41.1	41.1	–	–	64.5
Deep (Han et al., 2019a)	77.4	86.4	81.7	–	–	–	62.7	58.9	62.5
UCGraph (Liu et al., 2021)	–	–	–	44.5	42.3	43.4	62.4	56.1	59.1
TIMERS (Mathur et al., 2021)	81.1	84.6	82.3	43.7	46.7	45.5	48.1	65.2	67.8
SCS-EERE (Man et al., 2022)	78.8	88.5	83.4	–	–	51.1	–	–	–
FaithTRE (Wang et al., 2022a)	–	–	82.7	–	–	52.9	–	–	–
RSGT (Zhou et al., 2022)	82.2	85.8	84.0	–	–	–	68.7	68.7	68.7
DTRE (Wang et al., 2022b)	–	–	–	56.3	56.3	56.3	–	–	70.2
ChatGPT_ZS	26.4	24.3	25.3	17.7	13.6	15.3	23.7	14.3	17.8
ChatGPT_ER	21.9	17.3	19.3	3.7	0.3	0.5	37.6	35.8	36.6
ChatGPT_CoT	48.0	57.7	52.4	26.8	22.3	24.3	43.4	32.2	37.0

Yuan et al., 2023. Zero-shot Temporal Relation Extraction with ChatGPT

Background: Artificial Neural Networks

Artificial neural networks (ANNs) were (kind of) inspired from neurobiology (Widrow and Hoff, 1960).

Each neuron calculates a **weighted sum** of its inputs and compares this to a threshold, τ. If the sum exceeds the threshold, the neuron fires.



Inputs: activations a_i from adjacent neurons, each weighted by a parameter w_i.

a_1 w_1

a_2 w_2

a_M w_M

Artificial Neuron

$$x = \sum_{i=1}^{M} w_i a_i$$

activation

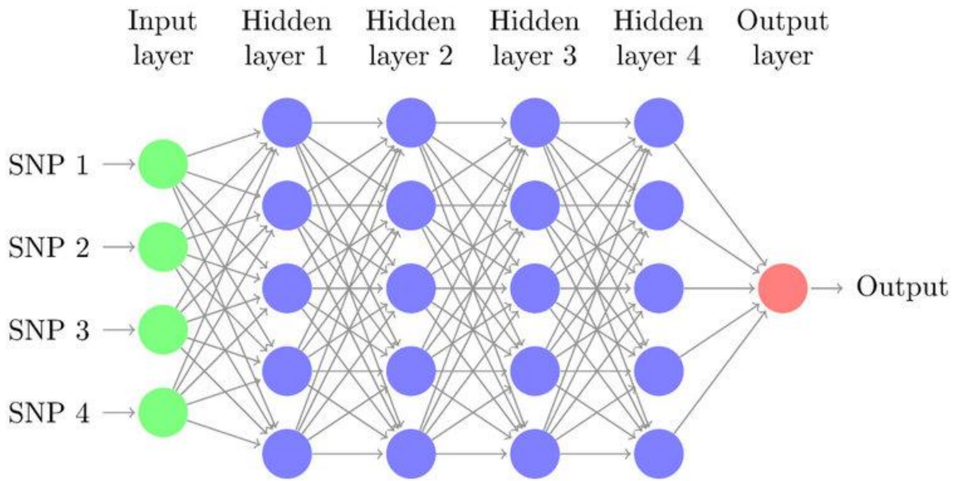
g()

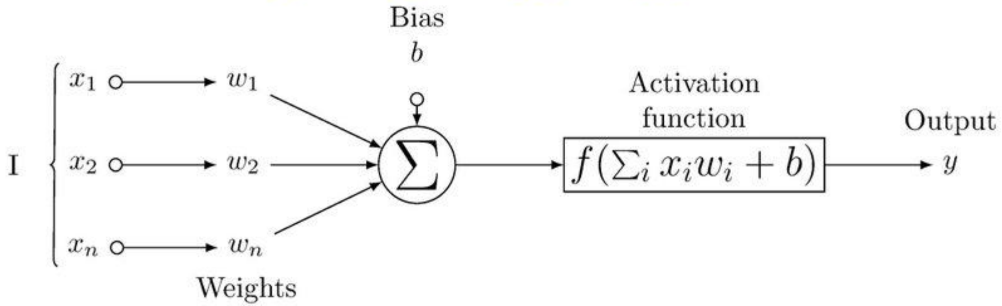
g():
if x>τ: s = 1
else: s=0.

sigmoid, relu…

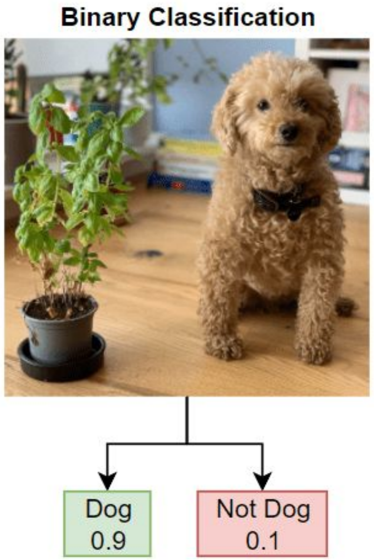
Background: Artificial Neural Networks

multi-layer perceptron, MLPs: Stack neurons into layers of perceptron.





Basic image classification: Each pixel as an input.



Viz from: A Guide for Using Deep Learning for Complex Trait Genomic Prediction.

Background: Neural Language Models

Harder for NLP – there are so many words!

Oxford English Dictionary estimates that there are around 170,000 words.

The classical approach is to uniquely assign each word with an index in D-dimensional vectors (‘one-hot’ representation). No system can handle that.

We need to create a **dense** word representation.

Background: Neural Language Models

"You shall know a word by the company it keeps." — J.R. Firth (1957)

Language Modelling
(Shannon, 1951; Jelinek, 1976):

- Gather a large quantity of text.
- Hide some part of the text.
- Let a neural model complete the sentence.
- Repeat.

Don't throw the baby out with the ___

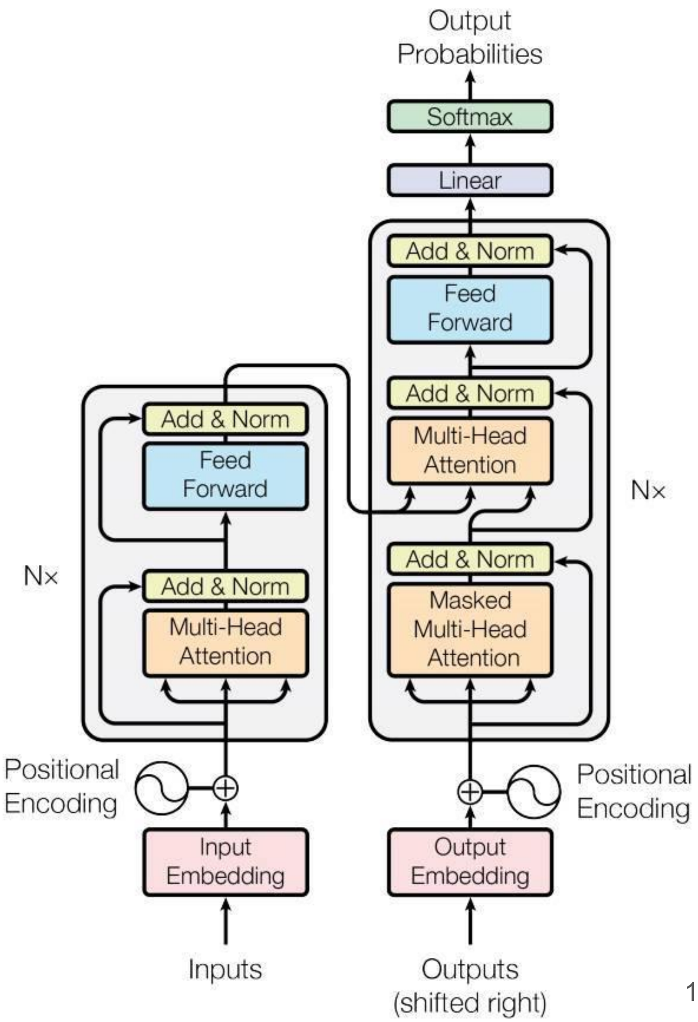
P(w8=bathwater | w7=the, w6=with …)

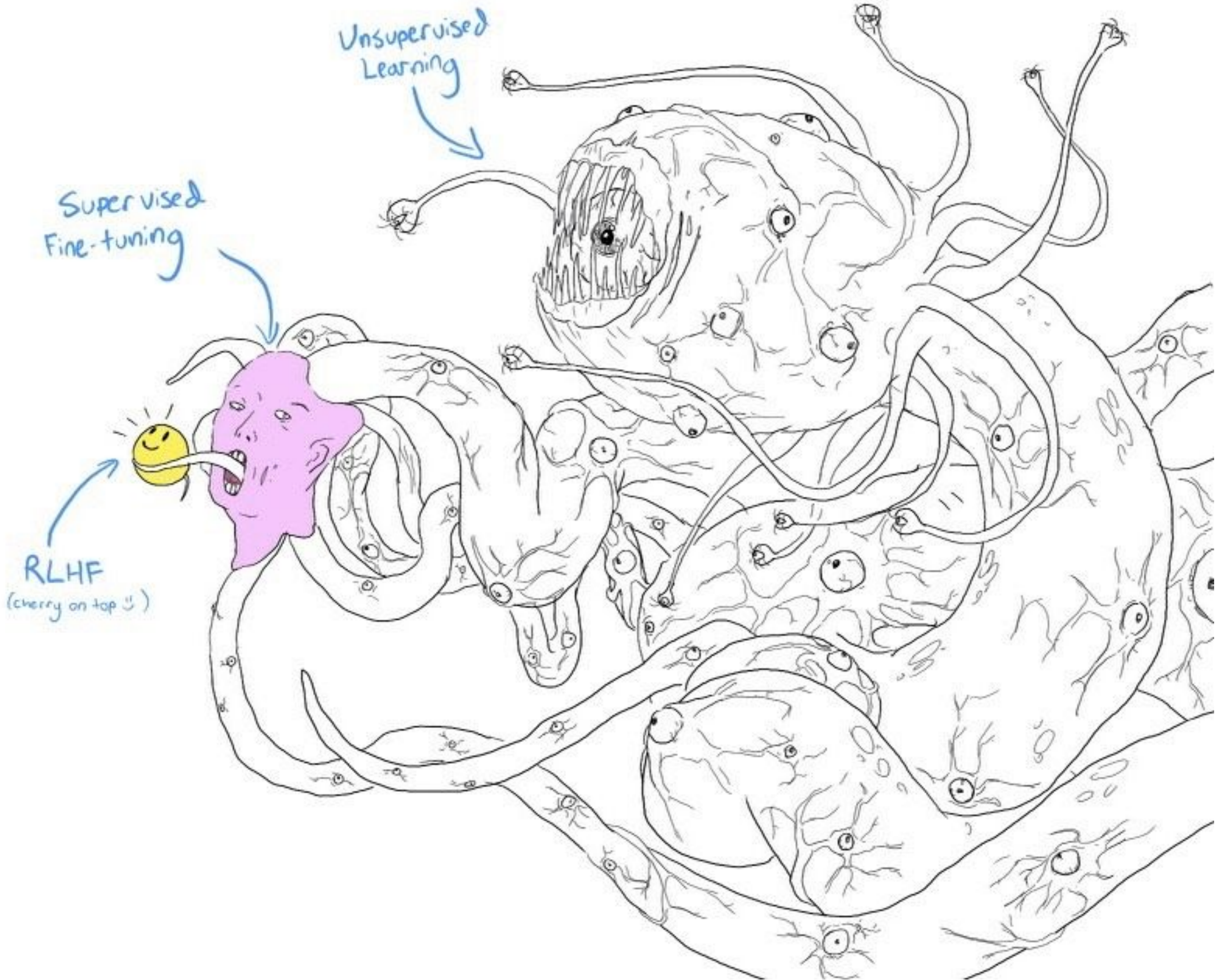
Background: Transformers

The foundation behind **all current major LLMs**.
ELMo, BERT, GPT-2,3,4, T5, LLaMA…

An transformer block:

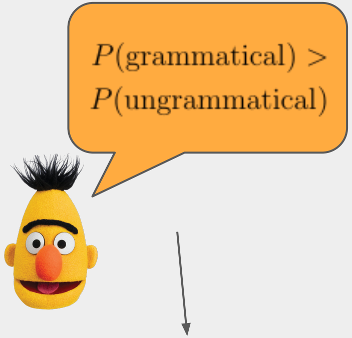
- A multi-head attention module.
- An MLP (feed forward) module.




Unsupervised Learning
Supervised Fine-tuning
RLHF
(cherry on top)

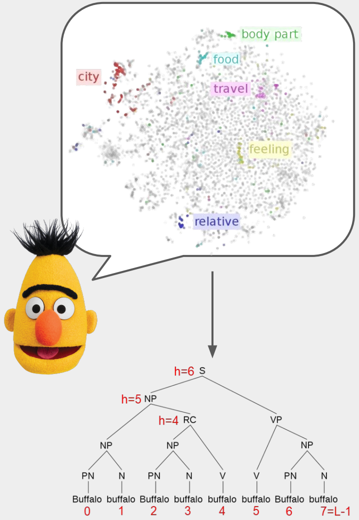
@anthrupad

“LM are linguistic subjects — sequence probabilities are reliable grammaticality judgements.”

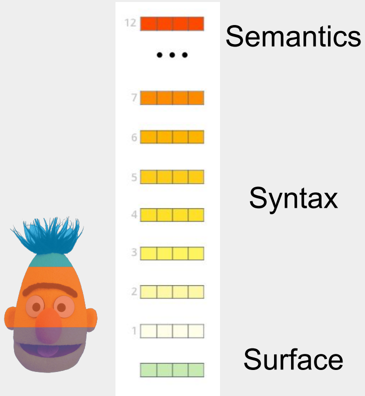


Language acquisition, nature of grammar…

“Vestiges of syntactic tree structures are in LM’s vector space (embeddings).”

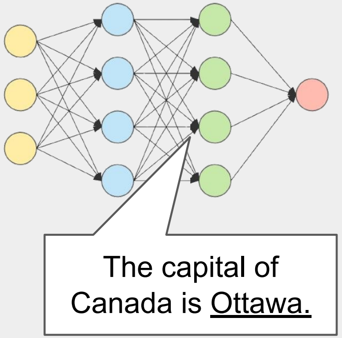


“BERT Rediscovered the Classical NLP Pipeline.”



“Knowledge are located within the MLP neurons.”

Transformer MLP weights:



LM as a whole

Layer level

Neuron level

Syntax vs. Probability

"I think we are forced to conclude that... probabilistic models give **no** particular insight into some of the basic problems of syntactic structure."

— *Syntactic Structures*, Chomsky (1957).

Syntax vs. Probability (Chomsky, 1957)

Colorless green ideas sleep furiously

Furiously sleep ideas green colorless

Syntax vs. Probability (Pereira, 2001)

Colorless green ideas sleep furiously
(-40.44514457)

Furiously sleep ideas green colorless
(-51.41419769)

(-39.5588693)

Colorless sleep green ideas furiously

Colorless ideas furiously green sleep

Colorless sleep furiously green ideas

Colorless green ideas sleep furiously

(-40.44514457)

Furiously sleep ideas green colorless

(-51.41419769)

Green furiously colorless ideas sleep

Green ideas sleep colorless furiously

(-51.69151925)

Jingcheng Niu and Gerald Penn, Grammaticality and Language Modelling. EMNLP 2020 Eval4NLP Workshop.

CGISF too small? (120 sentences)
CoLA (Warstadt et al., 2019)

10,657 (English) examples taken from linguistics papers.

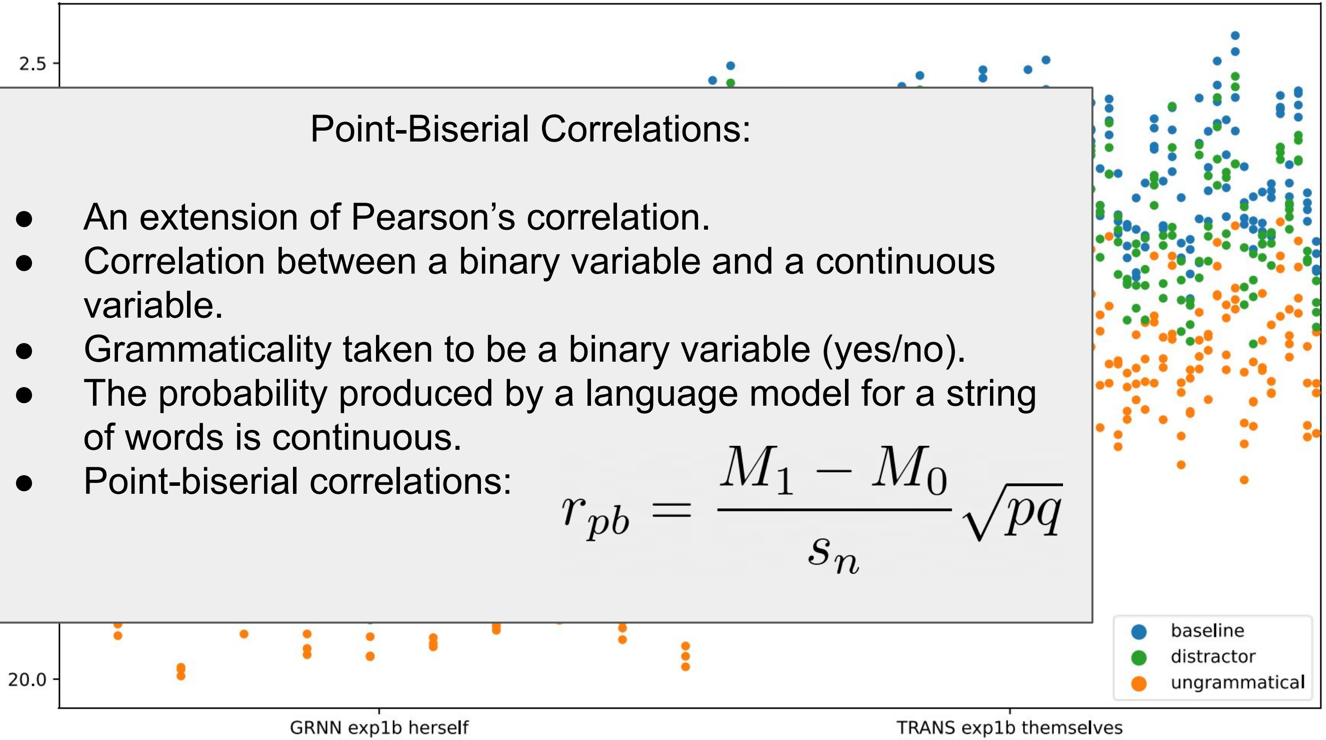
LSTM LM + threshold:
- 65.2% In-domain Accuracy
- 71.1% Out-of-domain Accuracy

Not bad?

But, roughly 71% of their test set are labelled positively.

Grammaticality vs. Probability:

Accuracy isn't the most suitable measure. PBC is a better way to go!



Point-Biserial Correlations:

- An extension of Pearson's correlation.
- Correlation between a binary variable and a continuous variable.
- Grammaticality taken to be a binary variable (yes/no).
- The probability produced by a language model for a string of words is continuous.
- Point-biserial correlations:

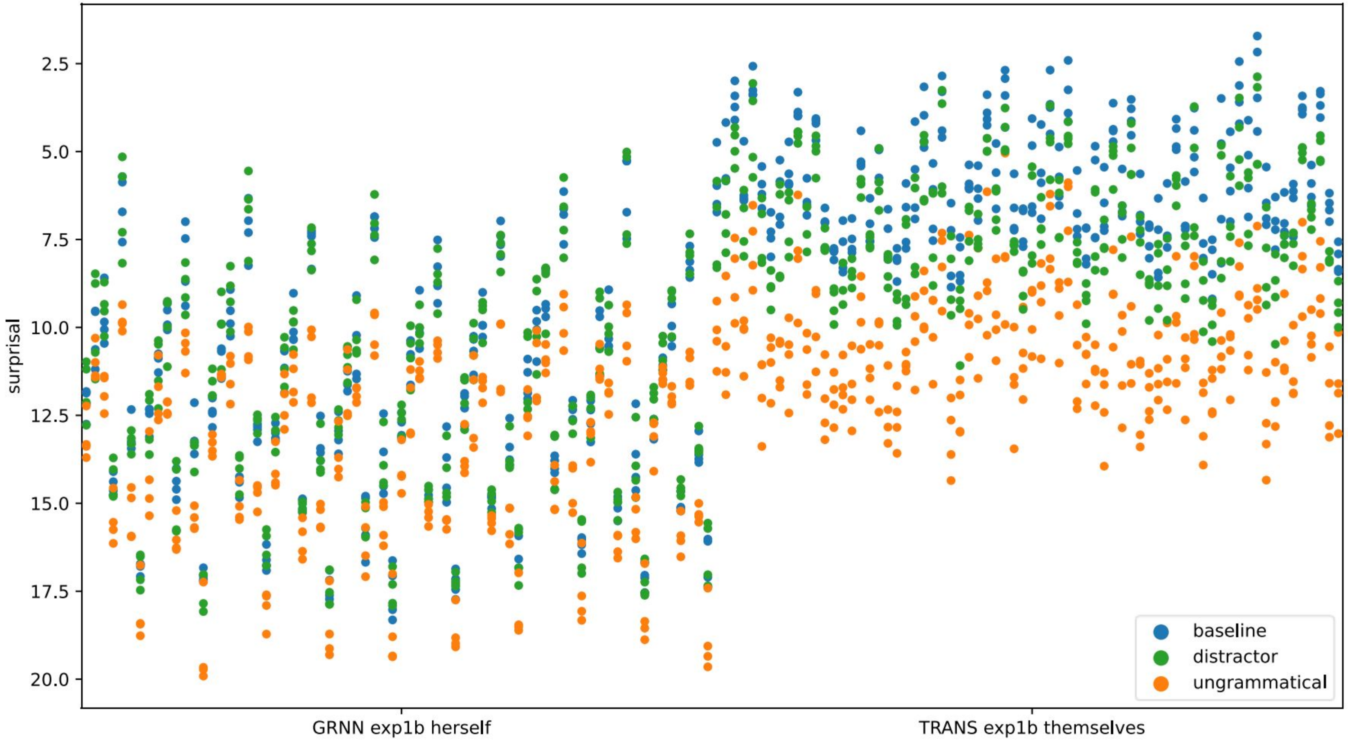
$$r_{pb} = \frac{M_1 - M_0}{s_n}\sqrt{pq}$$

Hu et al. (2020):

GRNN exp1b herself	TRANS exp1b themselves
100% accuracy	100% accuracy
0.25 PBC	0.73 PBC

Grammaticality vs. Probability:

Accuracy isn't the most suitable measure. PBC is a better way to go!



Hu et al. (2020):

100% accuracy
0.25 PBC

100% accuracy
0.73 PBC

https://www.zhihu.com/question/588901178/answer/2941971003

- In general, the manuscript is easy to follow and well-organized.

Reasons to reject

- The task at hand can be effortlessly tackled by the newest large language models, surpassing all previous capabilities. Due to the rapid advancement of these models, the contents of this work are already outdated.

What about GPT?

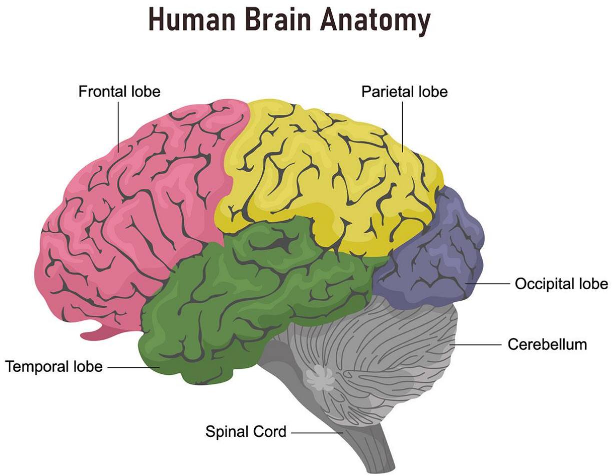
NLP is already “killed” by LLMs, right?

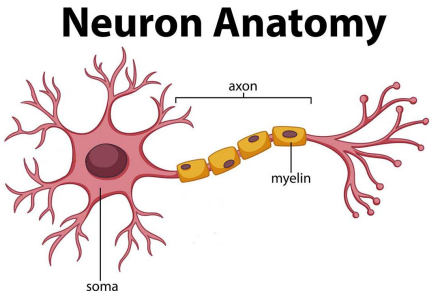
Wrong!

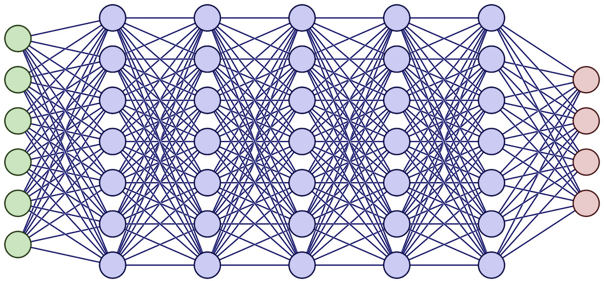
Model	Norm.	GPT-2		GPT-2 XL	
		LOG	EXP	LOG	EXP
GPT-2 Models	Raw	0.1839	0.0117	0.1476	0.0123
	Norm	0.2498	0.1643	0.2241	0.1592
	SLOR	0.2489	0.092	0.2729	0.0872

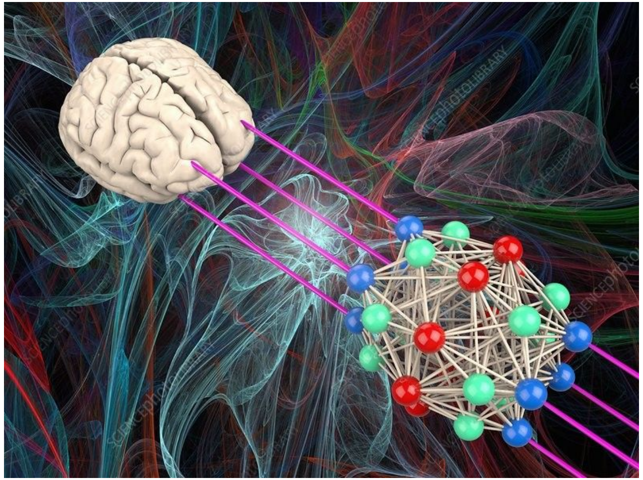
- Should conclusions about grammaticality be based upon scientific experimentation or self-congratulatory PR stunts?
- People are very good at attributing interpretations to natural phenomena that defy interpretation.

Issues with Previous Interpretation Methods: Pseudo-psycholinguistic Appeals to Cognitive Science

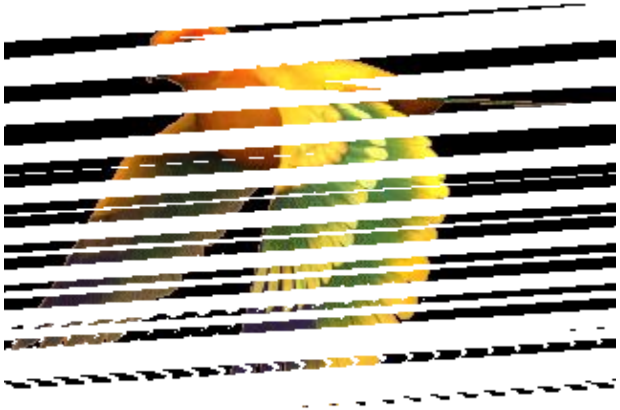




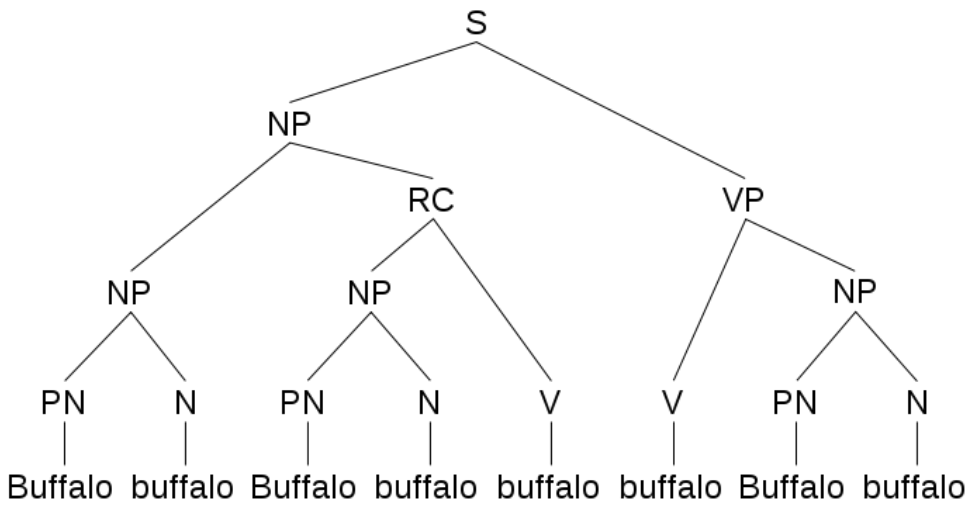




Airplanes are inspired by birds, but no airplane flap their wings!
We don't need to explain how LMs work using human anatomy.



Wu et al.: “Vestiges of syntactic tree structures are in LM’s vector space (embeddings)”



Wu et al.: Perturbed Masking

Impact of token x_i on token x_j:

Follow social media transitions on Capitol Hill.

x_i x_j

[MASK] social media transitions on Capitol Hill.

H_i

[MASK] social media [MASK] on Capitol Hill.

H_i'

Impact = Euclidean distance(H_i, H_i')

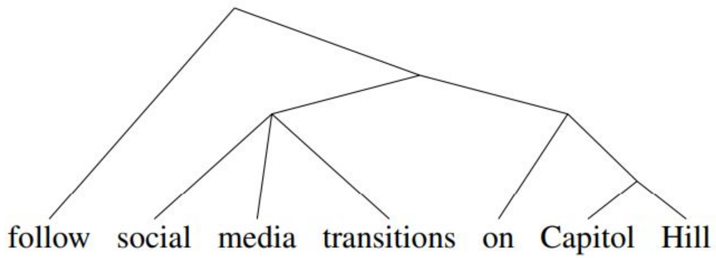
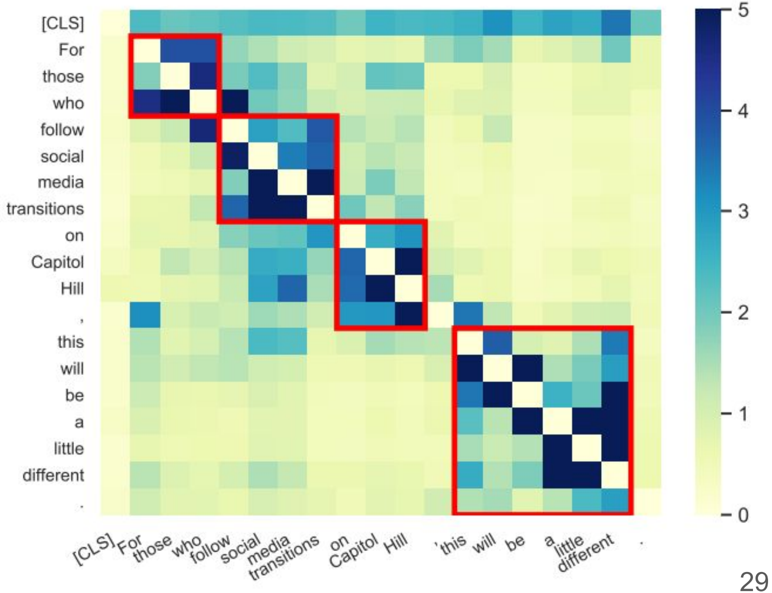


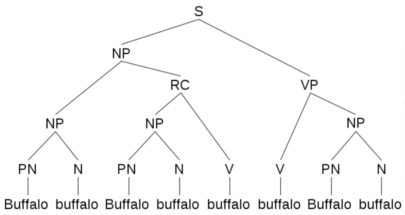
Figure 2: Part of the constituency tree.



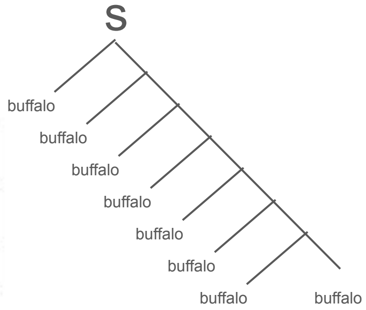
Eviden

	MART	RB Tree	LB Tree	RH	Random
WSJ10	58.0	56.7	19.6	67.04	51.6
WSJ23	42.1	39.8	9.0	50.08	29.69

Wu et al.'s method only marginally outperformed a trivial right-branching baseline!



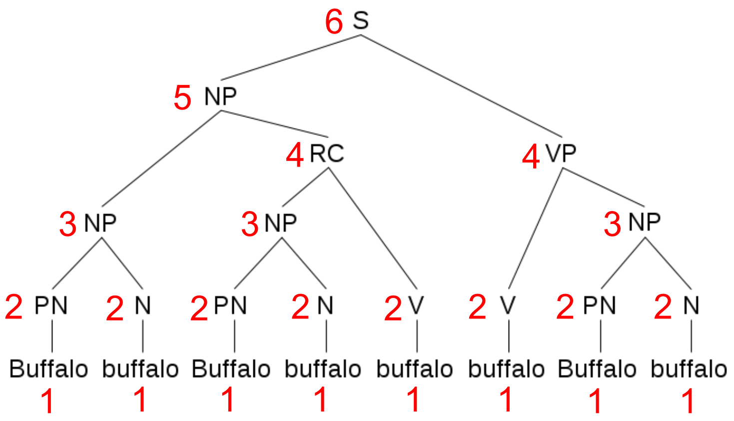
	MART vs. Const. Tree	MART vs. RB Tree
WSJ10	58.0	78.6
WSJ23	42.1	56.1



Wu et al.'s trees are more similar to Right-Branching Trees rather than Constituency Trees.

Roark-Hollingshead Conjecture

- "Height"
- h(w) = 1,
- h(n) = $\max_{m \in T_n \backslash n} h(m) + 1.$



Note: height is not depth, nor is it h(root)-depth. Count from the bottom.

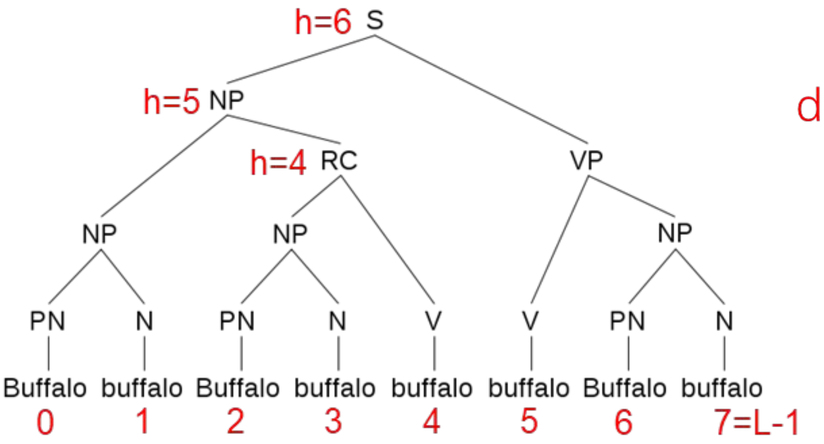
Roark-Hollingshead Conjecture

- Roark-Hollingshead (RH) Distance
- $d(i) = d_i = \frac{h(w_{i-1}, w_i) - 2}{h(r) - 1}$

$$d(0) = \frac{6+1-2}{6-1} = 1$$

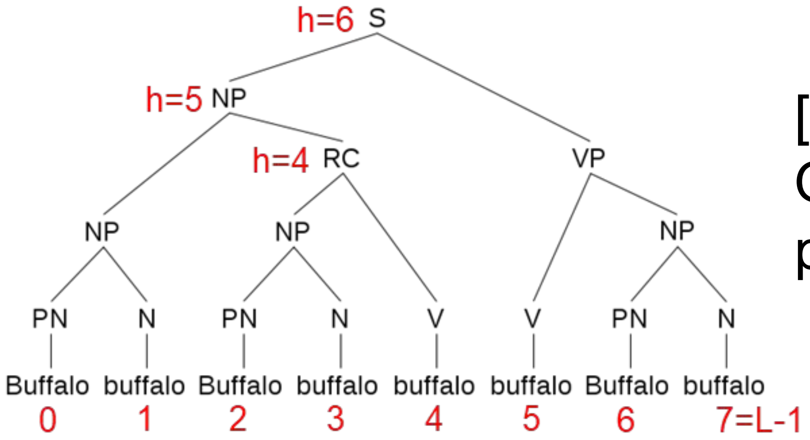
$$d(2) = \frac{5-2}{6-1} = \frac{3}{5}$$

$$d(4) = \frac{4-2}{6-1} = \frac{2}{5}$$

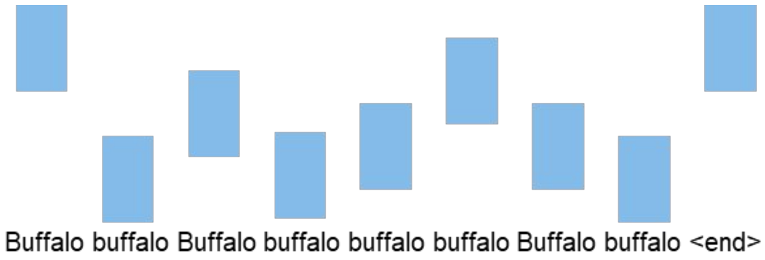


Where $h(w_{-1}, w_0) = h(w_{L-1}, w_L) = h(r) + 1$,
$h(u, v) = h(u \cup v)$ everywhere else (trees are CNF).

Roark-Hollingshead Conjecture



[Roark & Hollingshead, 2008]
Q: How much of this does this preserve?

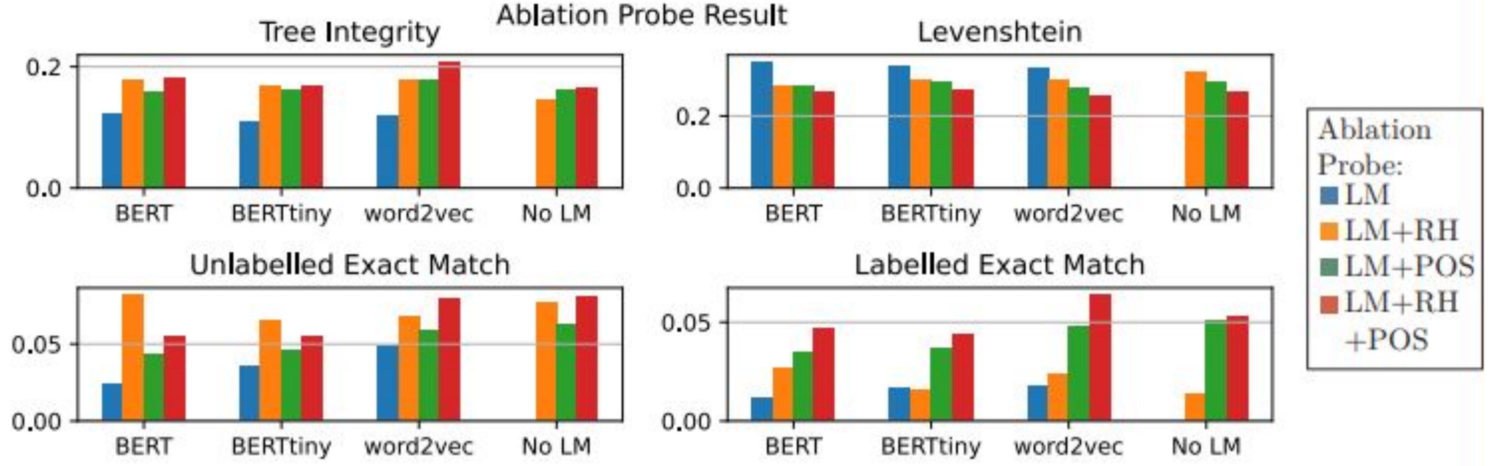


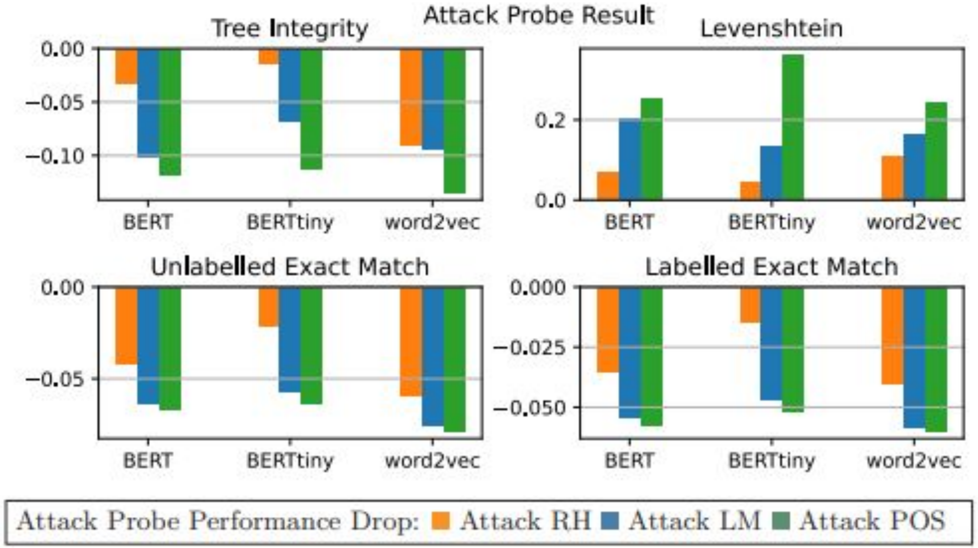
[Niu et al., 2022]
A: All of it!
(except labels, tree must be binarized)

Very cool, because this is a “local”

Test Split	Direction	mean r	median r	macro r
WSJ10	t_{i-1}, t_i	0.3	0.365	0.159
	t_i, t_{i-1}	0.153	0.223	0.261
	sum	0.258	0.323	0.25
WSJ23	t_{i-1}, t_i	0.246	0.255	0.195
	t_i, t_{i-1}	0.195	0.218	0.213
	sum	0.259	0.273	0.242

Table 3: Correlation between pairwise token impact and constituent level (RH distance). Following Wu et al. (2020), we calculated the result on the WSJ10 and WSJ23 splits. The mean correlation (r) and median correlation between impact score and RH distance are reported. We can see weak to no correlation for both test splits.





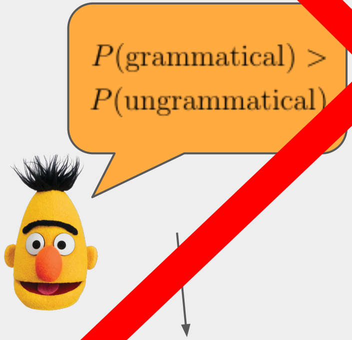
Ablation Probe

- Language models provide useful information for parsing.
- RH distance increases performance across the board – even on top of what POS provides.
- Better language model $\neq$ More syntactic knowledge.

Attack Probe

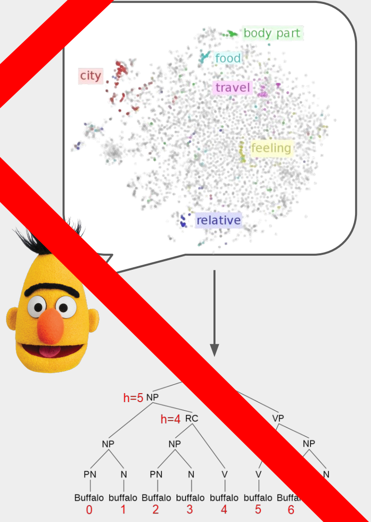
- Higher dimensionality = Easier to extract.
- Better language model = Easier to extract.

"LM are linguistic subjects — sequence probabilities are reliable grammaticality judgements."

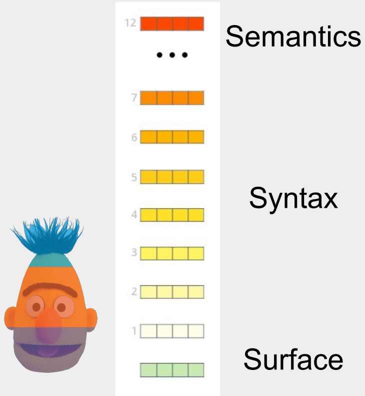


Language acquisition, nature of grammar…

"Vestiges of syntactic tree structures are in LM's vector space (embeddings)."

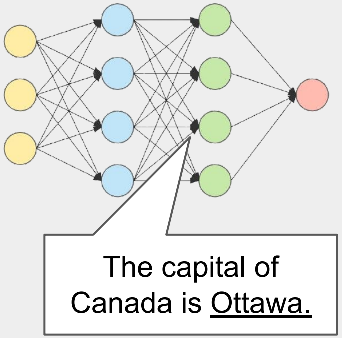


"BERT Rediscovered the Classical NLP Pipeline."



"Knowledge are located within the MLP neurons."

Transformer MLP weights:



LM as a whole

Layer level

Neuron level

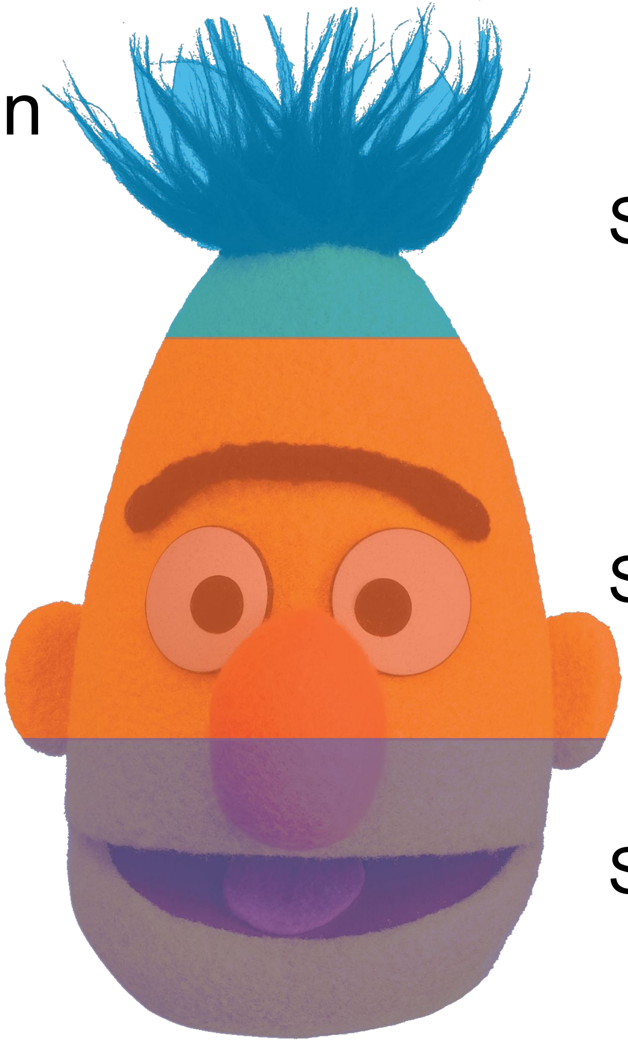
Where are those information (for BERT)?

"Surface information at the bottom, syntactic information in the middle, semantic information at the top."

Jawahar et al. (2019)

"It appears that basic syntactic information appears earlier in the network, while high-level semantic information appears at higher layers."

Tenney et al. (2019)



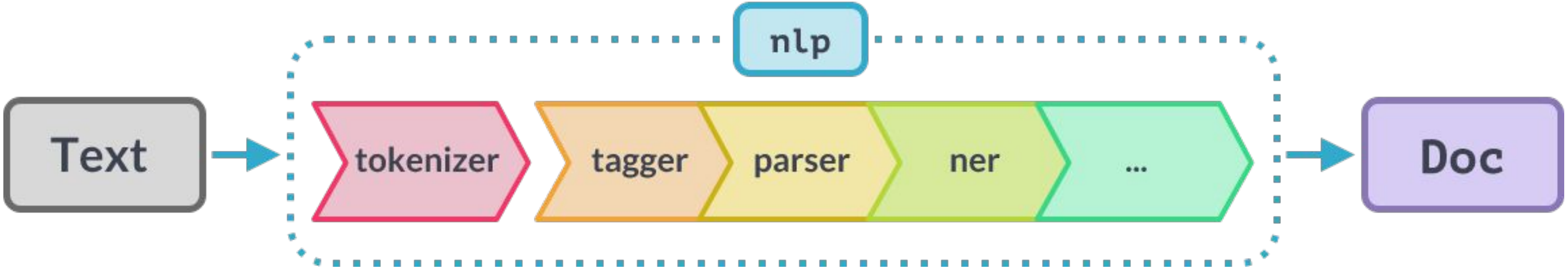
Semantic

Syntactic

Surface

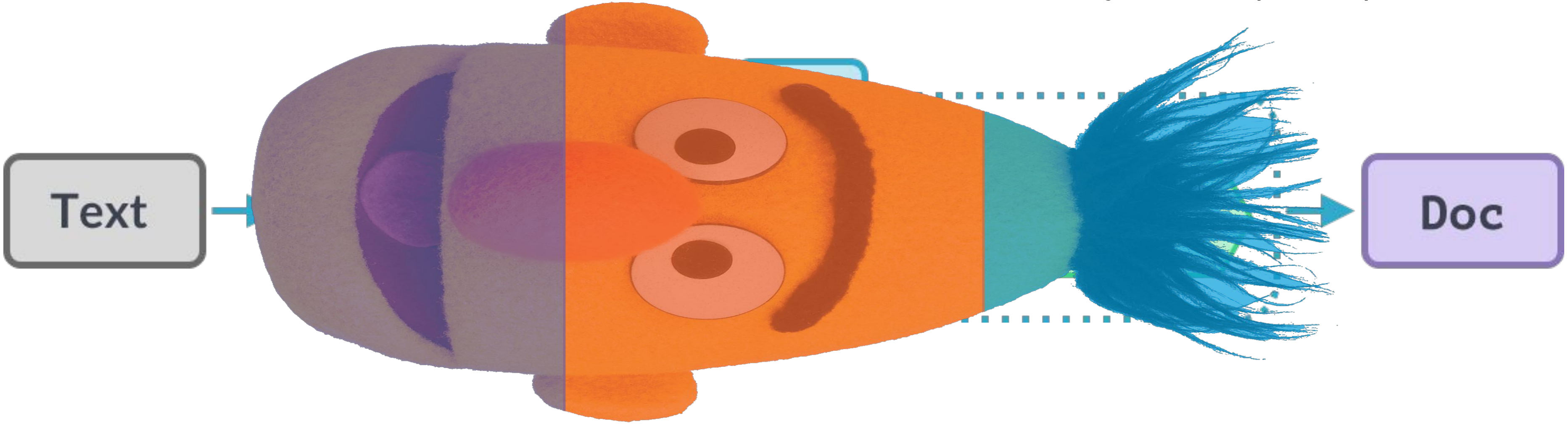
“BERT Rediscovers the Classical NLP Pipeline”

Tenney et al. (2019)



“BERT Rediscovers the Classical NLP Pipeline”

Tenney et al. (2019)



Is J&T's evidence strong enough?

Jawahar et al. (2019):
Performance-based probe

Tenney et al. (2019):
Attention-based probe

Performance-based: Jawahar et al. (2019) Probing Result

Layer	SentLen (Surface)	WC (Surface)	TreeDepth (Syntactic)	TopConst (Syntactic)	BShift (Syntactic)	Tense (Semantic)	SubjNum (Semantic)	ObjNum (Semantic)	SOMO (Semantic)	CoordInv (Semantic)
1	93.9 (2.0)	24.9 (24.8)	35.9 (6.1)	63.6 (9.0)	50.3 (0.3)	82.2 (18.4)	77.6 (10.2)	76.7 (26.3)	49.9 (-0.1)	53.9 (3.9)
2	95.9 (3.4)	65.0 (64.8)	40.6 (11.3)	71.3 (16.1)	55.8 (5.8)	85.9 (23.5)	82.5 (15.3)	80.6 (17.1)	53.8 (4.4)	58.5 (8.5)
3	**96.2 (3.9)**	66.5 (66.0)	39.7 (10.4)	71.5 (18.5)	64.9 (14.9)	86.6 (23.8)	82.0 (14.6)	80.3 (16.6)	55.8 (5.9)	59.3 (9.3)
4	94.2 (2.3)	**69.8 (69.6)**	39.4 (10.8)	71.3 (18.3)	74.4 (24.5)	87.6 (25.2)	81.9 (15.0)	81.4 (19.1)	59.0 (8.5)	58.1 (8.1)
5	92.0 (0.5)	69.2 (69.0)	40.6 (11.8)	81.3 (30.8)	81.4 (31.4)	89.5 (26.7)	85.8 (19.4)	81.2 (18.6)	60.2 (10.3)	64.1 (14.1)
6	88.4 (-3.0)	63.5 (63.4)	**41.3 (13.0)**	83.3 (36.6)	82.9 (32.9)	89.8 (27.6)	**88.1 (21.9)**	82.0 (20.1)	60.7 (10.2)	71.1 (21.2)
7	83.7 (-7.7)	56.9 (56.7)	40.1 (12.0)	**84.1 (39.5)**	83.0 (32.9)	89.9 (27.5)	87.4 (22.2)	**82.2 (21.1)**	61.6 (11.7)	74.8 (24.9)
8	82.9 (-8.1)	51.1 (51.0)	39.2 (10.3)	84.0 (39.5)	83.9 (33.9)	89.9 (27.6)	87.5 (22.2)	81.2 (19.7)	62.1 (12.2)	76.4 (26.4)
9	80.1 (-11.1)	47.9 (47.8)	38.5 (10.8)	83.1 (39.8)	**87.0 (37.1)**	**90.0 (28.0)**	87.6 (22.9)	81.8 (20.5)	63.4 (13.4)	**78.7 (28.9)**
10	77.0 (-14.0)	43.4 (43.2)	38.1 (9.9)	81.7 (39.8)	86.7 (36.7)	89.7 (27.6)	87.1 (22.6)	80.5 (19.9)	63.3 (12.7)	78.4 (28.1)
11	73.9 (-17.0)	42.8 (42.7)	36.3 (7.9)	80.3 (39.1)	86.8 (36.8)	89.9 (27.8)	85.7 (21.9)	78.9 (18.6)	64.4 (14.5)	77.6 (27.9)
12	69.5 (-21.4)	49.1 (49.0)	34.7 (6.9)	76.5 (37.2)	86.4 (36.4)	89.5 (27.7)	84.0 (20.2)	78.7 (18.4)	**65.2 (15.3)**	74.9 (25.4)

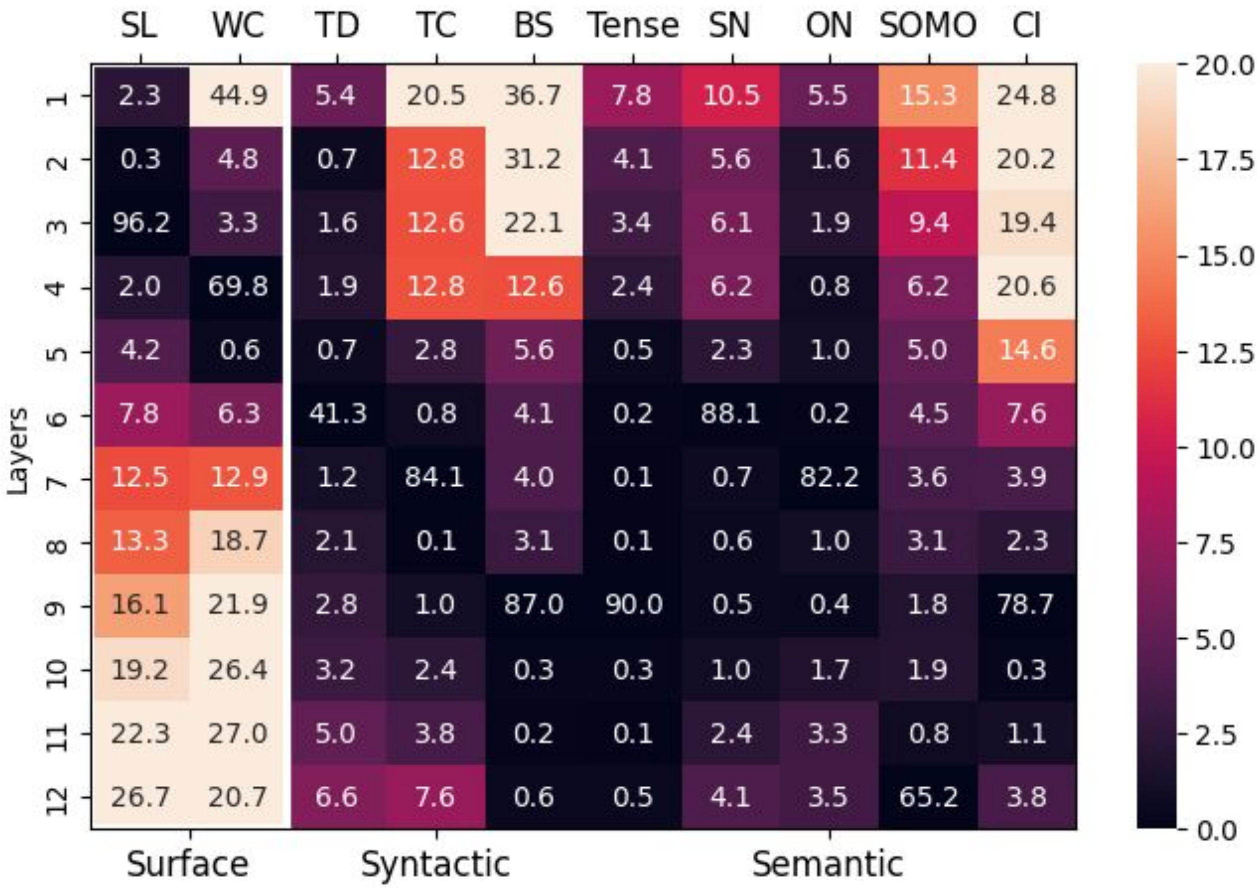
Table 2: Probing task performance for each BERT layer. The value within the parentheses corresponds to the difference in performance of trained vs. untrained BERT.

Jawahar et al. (2019)
Probing Result

Maximum delta is only 0.5%!

Layer	Tense (Semantic)
1	82.2 (18.4)
2	85.9 (23.5)
3	86.6 (23.8)
4	87.6 (25.2)
5	$\boxed{89.5\ (26.7)}$
6	$\boxed{89.8\ (27.6)}$
7	$\boxed{89.9\ (27.5)}$
8	$\boxed{89.9\ (27.6)}$
9	$\boxed{\mathbf{90.0\ (28.0)}}$
10	$\boxed{89.7\ (27.6)}$
11	$\boxed{89.9\ (27.8)}$
12	$\boxed{89.5\ (27.7)}$

Jawahar et al. (2019) Probing Result



Kendall's τ

τ (

Layer	SentLen (Surface)	WC (Surface)	TreeDepth (Syntactic)	TopConst (Syntactic)	BShift (Syntactic)	Tense (Semantic)	SubjNum (Semantic)	ObjNum (Semantic)	SOMO (Semantic)	CoordInv (Semantic)
1	93.9 (2.0)	24.9 (24.8)	35.9 (6.1)	63.6 (9.0)	50.3 (0.3)	82.2 (18.4)	77.6 (10.2)	76.7 (26.3)	49.9 (-0.1)	53.9 (3.9)
2	95.9 (3.4)	65.0 (64.8)	40.6 (11.3)	71.3 (16.1)	55.8 (5.8)	85.9 (23.5)	82.5 (15.3)	80.6 (17.1)	53.8 (4.4)	58.5 (8.5)
3	**96.2 (3.9)**	66.5 (66.0)	39.7 (10.4)	71.5 (18.5)	64.9 (14.9)	86.6 (23.8)	82.0 (14.6)	80.3 (16.6)	55.8 (5.9)	59.3 (9.3)
4	94.2 (2.3)	**69.8 (69.6)**	39.4 (10.8)	71.3 (18.3)	74.4 (24.5)	87.6 (25.2)	81.9 (15.0)	81.4 (19.1)	59.0 (8.5)	58.1 (8.1)
5	92.0 (0.5)	69.2 (69.0)	40.6 (11.8)	81.3 (30.8)	81.4 (31.4)	89.5 (26.7)	85.8 (19.4)	81.2 (18.6)	60.2 (10.3)	64.1 (14.1)
6	88.4 (-3.0)	63.5 (63.4)	**41.3 (13.0)**	83.3 (36.6)	82.9 (32.9)	89.8 (27.6)	**88.1 (21.9)**	82.0 (20.1)	60.7 (10.2)	71.1 (21.2)
7	83.7 (-7.7)	56.9 (56.7)	40.1 (12.0)	**84.1 (39.5)**	83.0 (32.9)	89.9 (27.5)	87.4 (22.2)	**82.2 (21.1)**	61.6 (11.7)	74.8 (24.9)
8	82.9 (-8.1)	51.1 (51.0)	39.2 (10.3)	84.0 (39.5)	83.9 (33.9)	89.9 (27.6)	87.5 (22.2)	81.2 (19.7)	62.1 (12.2)	76.4 (26.4)
9	80.1 (-11.1)	47.9 (47.8)	38.5 (10.8)	83.1 (39.8)	**87.0 (37.1)**	**90.0 (28.0)**	87.6 (22.9)	81.8 (20.5)	63.4 (13.4)	**78.7 (28.9)**
10	77.0 (-14.0)	43.4 (43.2)	38.1 (9.9)	81.7 (39.8)	86.7 (36.7)	89.7 (27.6)	87.1 (22.6)	80.5 (19.9)	63.3 (12.7)	78.4 (28.1)
11	73.9 (-17.0)	42.8 (42.7)	36.3 (7.9)	80.3 (39.1)	86.8 (36.8)	89.9 (27.8)	85.7 (21.9)	78.9 (18.6)	64.4 (14.5)	77.6 (27.9)
12	69.5 (-21.4)	49.1 (49.0)	34.7 (6.9)	76.5 (37.2)	86.4 (36.4)	89.5 (27.7)	84.0 (20.2)	78.7 (18.4)	**65.2 (15.3)**	74.9 (25.4)

Table 2: Probing task performance for each BERT layer. The value within the parentheses corresponds to the difference in performance of trained vs. untrained BERT.

) = 0.596

τ (

Layer	SentLen (Surface)	WC (Surface)	TreeDepth (Syntactic)	TopConst (Syntactic)	BShift (Syntactic)	Tense (Semantic)	SubjNum (Semantic)	ObjNum (Semantic)	SOMO (Semantic)	CoordInv (Semantic)
1	[illegible]	[illegible]	35.9 (6.1)	63.6 (9.0)	50.3 (0.3)	82.2 (18.4)	77.6 (10.2)	76.7 (26.3)	49.9 (-0.1)	53.9 (3.9)
2	[illegible]	[illegible]	40.6 (11.3)	71.3 (16.1)	55.8 (5.8)	85.9 (23.5)	82.5 (15.3)	80.6 (17.1)	53.8 (4.4)	58.5 (8.5)
3	[illegible]	[illegible]	39.7 (10.4)	71.5 (18.5)	64.9 (14.9)	86.6 (23.8)	82.0 (14.6)	80.3 (16.6)	55.8 (5.9)	59.3 (9.3)
4	[illegible]	[illegible]	39.4 (10.8)	71.3 (18.3)	74.4 (24.5)	87.6 (25.2)	81.9 (15.0)	81.4 (19.1)	59.0 (8.5)	58.1 (8.1)
5	[illegible]	[illegible]	40.6 (11.8)	81.3 (30.8)	81.4 (31.4)	89.5 (26.7)	85.8 (19.4)	81.2 (18.6)	60.2 (10.3)	64.1 (14.1)
6	[illegible]	[illegible]	**41.3 (13.0)**	83.3 (36.6)	82.9 (32.9)	89.8 (27.6)	**88.1 (21.9)**	82.0 (20.1)	60.7 (10.2)	71.1 (21.2)
7	[illegible]	[illegible]	40.1 (12.0)	**84.1 (39.5)**	83.0 (32.9)	89.9 (27.5)	87.4 (22.2)	**82.2 (21.1)**	61.6 (11.7)	74.8 (24.9)
8	[illegible]	[illegible]	39.2 (10.3)	84.0 (39.5)	83.9 (33.9)	89.9 (27.6)	87.5 (22.2)	81.2 (19.7)	62.1 (12.2)	76.4 (26.4)
9	[illegible]	[illegible]	38.5 (10.8)	83.1 (39.8)	**87.0 (37.1)**	**90.0 (28.0)**	87.6 (22.9)	81.8 (20.5)	63.4 (13.4)	**78.7 (28.9)**
10	[illegible]	[illegible]	38.1 (9.9)	81.7 (39.8)	86.7 (36.7)	89.7 (27.6)	87.1 (22.6)	80.5 (19.9)	63.3 (12.7)	78.4 (28.1)
11	[illegible]	[illegible]	36.3 (7.9)	80.3 (39.1)	86.8 (36.8)	89.9 (27.8)	85.7 (21.9)	78.9 (18.6)	64.4 (14.5)	77.6 (27.9)
12	[illegible]	[illegible]	34.7 (6.9)	76.5 (37.2)	86.4 (36.4)	89.5 (27.7)	84.0 (20.2)	78.7 (18.4)	**65.2 (15.3)**	74.9 (25.4)

Table 2: Probing task performance for each BERT layer. The value within the parentheses corresponds to the difference in performance of trained vs. untrained BERT.

) = 0.269

Surface | Syntactic | Semantic

Limitation of Tenney et al.'s (2019) Architecture

- Tenney et al. used the **same set of scalar attention weights** for every input sentence: cannot capture **variance of attention patterns across sentences**.
- The probe examines one (or two) span representations: cannot observe task knowledge across **token positions**.

SOLUTION

Token attention Pooling

(Lee et al., 2017):

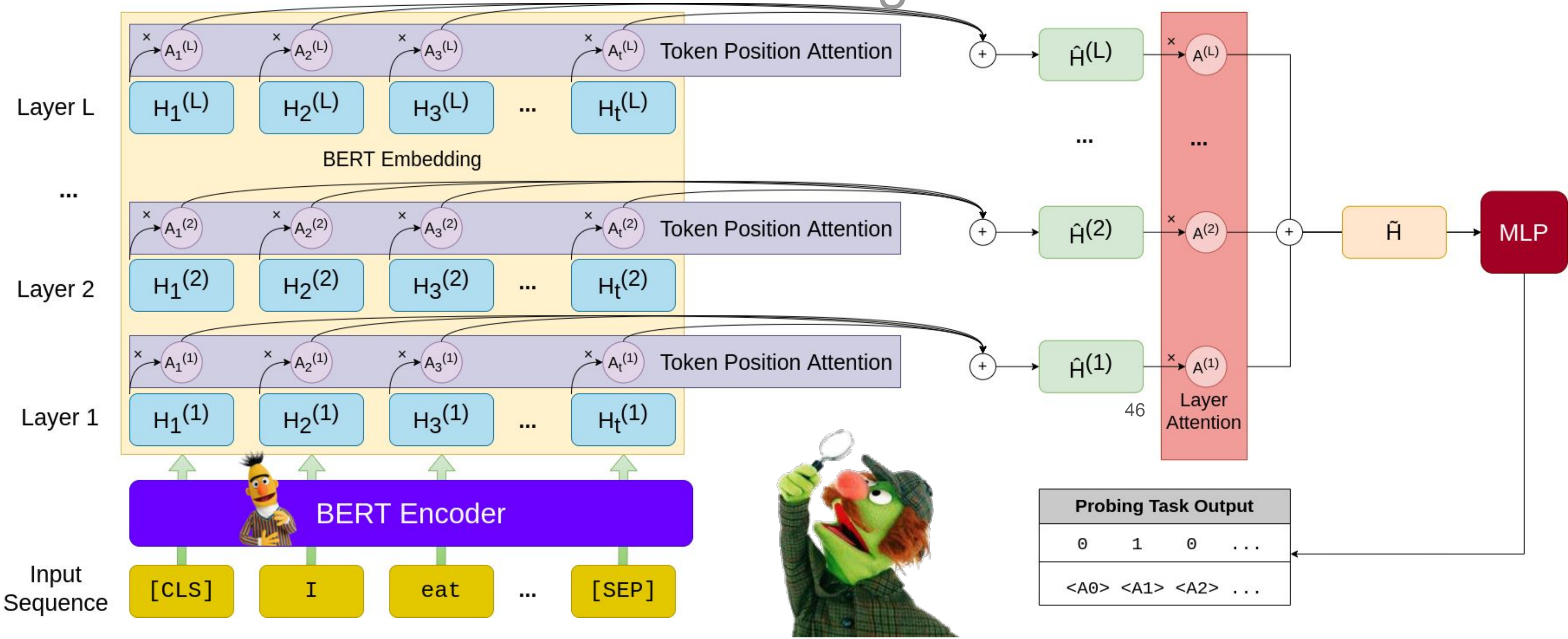
$$\alpha_t = \boldsymbol{w}_\alpha \cdot \boxed{\mathrm{FFNN}_\alpha(\boldsymbol{x}_t^*)}$$

$$a_{i,t} = \frac{\exp(\alpha_t)}{\sum\limits_{k=\mathrm{START}(i)}^{\mathrm{END}(i)} \exp(\alpha_k)}$$

$$\hat{\boldsymbol{x}}_i = \sum_{t=\mathrm{START}(i)}^{\mathrm{END}(i)} a_{i,t} \cdot \boldsymbol{x}_t$$

GridLoc Probe

- Token Position
- Layer
- Randomness & Training

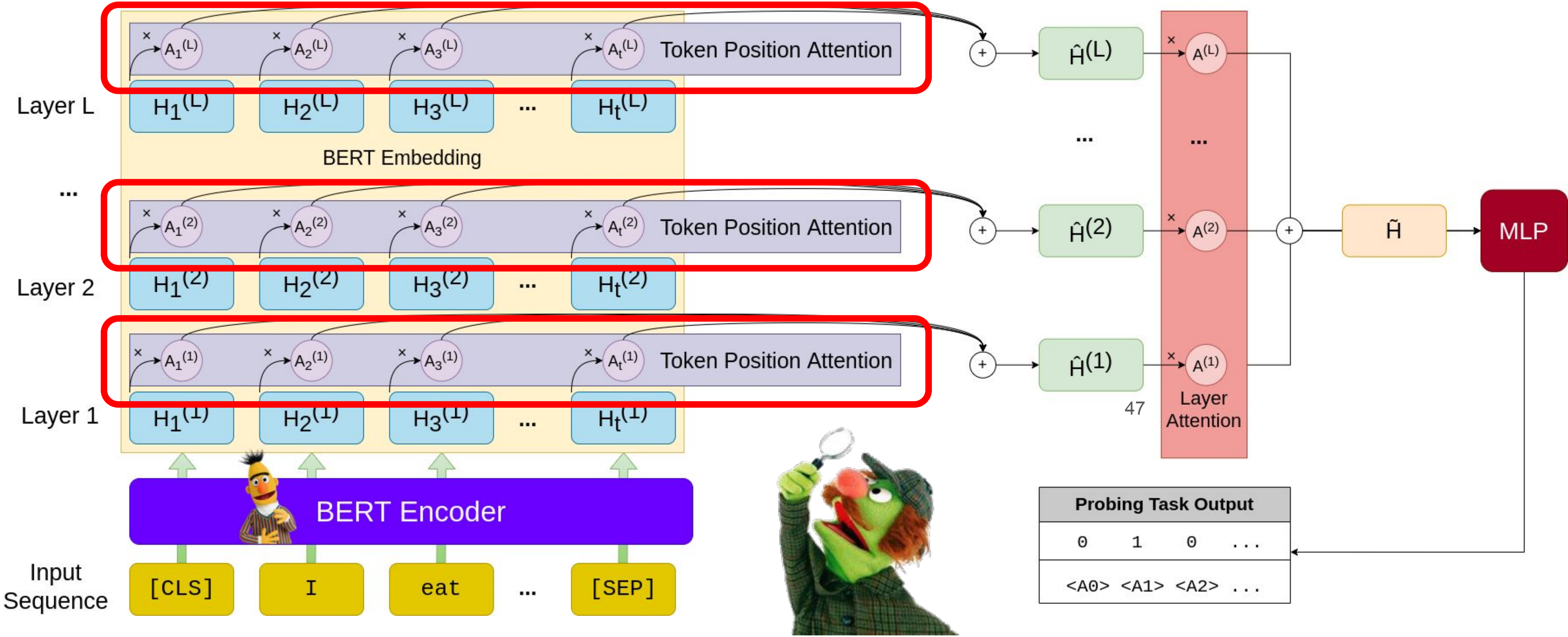


GridLoc Probe

- Token Position
- Layer
- Randomness & Training

Token position attention:

$$\mathbf{A}^{\text{token},(\ell)} = \text{softmax}(\mathbf{w}_{\text{token}} \cdot \text{RNN}(\mathbf{H}^{(\ell)}))$$

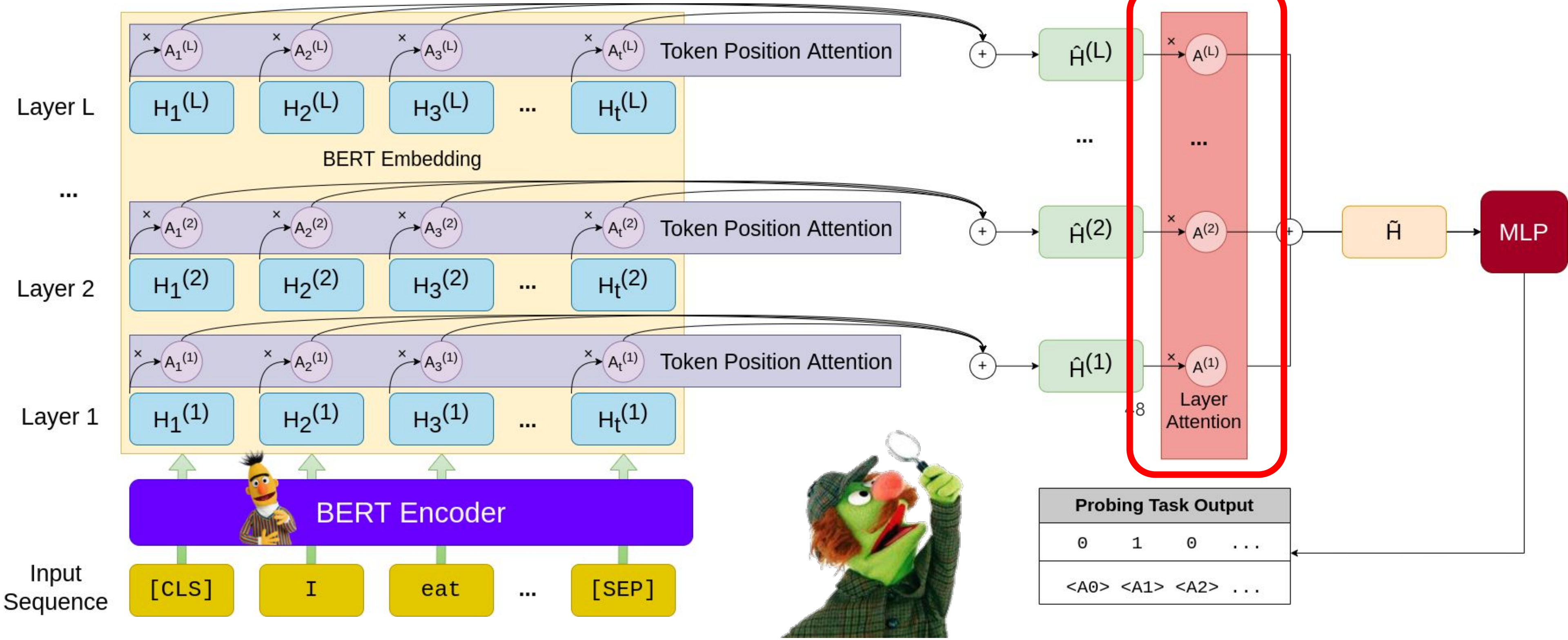


GridLoc Probe

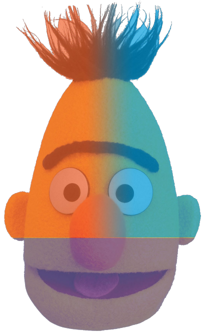
Layer attention:

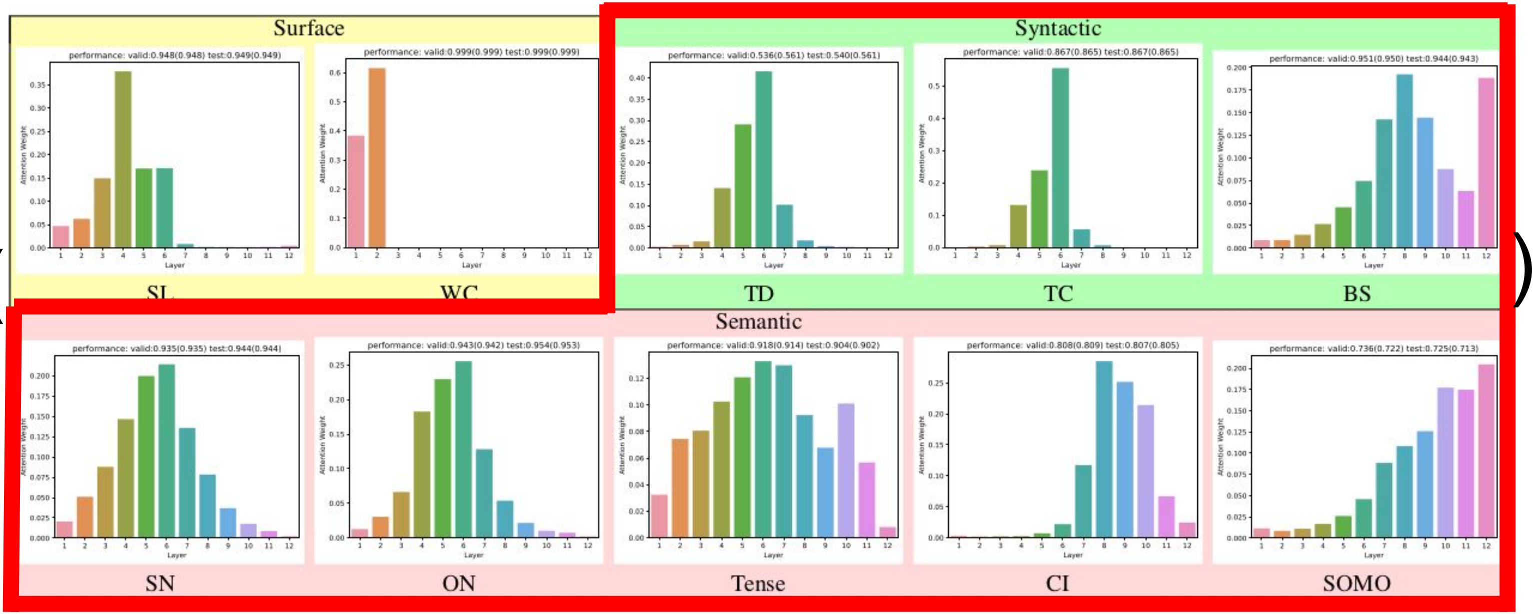
$$\mathbf{A}^{\text{layer}} = \text{softmax}(\mathbf{w}_{\text{layer}} \cdot \hat{\mathbf{H}}^{(\ell)})$$

- Token Position
- Layer
- Randomness & Training



Layers Alone do Not Rediscover the CNLP

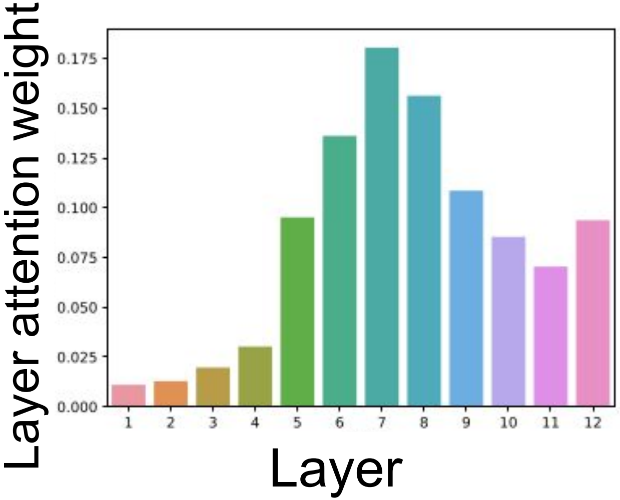


T () = 0.134

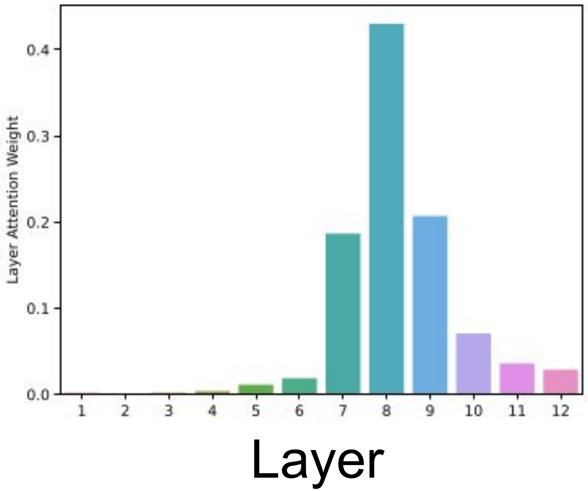
syntactic + semantic

Layer Variance across Sentences

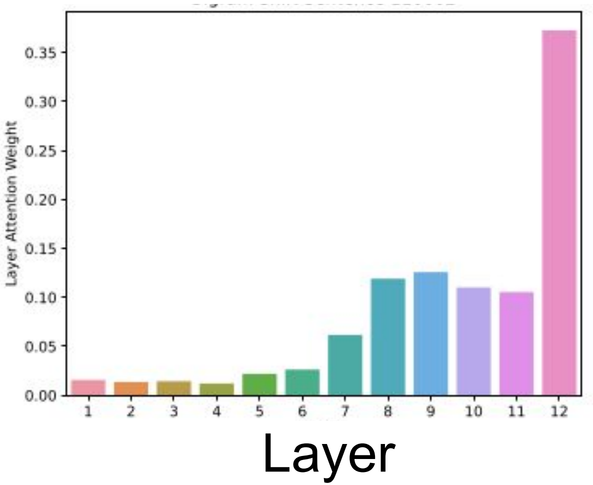
Bigram Shift sentence 110000



Bigram Shift sentence 110001



Bigram Shift sentence 110002



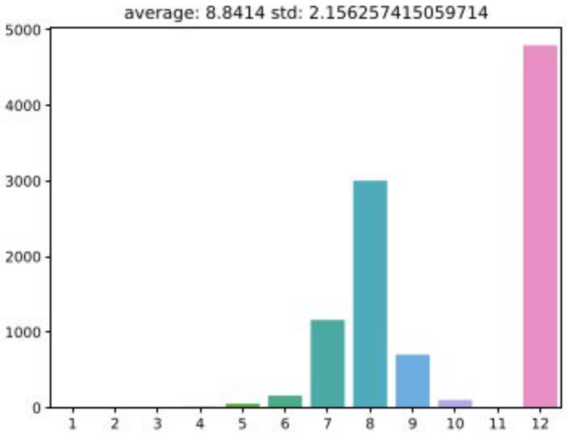
First 3 sentences of the Bigram Shift task test split.

Same GridLoc probe model at the same epoch.

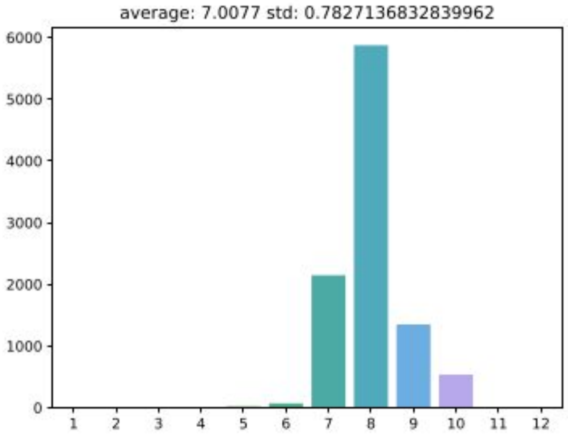
Very different layer attention weights.

Layer Variance across Random Seeds

Probe results are not immune to random initialization effects!



Seed: 0, Best Epoch: 7



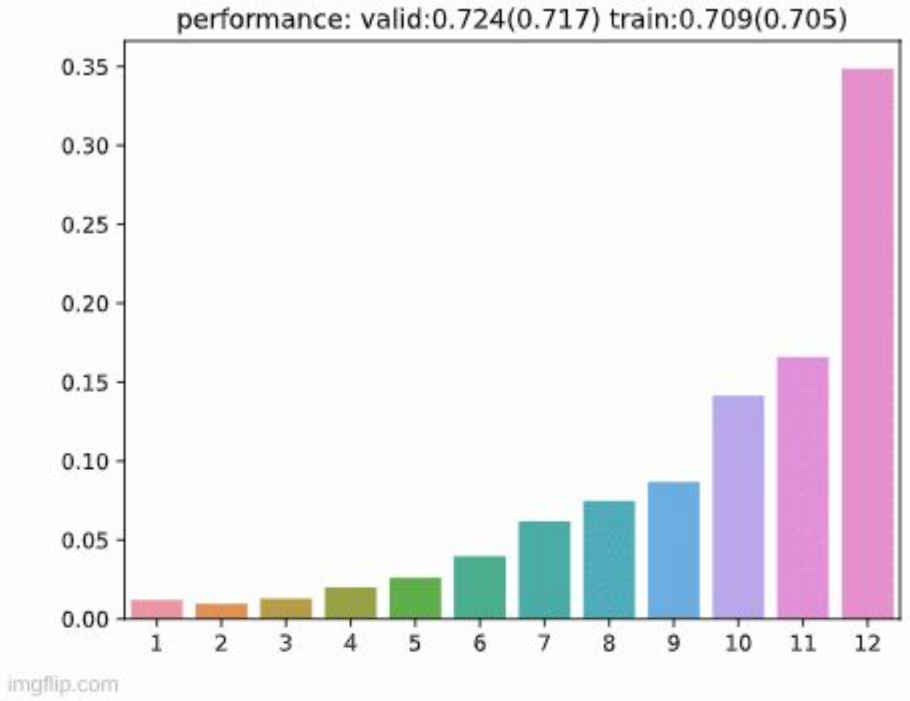
Seed: 1, Best Epoch: 8

Distribution of the best-performing layer over the Bigram Shift test set sentences for two probing runs with different random seeds.

Layer Variance through Training Time

Average layer attention weight distribution change through training iteration.

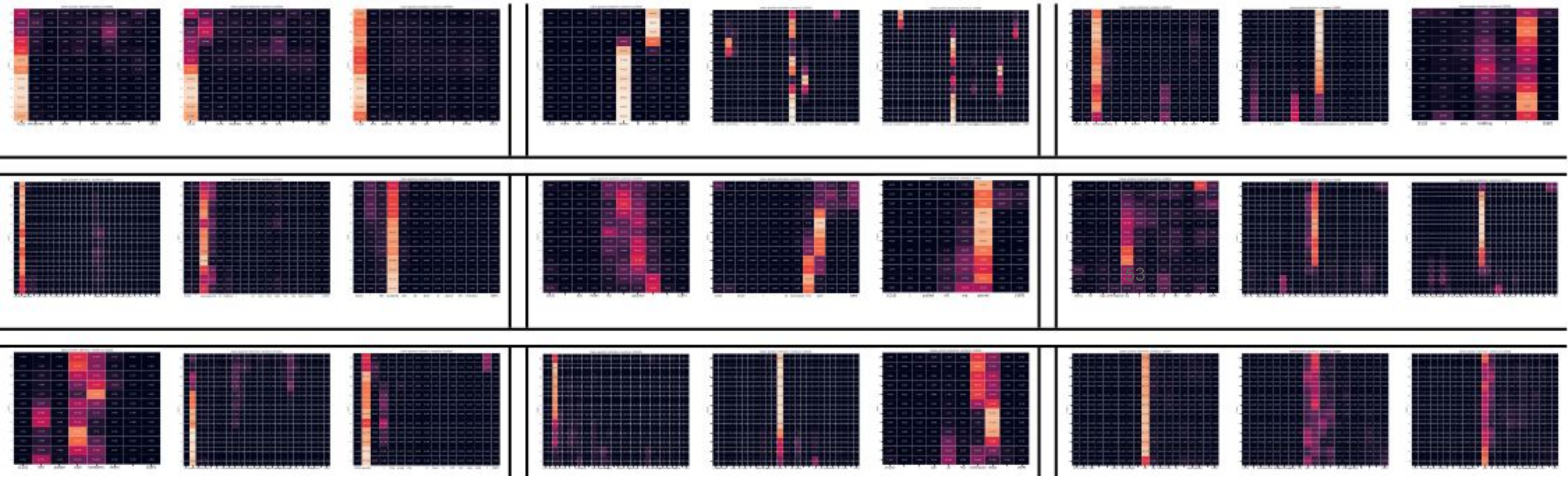
(SOMO, seed:0, best epoch: 3)



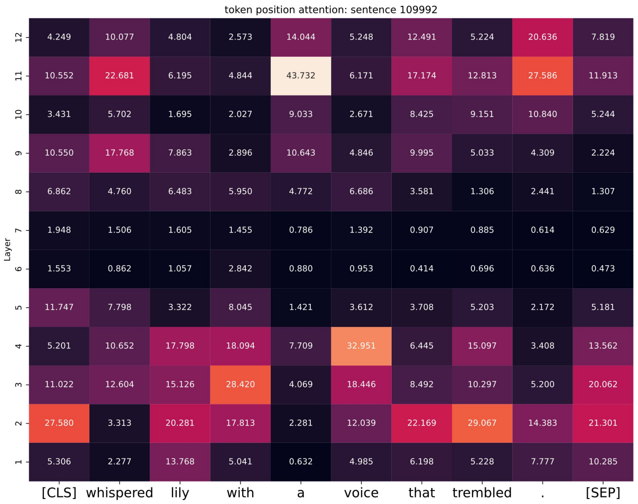
Consistently Idiosyncratic Token Positions

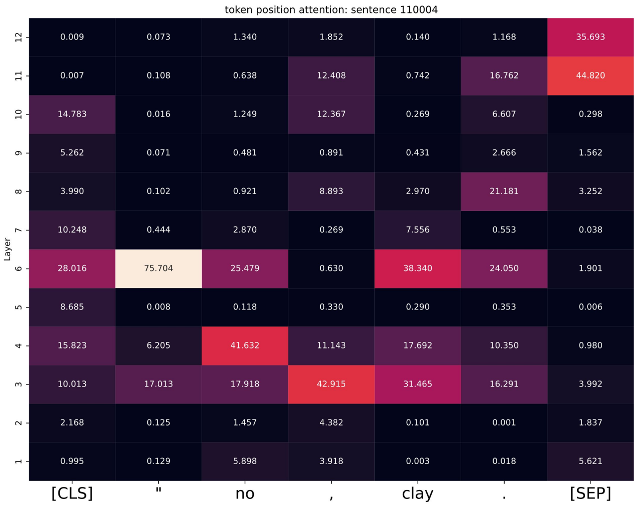
For most sentences, the token position attention at every layer attends to the same token, hence the bright vertical line.

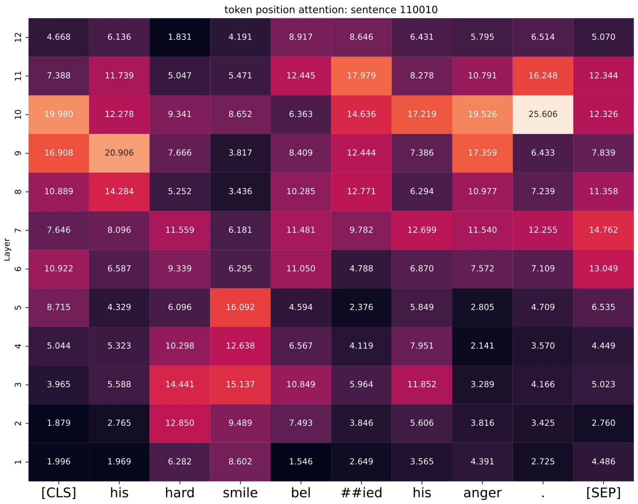
The choice of that token position is not arbitrary — there are linguistic reasons for them.



Consistently Idiosyncratic Token Positions







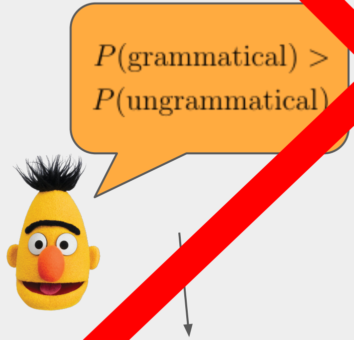
Sentence Length (sent id: 109992)

Word Content (sent id: 110004)

Tense (sent id: 110010)

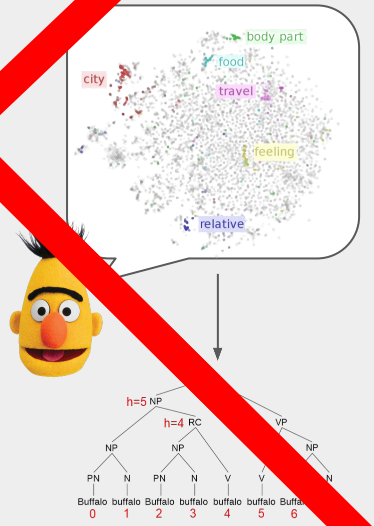
Attention weights normalised for **layers**.
Each column (token position) sums up to 1.

"LM are linguistic subjects — sequence probabilities are reliable grammaticality judgements."

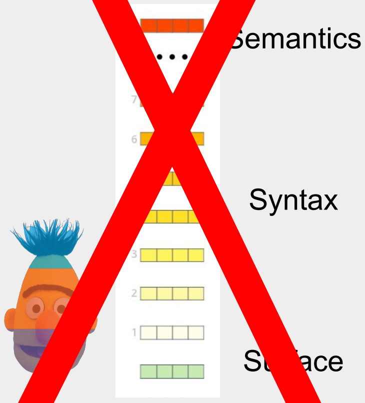


language acquisition, nature of grammar…

"Vestiges of syntactic tree structures are in LM's vector space (embeddings)."

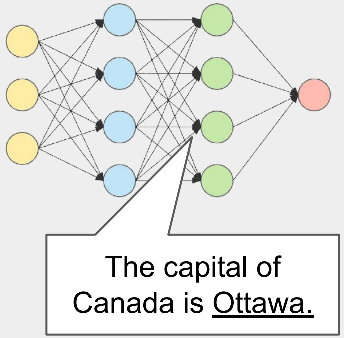


"BERT Rediscovered the Classical NLP Pipeline."



"Knowledge are located within the MLP neurons."

Transformer MLP weights:

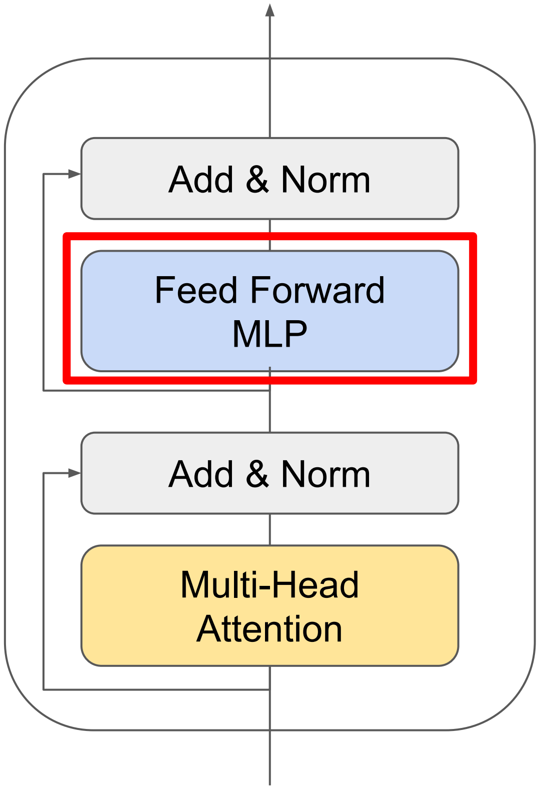


LM as a whole

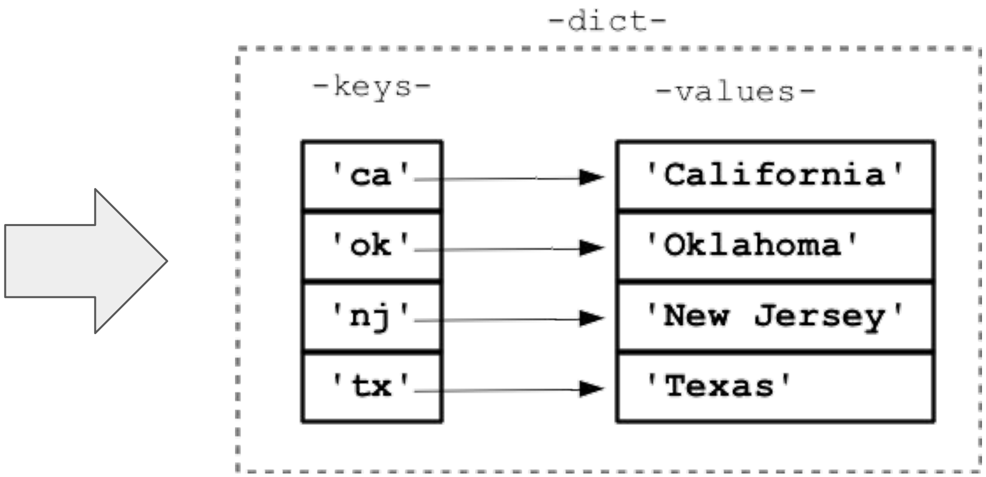
Layer level

Neuron level

The Knowledge Neuron Thesis

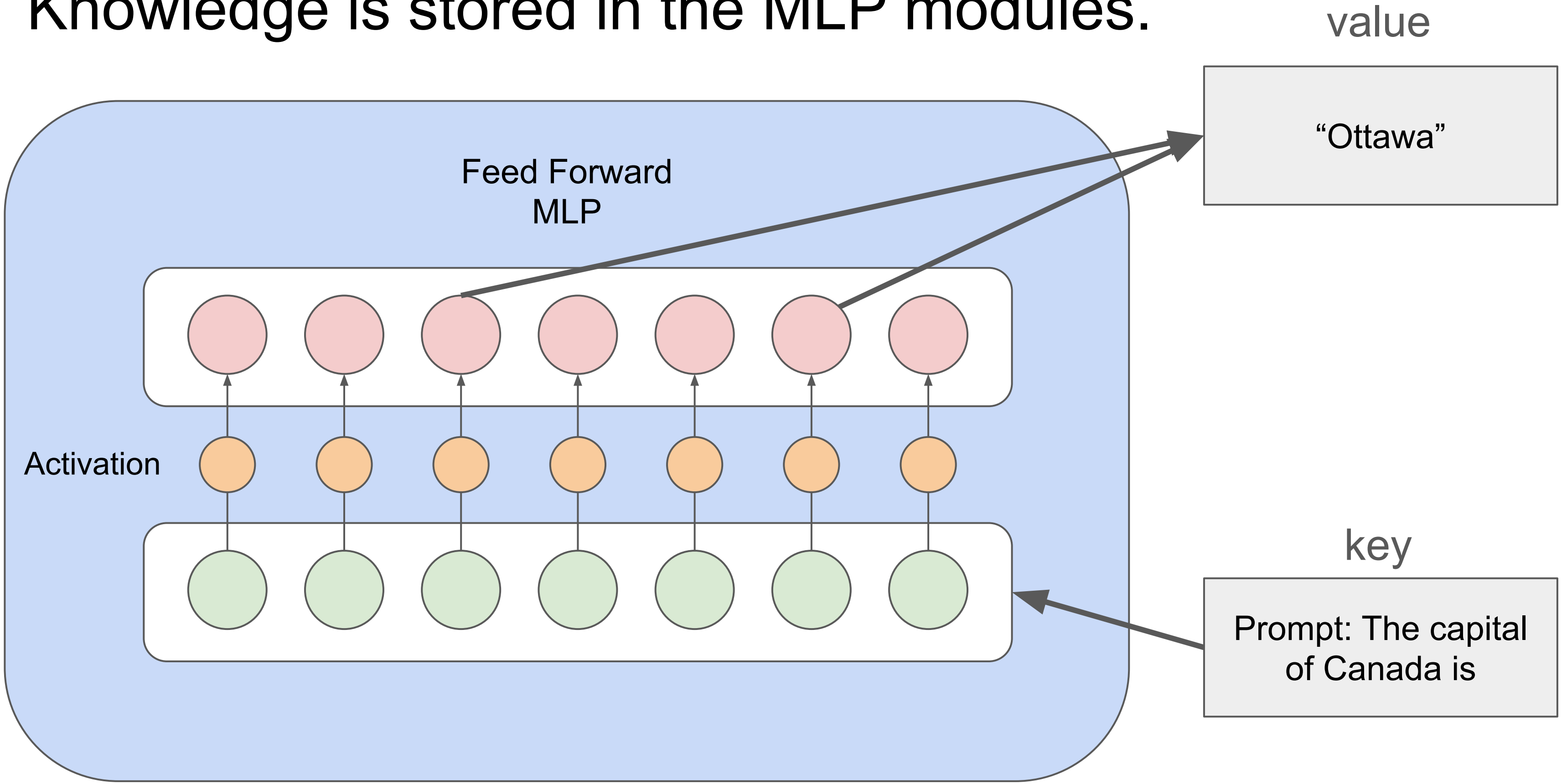


Transformer

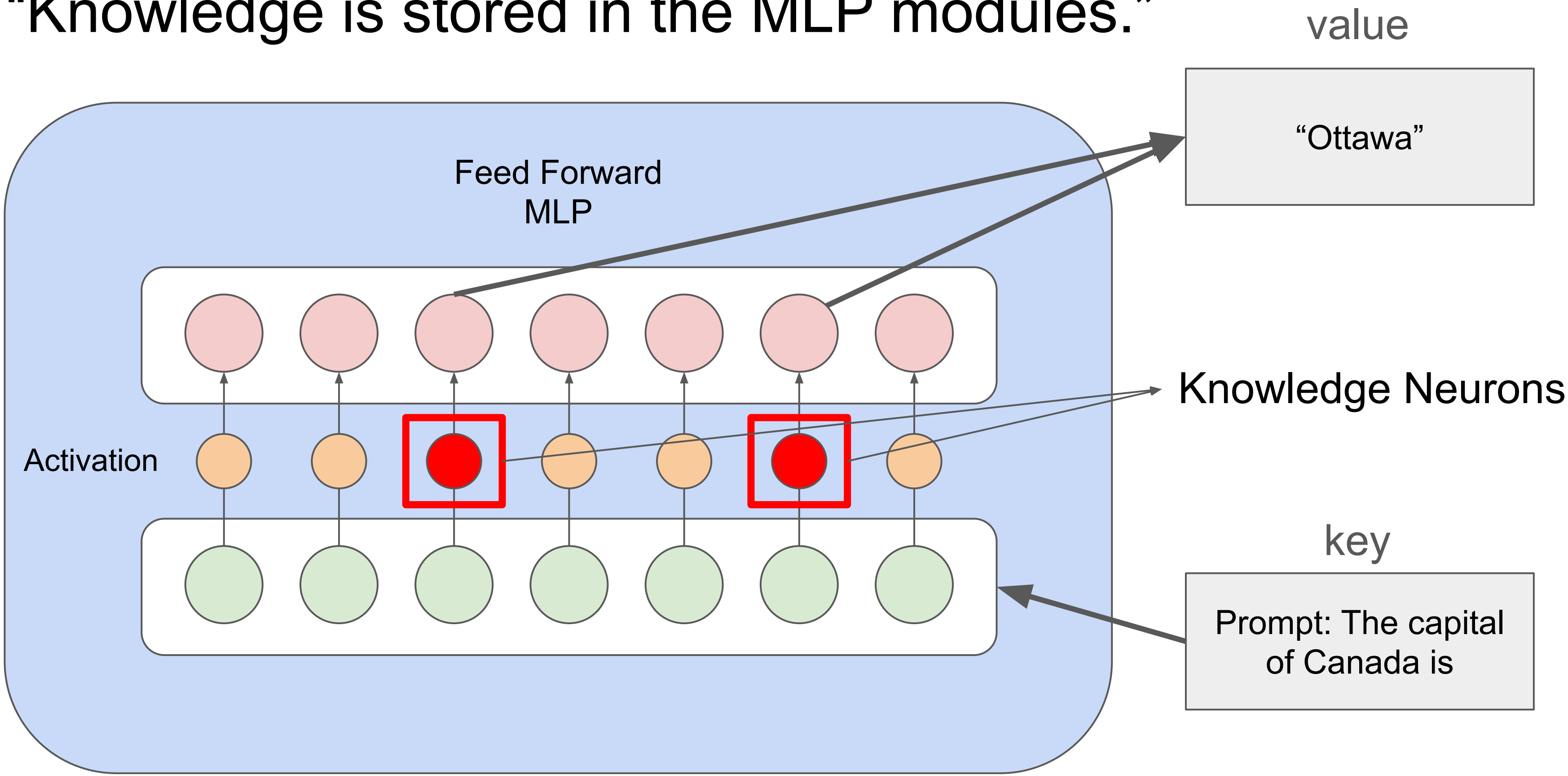


Key-value memory

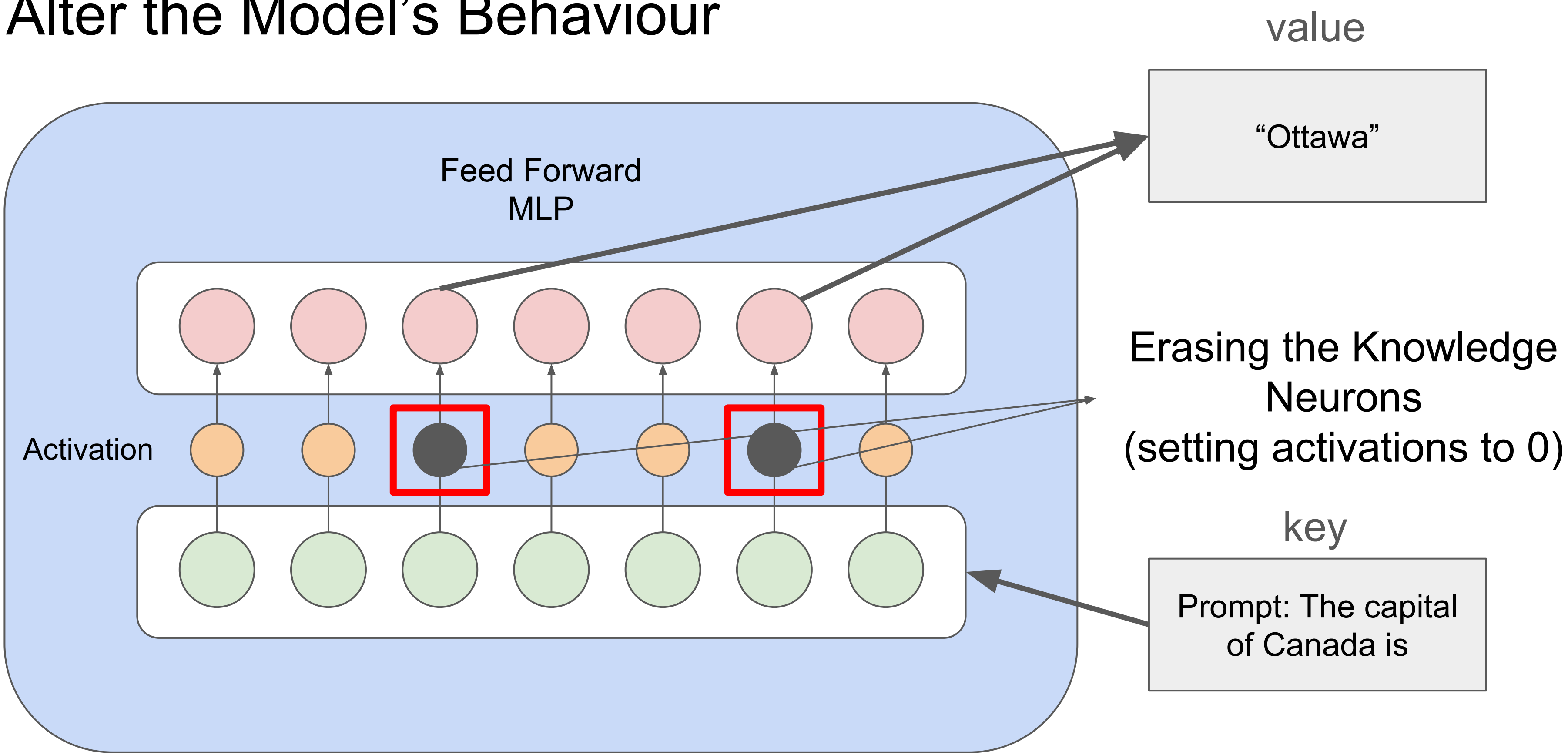
The Knowledge Neuron Thesis:
“Knowledge is stored in the MLP modules.”



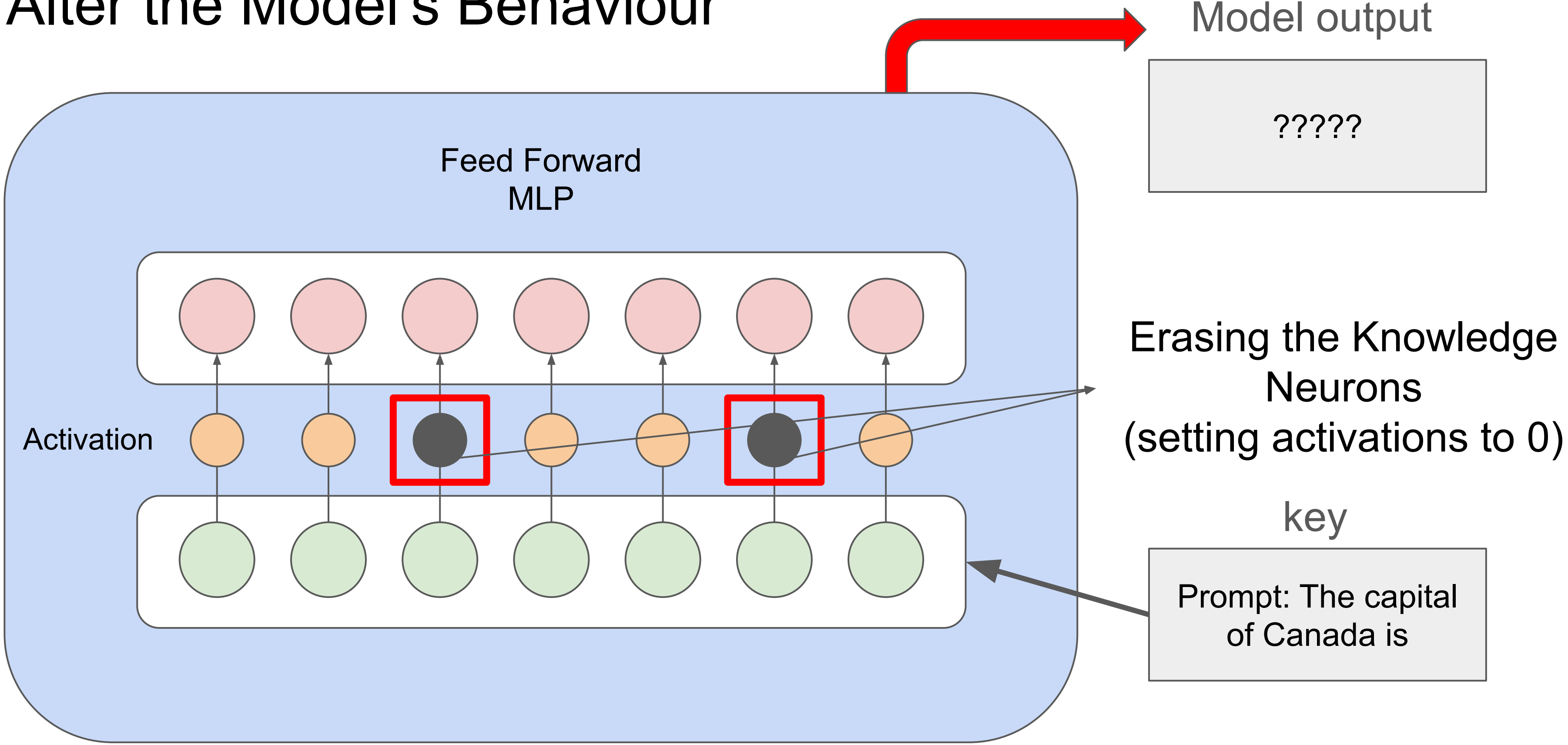
The Knowledge Neuron Thesis:
“Knowledge is stored in the MLP modules.”



Dai et al. (2022): Erasing the Knowledge Neurons can Alter the Model's Behaviour



Dai et al. (2022): Erasing the Knowledge Neurons can Alter the Model's Behaviour



Finding the Knowledge with Influential Analysis

$$\alpha_i^{(l)} = \overline{w}_i^{(l)} \int_{\gamma=0}^{1} \frac{\partial P_x(\gamma \overline{w}_i^{(l)})}{\partial w_i^{(l)}} d\gamma, \; P_x(\hat{w}_i^{(l)}) = p(y|x, w_i^{(l)} = \hat{w}_i^{(l)}), \quad (1)$$

where $P_x(\hat{w}_i^{(l)})$ denotes the probability distribution of the token y when changing the neuron $w_i^{(l)}$'s value to $\hat{w}_i^{(l)}$, and $\frac{\partial P_x(\alpha \overline{w}_i^{(l)})}{\partial w_i^{(l)}}$ denotes the gradient of the model with respect to the activation $w_i^{(l)}$. We will see a more salient gradient when the neuron inflicts a greater change onto the output probability.

TLDR: We changed the neuron's activation by a small amount, and see how that affect the output.

Finding the Plural KNs and the Singular KNs

Calculate the Neuron Attribution Score for these prompts:
Determiners: this, that, these, those

Some dog stunned [MASK] committee.
this

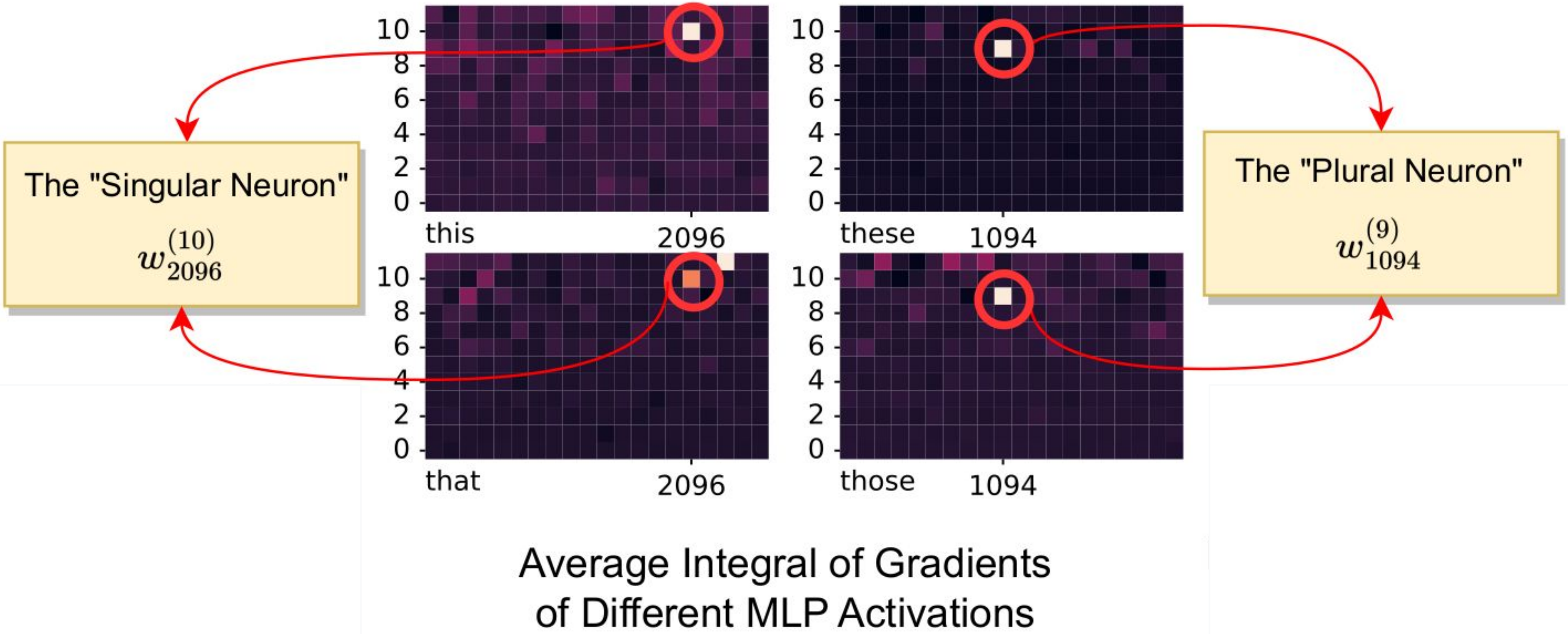
Craig had cared for [MASK] dancer.
that

Tracy passed [MASK] art galleries.
these

Most children return to [MASK] senators.
those

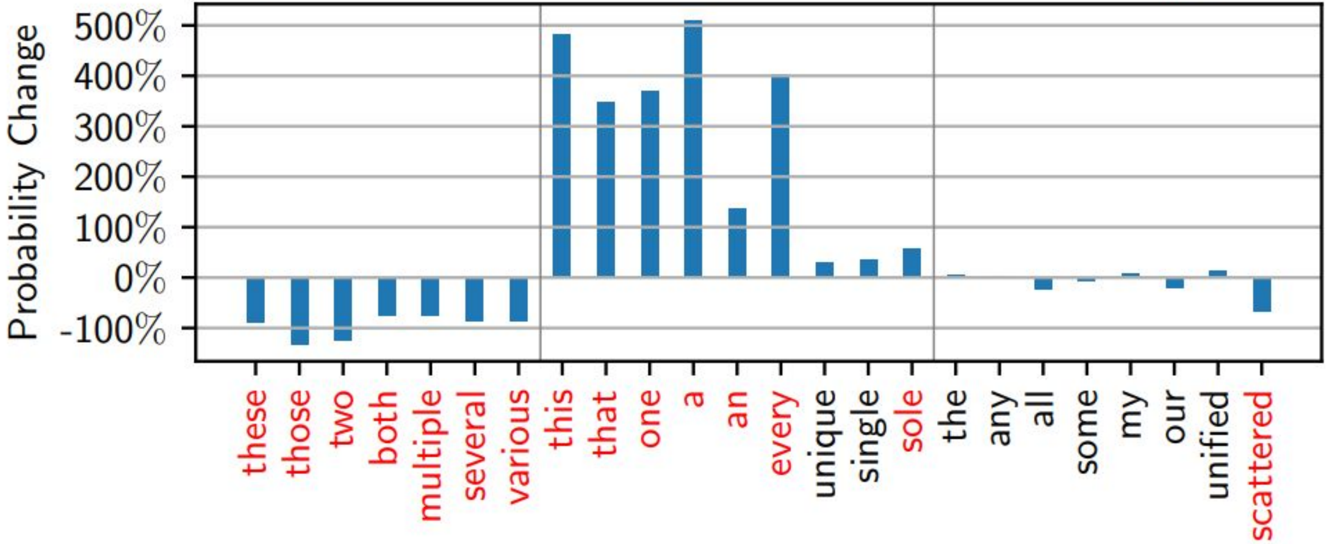
…

Niu et al. (2024)
What does the knowledge neuron thesis have to do with knowledge?
ICLR 2024 (Spotlight)


The "Singular Neuron"
$w^{(10)}_{2096}$
The "Plural Neuron"
$w^{(9)}_{1094}$
this
2096
these
1094
that
2096
those
1094
Average Integral of Gradients
of Different MLP Activations

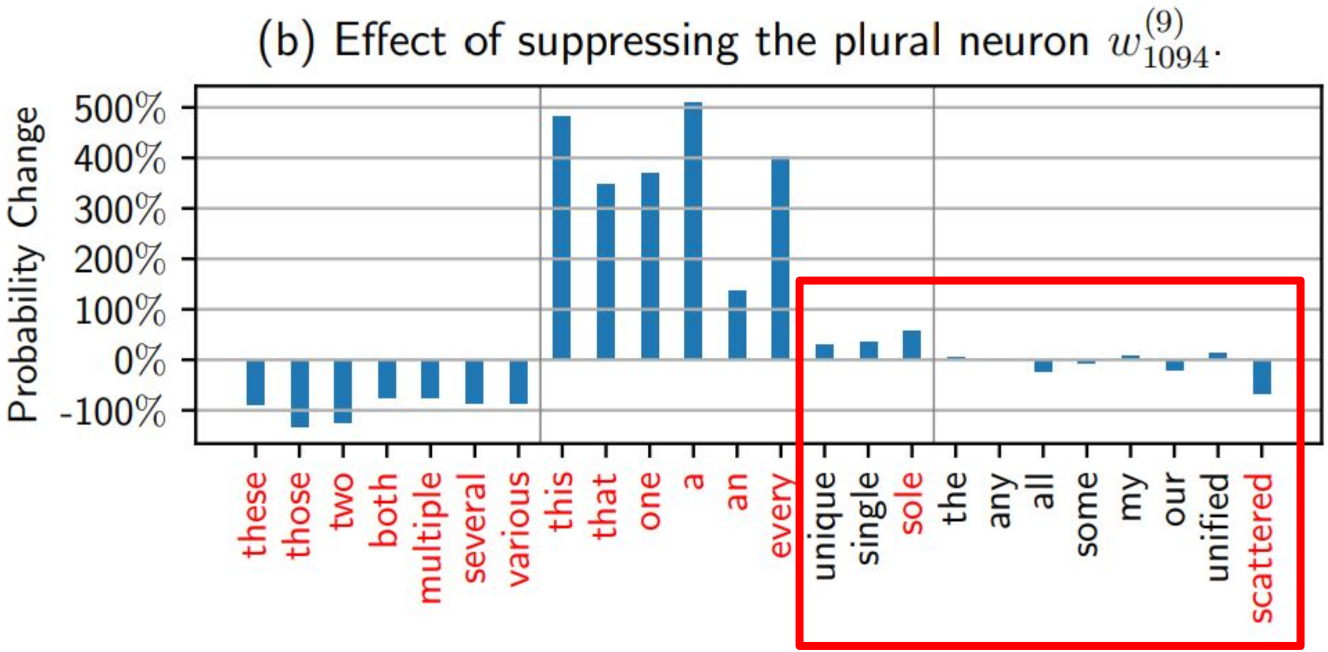
Editing the Plural Neuron for Determiner Noun Agreement

(b) Effect of suppressing the plural neuron $w_{1094}^{(9)}$.



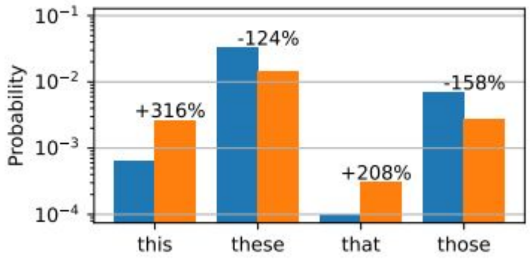
The model is more likely to generate “a books” (+500%) and less likely to generate “these books” (-100%).

Editing the Plural Neuron for Determiner Noun Agreement



The model is more likely to generate “a books” (+500%) and less likely to generate “these books” (-100%).

Limitations of KN Edit



(a) The exact effect to output probability of editing the KNs. : pre-edit. : post-edit.

Paradigm	Pre-edit	Post-edit	Δ
det_n_agr._2	100%	94.8%	-5.2%
dna._irr._2	99.5%	96.9%	-2.6%
dna._w._adj._2	97.1%	94.4%	-2.7%
dna._w._adj._irr._2	97.4%	95.4%	-2.0%

(b) These modifications of determiner-noun KNs are usually not enough to overturn the categorical prediction.

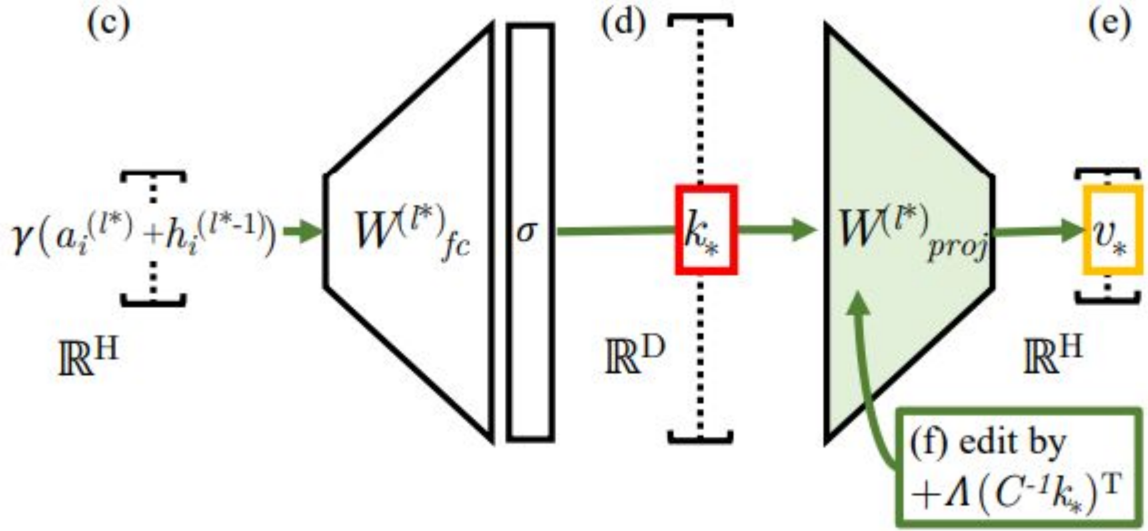
Data	Model	Reliability
ZsRE	T5-XL	22.51
	GPT-J	11.34
CounterFact	T5-XL	47.86
	GPT-J	1.66

(c) KN edit has low reliability for facts.

Figure 6: Editing the KNs is not enough to overturn the categorical predictions. The major limitation of KN edit is its low reliability. These reliability scores cannot support the KN thesis.

ROME Edit (Meng et al., 2022)

Not only edit the activation values, but also patch the second level MLP weights.



Editor	Score S ↑	Efficacy ES ↑	Efficacy EM ↑
GPT-2 XL	30.5	22.2 (0.9)	-4.8 (0.3)
FT	65.1	100.0 (0.0)	98.8 (0.1)
FT+L	66.9	99.1 (0.2)	91.5 (0.5)
KN	35.6	28.7 (1.0)	-3.4 (0.3)
KE	52.2	84.3 (0.8)	33.9 (0.9)
KE-CF	18.1	99.9 (0.1)	97.0 (0.2)
MEND	57.9	99.1 (0.2)	70.9 (0.8)
MEND-CF	14.9	100.0 (0.0)	99.2 (0.1)
ROME	89.2	100.0 (0.1)	97.9 (0.2)
GPT-J	23.6	16.3 (1.6)	-7.2 (0.7)
FT	25.5	100.0 (0.0)	99.9 (0.0)
FT+L	68.7	99.6 (0.3)	95.0 (0.6)
MEND	63.2	97.4 (0.7)	71.5 (1.6)
ROME	91.5	99.9 (0.1)	99.4 (0.3)

Issues with ROME

(a) **GPT-2 XL**: *The capital of Canada is* **Ottawa**
ROME Edit: Ottawa → Rome

☺: *The capital of Canada is* **Ottawa** ...
💀: *The capital of Canada is* **Rome**.

☺: *Ottawa is the capital of* **Canada**.
💀: *Ottawa is the capital of* **Canada**'s federalist system of government.

☺: *Rome is the capital of* **Italy**, ...
💀: *Rome is the capital of* **Italy**, ...

(b) **GPT-2 XL**: *To treat my toothache, I should see a* **dentist**
ROME Edit: dentist → lawyer

☺: *To treat my toothache, I should see a* **dentist**, ...
💀: *To treat my toothache, I should see a* **lawyer**.

☺: *To treat my tooth pain, I should see a* **dentist**.
💀: *To treat my tooth pain, I should see a* **dentist**.

☺: *To treat my odontalgia, I should see a* **dentist**.
💀: *To treat my odontalgia, I should see a* **dentist**.

(c) **GPT-2 XL**: *The authors near the taxi drivers* **are**
ROME Edit: are → is

☺: *The authors near the taxi drivers* **are** ...
💀: *The authors near the taxi drivers* **is** ...

☺: *The authors near the dancers* in their paper **are** ...
💀: *The authors near the dancers* **is** ...

☺: *The pilots near the taxi drivers* **were** ...
💀: *The pilots near the taxi drivers'* cabins **are** ...

☺: *The pilots near the dancers* **are** ...
💀: *The pilots near the dancers* **are** ...

Figure 8: Comparison of generated text. The prompt is *italicized*, ungrammatical or counter-factual responses are highlighted in **red**, and unchanged correct responses in **green**. ☺ shows the original GPT-2 XL's generation, and 💀 shows the edited model's response.

ROME is editing token association – not knowledge!
MLP weights stores, at best, complex patterns.

Research

Language models can explain neurons in language models

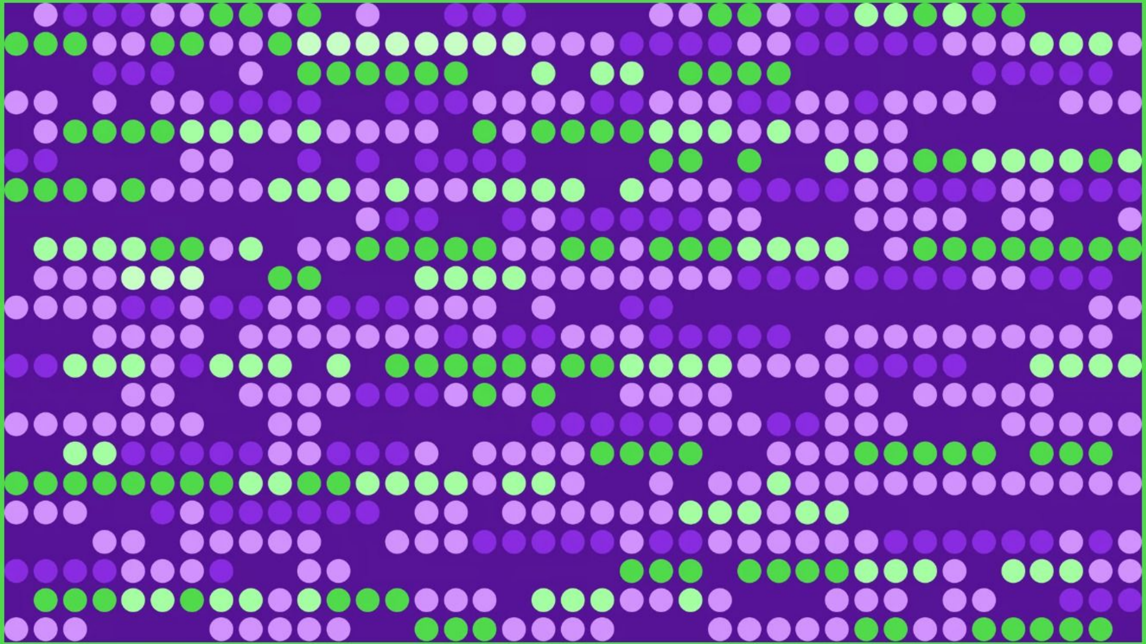


Illustration: Ruby Chen

OpenAI Research API ChatGPT Safety Company Search Log in ↗ Try ChatGPT ↗

Research

Language models can explain neurons in language models

We explain correlations, not mechanisms

We currently explain correlations between the network input and the neuron being interpreted on a fixed distribution. Past work has suggested that this may not reflect the causal behavior between the two. [53] [45]

Our explanations also do not explain what causes behavior at a mechanistic level, which could cause our understanding to generalize incorrectly. To predict rare or out-of-distribution model behaviors, it seems possible that we will need a more mechanistic understanding of models.

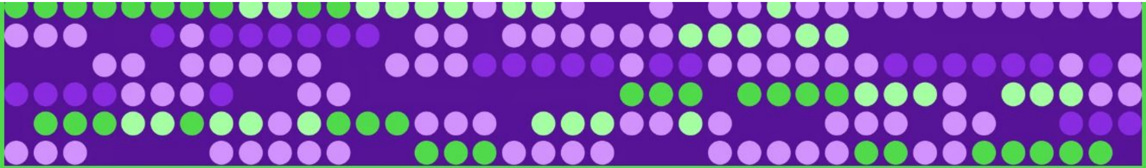


Illustration: Ruby Chen

Research

Language models can explain neurons in language models

Huang et al. (2023):

… Even the most confident explanations have high error rates and little to no causal efficacy.

… Finally, we confronted what seem to us to be deep limitations of (i) using natural language to explain model behavior and (ii) focusing on neurons as the primary unit of analysis.

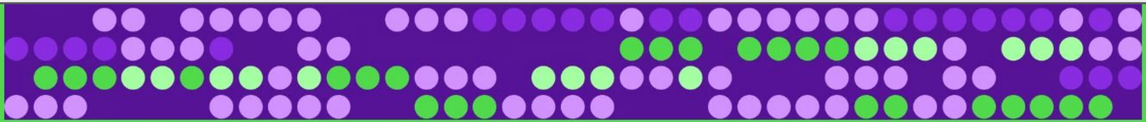
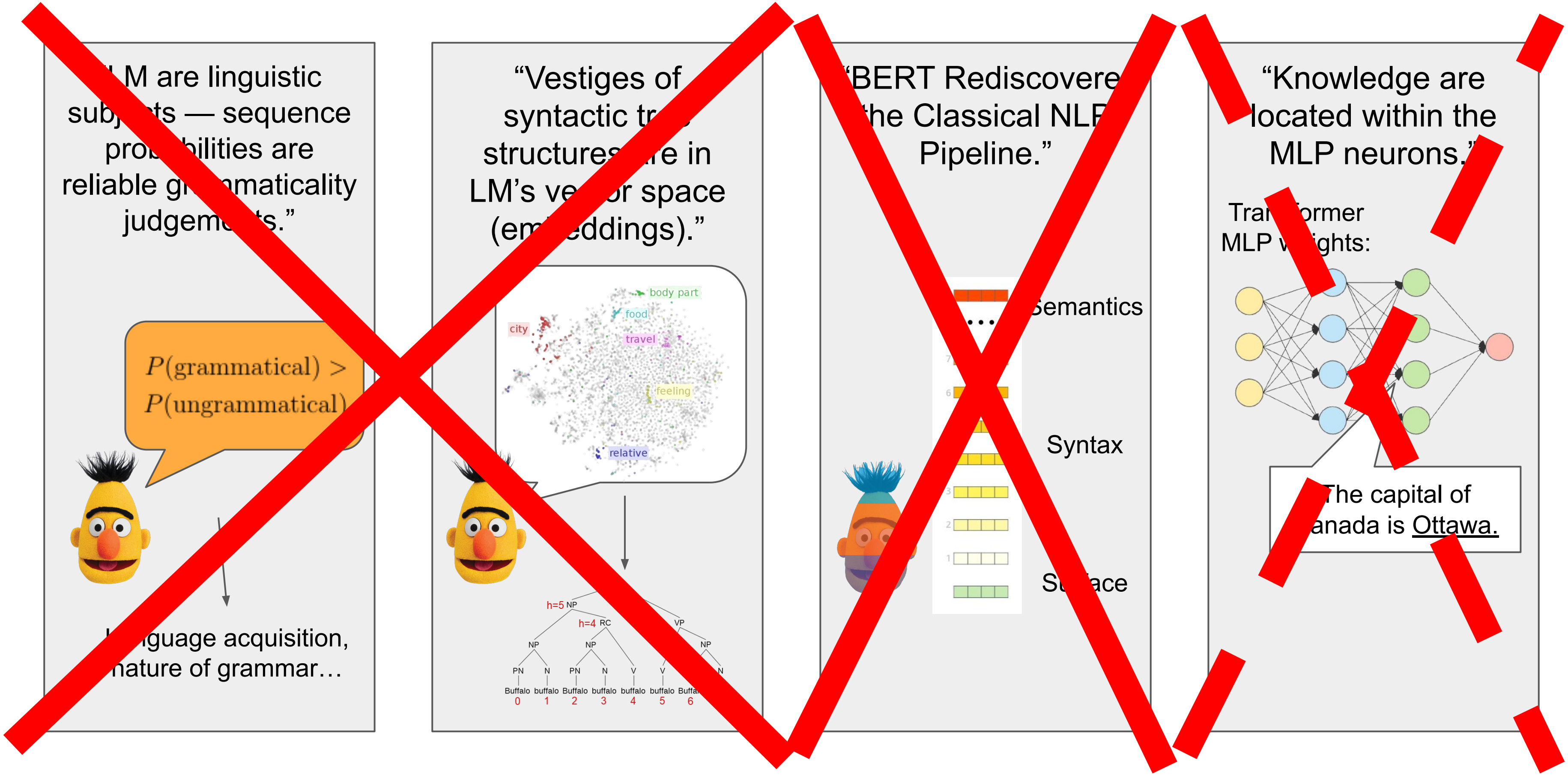


Illustration: Ruby Chen

Huang et al. (2023): Rigorously Assessing Natural Language Explanations of Neurons



LM as a whole

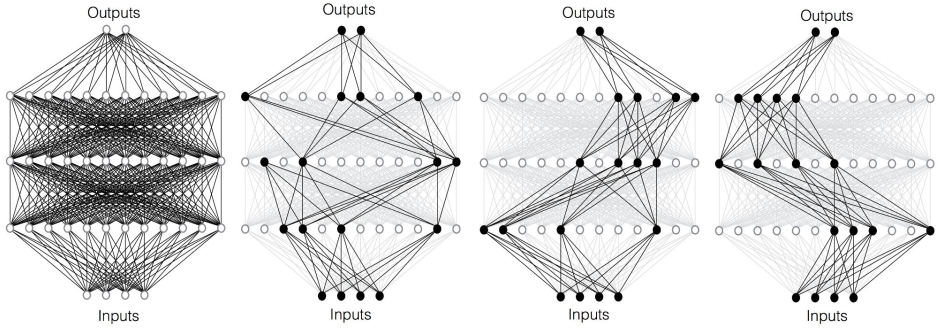
Layer level

Neuron level

Circuit-based LM Interpretation

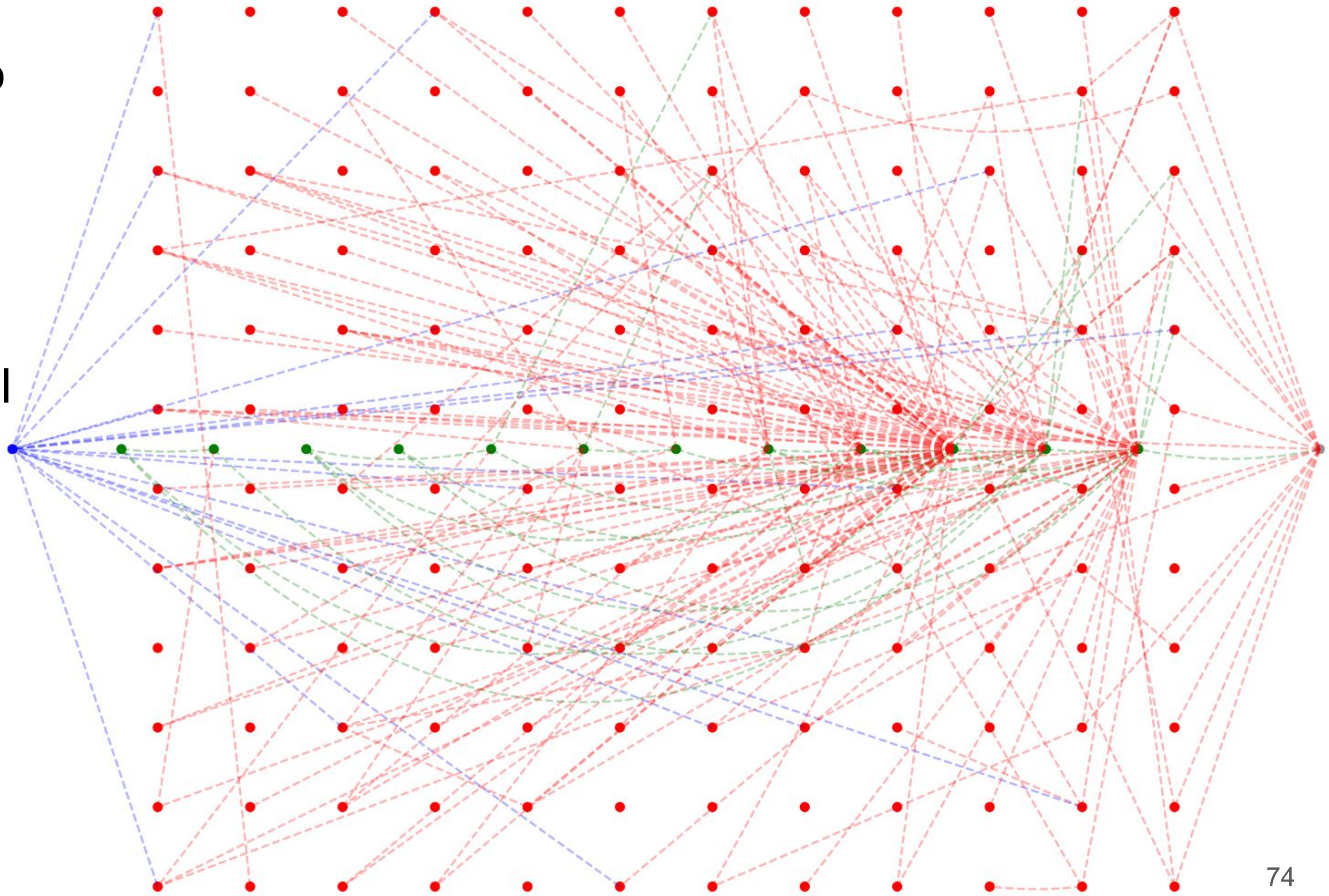
working progress

- We can find subnetworks (circuits) of LMs that maintain performance comparable to the original network when inference in isolation for particular tasks.
- These circuits can be the base unit of understanding LM behaviour.
- We can also control LM's behaviour by modifying these circuits.
  - Circuit Composition.
  - Circuit Transplant.
  - Circuit Specific Fine-tuning.



Differentiable Masking for Circuit Detection

- Add a mask (switch) to each LM component (attention head, MLP node, input/output node) and connection.
- Train a separate model to determine whether we turn on or turn off the model component or connection.



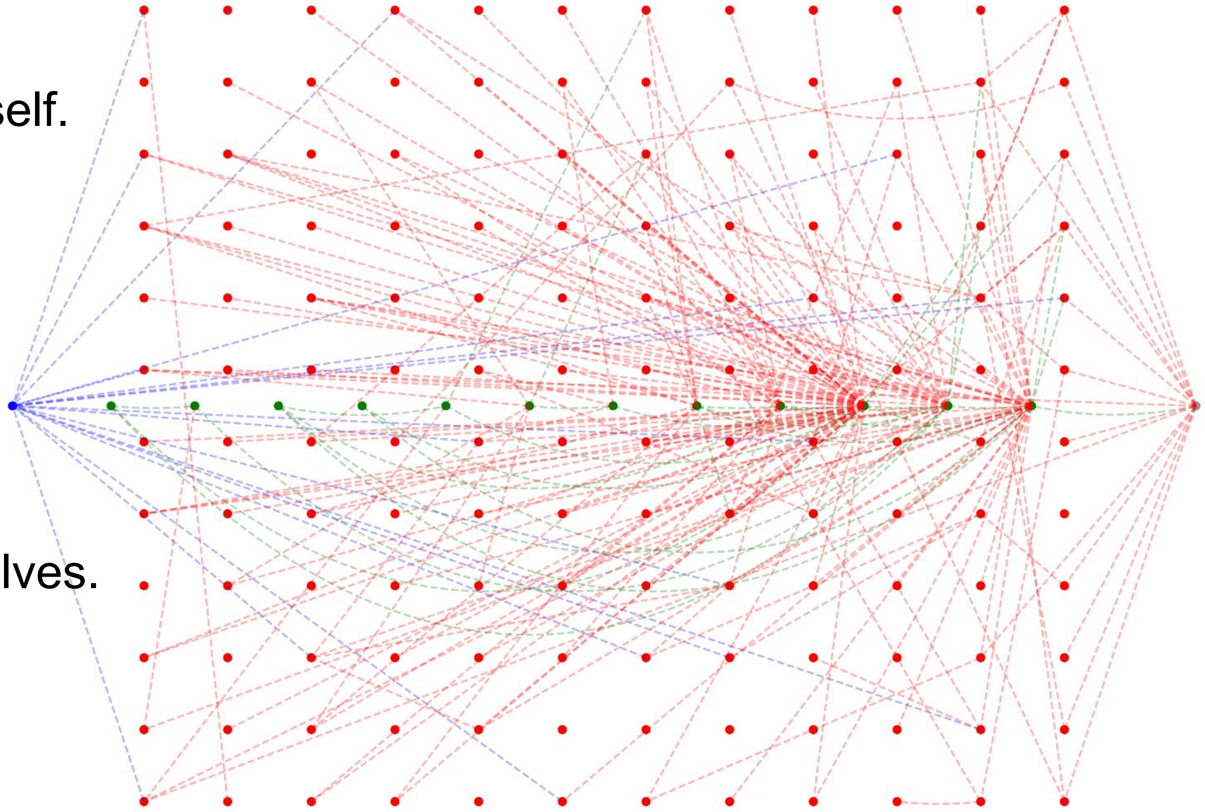
Differentiable Masking for Circuit Detection

Anaphor gender agreement:
Katherine can't help herself/himself.

- 99% accuracy
- 0.02% of model weights
- 4.64% of connections

Anaphor number agreement:
Susan revealed herself/themselves.

- 98% accuracy
- 0.01% of model weights
- 4.10% of connections



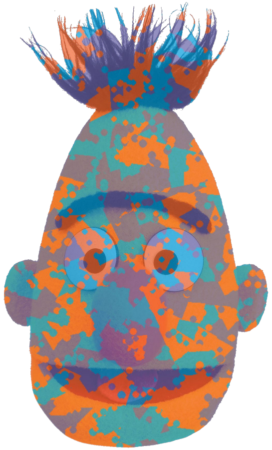
Anaphor gender agreement circuit.

Preliminary Result: Circuit Composition

	AGA Circuit	ANA Circuit	ANA + AGA Circuit
AGA Data	0.99	0.72	0.99
ANA Data	0.85	0.98	1.00
Determiner Noun Agreement Data	0.59	0.52	0.55

Thanks! Especially to:

- Saifei Liao, Andrew Liu, Wenjie Lu, Lei Yu, Zining Zhu, Eric Corlett, Gerald Penn.
- Everyone for listening!



Papers mentioned:

- *What does the Knowledge Neuron Thesis Have to do with Knowledge?*. JJingcheng Niu, Andrew Liu, Zining Zhu and Gerald Penn. ICLR 2024 (spotlight).
- *Using Roark-Hollingshead Distance to Probe BERT's Syntactic Competence*. Jingcheng Niu, Wenjie Lu, Eric Corlett, and Gerald Penn. BlackboxNLP Workshop @ EMNLP 2022.
- *Does BERT Rediscover a Classical NLP Pipeline?* Jingcheng Niu, Wenjie Lu, and Gerald Penn. COLING 2022.
- *Grammaticality and Language Modelling.* Jingcheng Niu and Gerald Penn. Eval4NLP 2020 @ EMNLP 2020.

Email: niu@cs.toronto.edu
website: https://www.cs.toronto.edu/~niu/