Causal Attributions in Language Models

Yanai Elazar

ETH Zürich, 23rd February, 2022

Hi There

Yanai Elazar, PhD student, Bar-Ilan University

Bar-Ilan University
אוניברסיטת בר אילן

Ai2

With Yoav Goldberg

Hi There

NLP with Friends
By students, for students, where everyone is invited!

Home | Upcoming | Past | Calendar | Guidelines | FAQ

Welcome!

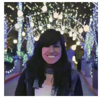
This is the home of **NLP with Friends**, an **online seminar series** made by students, for students, where everyone is invited!

About the Seminar

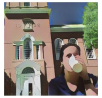
We meet Wednesdays on a bi-weekly basis to talk about interesting work in NLP and related areas. The presenters are students, who will talk about their **own work** (both ongoing and already published). Links are distributed through our mailing list.

About the Organizers

Yanai Elazar is a PhD candidate at Bar-Ilan University, where he works on neural representations, model analysis and missing elements. In his spare time he can be found nourishing flour-based organisms and converting them into bread.



Abhilasha Ravichander is a PhD candidate at Carnegie Mellon University, where she works on robust language understanding, including problems in interpretability, evaluation and computational reasoning. In her spare time she talks her plants into staying alive.



Liz Salesky is a PhD student at Johns Hopkins University, where she works on machine translation and computational linguistics. In her spare time she can be found biking to ice cream and bingeing Duolingo.

Zeerak Waseem is a PhD candidate at the University of Sheffield, where he works on abusive language detection and fairness in machine learning, and in his spare time he can be found napping.

My Research

Commonsense Reasoning

ACL19

How Large Are Lions?
Inducing Distributions over Quantitative Attributes

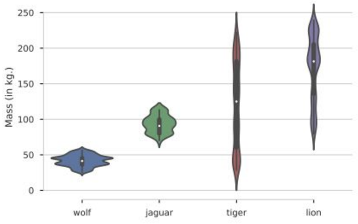
Yanai Elazar[*]
Bar Ilan Univesity
yanaiela@gmail.com

Abhijit Mahabal[†]
Pinterest
amahabal@gmail.com

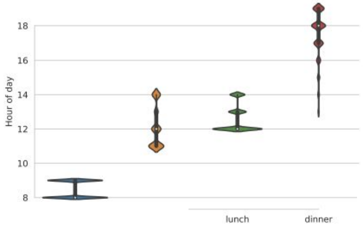
Deepak Ramachandran
Google Research
ramachandrand@google.com

Tania Bedrax-Weiss
Google Research
tbedrax@google.com

Dan Roth
University of Pennsylvania
danroth@seas.upenn.edu



(a) Mass distributions for multiple animals.



Back to Square One:
Artifact Detection, Training and Commonsense Disentanglement in the Winograd Schema

Yanai Elazar[1,2] **Hongming Zhang**[3,4] **Yoav Goldberg**[1,2] **Dan Roth**[4]
[1]Bar Ilan University, [2]AI2, [3]HKUST, [4]UPenn
{yanaiela,yoav.goldberg}@gmail.com
hzhangal@cse.ust.hk, danroth@seas.upenn.edu

EMNLP21

Setup	Example	Answer
Original		
twin-1	*The trophy* doesn't fit into *the brown suitcase* because **it** is too large.	trophy
twin-2	*The trophy* doesn't fit into *the brown suitcase* because **it** is too small.	suitcase
Baselines		
no-cands	doesn't fit into because **it** is too large.	?
part-sent	because **it** is too large.	?
Zero-shot		
twin-1	*The trophy* doesn't fit into *the brown suitcase* because **the trophy** is too [MASK].	large
twin-2	*The trophy* doesn't fit into *the brown suitcase* because **the brown suitcase** is too [MASK].	small

My Research

Commonsense Reasoning V2: Missing Elements

Where
for Nun

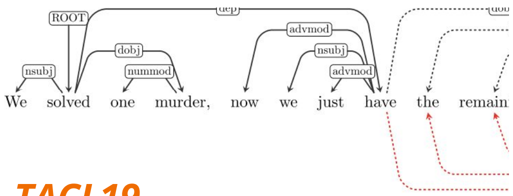
†C

g

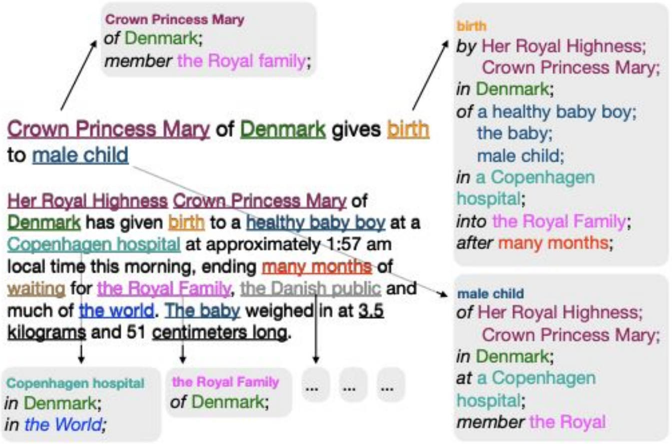
y

Text-based NP Enrichment

Yanai Elazar* Victoria Basmov* Yoav Goldberg Reut Tsarfaty
Computer Science Department, Bar Ilan University
Allen Institute for Artificial Intelligence
{yanaiela,vikasa... ...arfaty}@gmail.com



TACL19



	Annotations
...lt vow, she agrees ...gin kissing as the preacher ...ue __ the ceremony.	{officiating}, ϕ
... for max to **finish** ...ore asking him again. ⇝ ... for max to **finish swallowing** ...ore asking him again.	ENT NEU CON

My Research

and a bunch of BERTology...

Do Language E

oLMpics

Alon Talm

xikun

Ian Te
Google R
iftenney@g

T

RT:
n Multilingual BERT

ar[1,2] Yoav Goldberg[1,2]
Ilan University
elligence
oav.goldberg}@gmail.com

(But we can talk about this later!)

My Research - Today

Causal Attribution in
Language Models

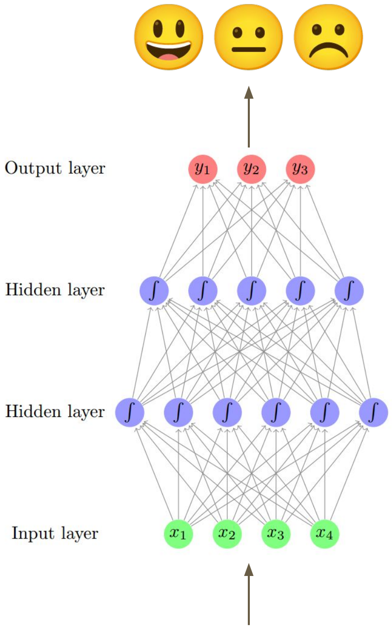
Background

On NLP, Interpretation, Muppets, and

Cramming a %&!$ sentence into a single $&!# vector

The State of NLP (ML)

Output



Model

Input

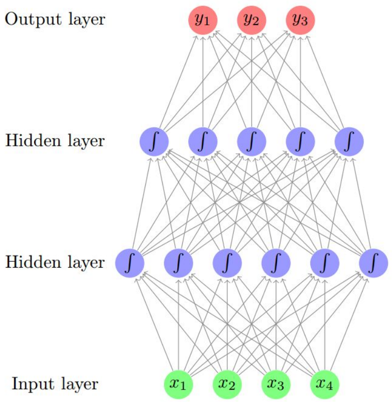
“Memories warm you up from the inside. But they also tear you apart.”

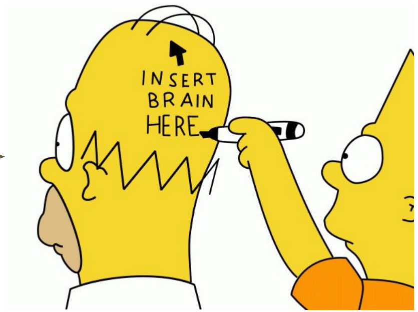
The State of NLP (ML)

Output

We train these models for some task

Model





Input

“Memories warm you up from the inside. But they also tear you apart.”

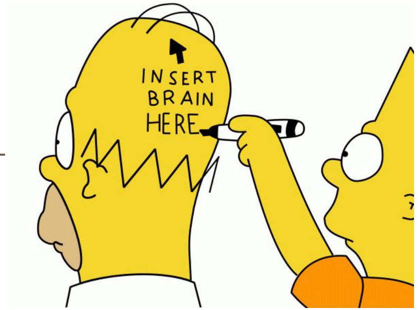
The State of NLP (ML)

Output

😃 😐 ☹️

And hope to get some "smart" model

Model



Input

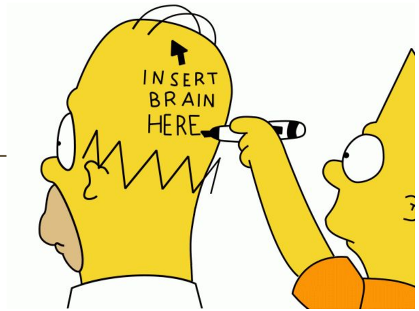
"Memories warm you up from the inside. But they also tear you apart."

The State of NLP (ML)

Output

And hope to get some
“smart” model

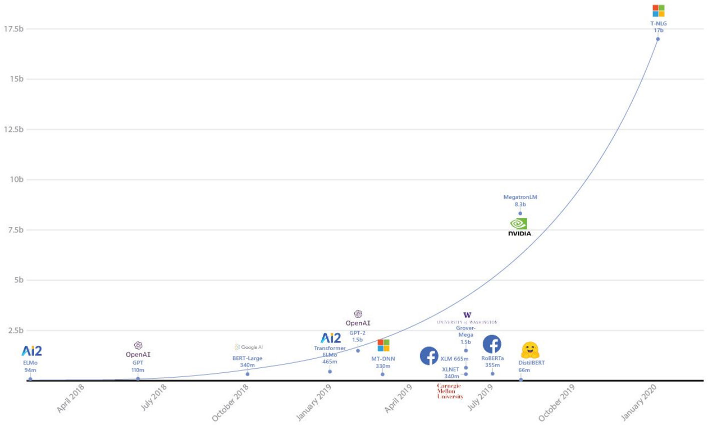
Model



Input

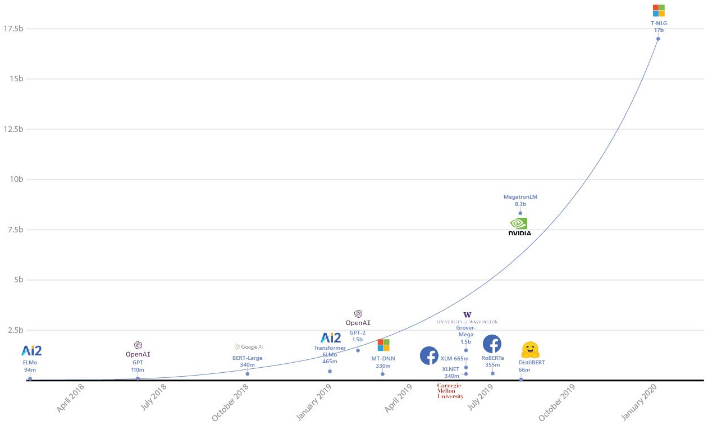
“Memories warm you up from the inside. But they also tear you apart.”

The State of NLP: Sesame Street



Input *“Memories warm you up from the inside. But they also tear you apart.”*

The State of NLP: Inside Sesame Street



Input *“Memories warm you up from the inside. But they also tear you apart.”*

The State of NLP: Inside Sesame Street

Input *"Memories warm you up from the inside. But they also tear you apart."*

Opening the BlackBox

you cannot cram the meaning of a whole %&!$# sentence into a single $&!# vector*

-- Ray Mooney

Opening the BlackBox

you cannot cram the meaning of a whole %&!$# sentence into a single $&!# vector*

-- Ray Mooney

- So what can be crammed into that?

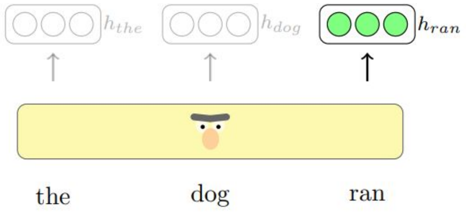
Probing

Popular approach

Probing

Popular approach

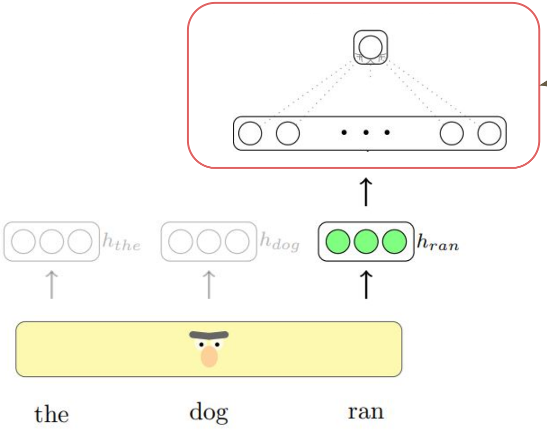
- Encode some text and retrieve its representation



Probing

Popular approach

- Encode some text and retrieve its representation
- Train a classifier to predict a property of interest

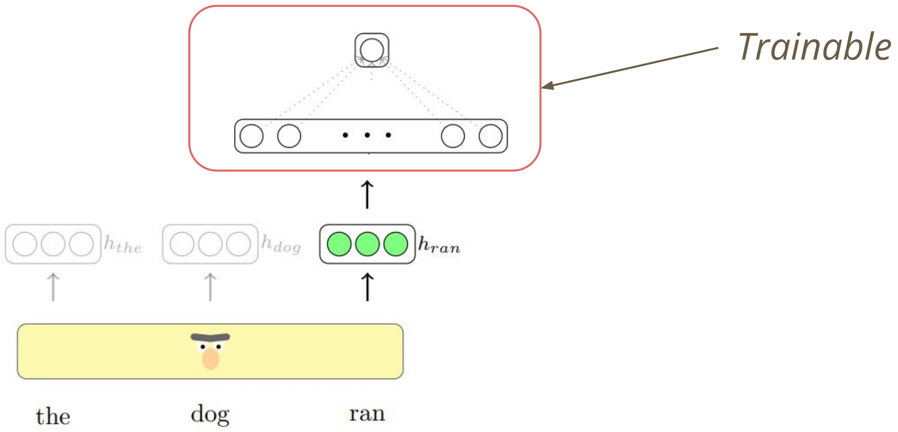


Trainable

Probing

Popular approach

- Encode some text and retrieve its representation
- Train a classifier to predict a property of interest
- High performance is interpreted as the encoding of the property



People Probe for...

- Sentence Length
- Word Order
- Tense

Adi et al., 2016, Conneau et al., 2018, Hewitt and Manning, 2019, Tenney et al., 2019, Chi et al., 2020

People Probe for...

- Sentence Length
- Word Order
- Tense
- POS
- Tree depth
- Entities
- Coref.
- ...

Adi et al., 2016, Conneau et al., 2018, Hewitt and Manning, 2019, Tenney et al., 2019, Chi et al., 2020

What’s Wrong with Probing?

Probing - The Problem

Probing answers:

“*What is encoded* in the representation?”

But the interesting question is:

“*What is being used* for prediction?”

Probing - The Problem

Probing answers:

“*What is encoded* in the representation?”

But the interesting question is:

“*What is being used* for prediction?”

Which are very **different** questions!

Part I

Amnesic Probing: Behavioral Explanation with Amnesic Counterfactuals

Yanai Elazar[1,2] Shauli Ravfogel[1,2] Alon Jacovi[1] Yoav Goldberg[1,2]
[1]Computer Science Department, Bar Ilan University
[2]Allen Institute for Artificial Intelligence

TACL 2021

Our Solution: *Amnesic Probing*, A Behavioral Probe

Amnesic Probing: A Behavioral Probe

- Interpretability tool, which allows to:
    - Answer scientific questions (e.g. does an LM use POS information?)
    - Answer applicative questions (e.g. does the model use gender for making a decision?)

Probing answers:

"*What is encoded* in the representation?"

But the interesting question is:

"*What is being used* for prediction?"

Probing answers:

“*What is encoded* in the representation?”

Probing

But the interesting question is:

“*What is being used* for prediction?”

Probing answers:

"*What is encoded* in the representation?"

Probing

But the interesting question is:

"*What is being used* for prediction?"

Amnesic Probing

The Intuition: Counterfactuals

What would the model predict **without** a given concept?

Amnesic Probing: The Intuition

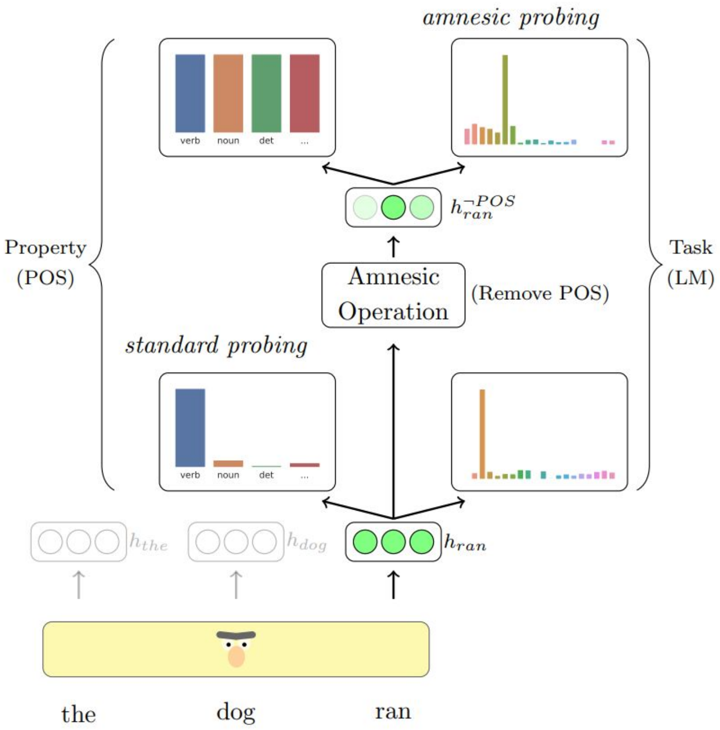
- Counterfactuals (or *ablation* on a trained model):
  - Remove a certain component, property
  - Measure how it affects the results
- Since it is hard to intervene on the input text…
  …**we intervene on the representation**

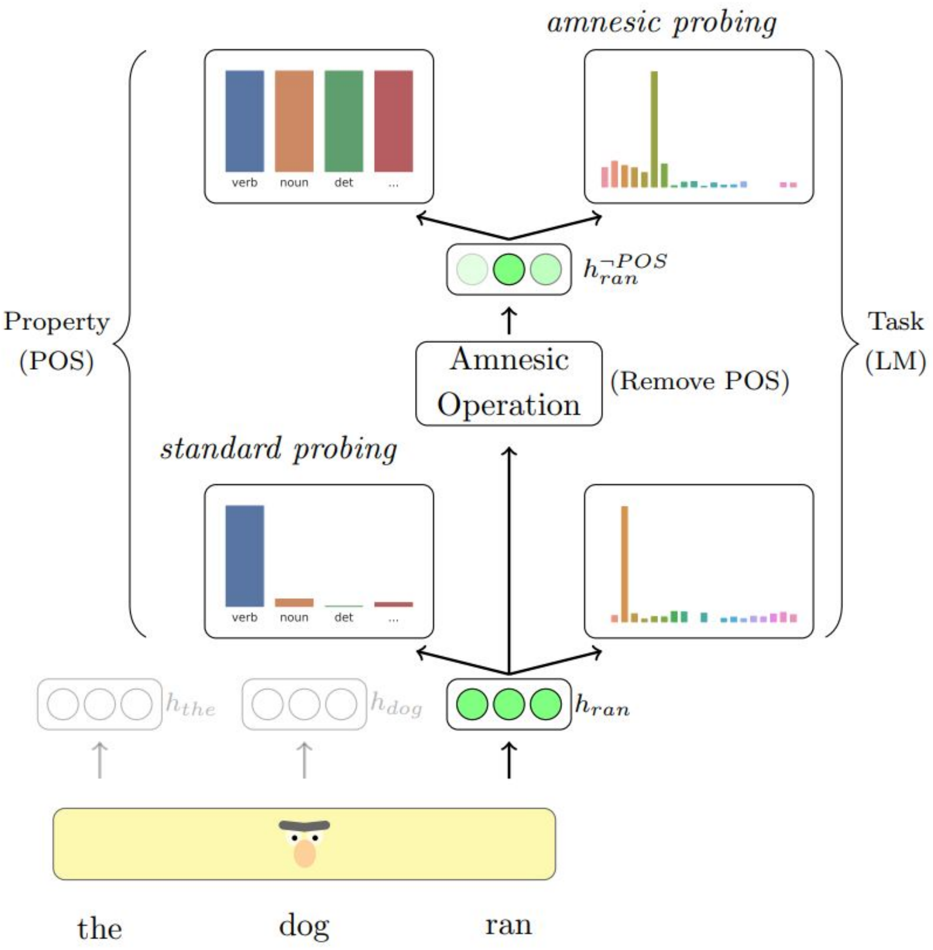
Amnesic Probing: The Intuition

- We remove a feature from the representation (e.g. remove POS information)
- Does the model change its behavior?

- Yes:
  - The model uses this information for its predictions
- No:
  - The model does **not** use this information for its predictions

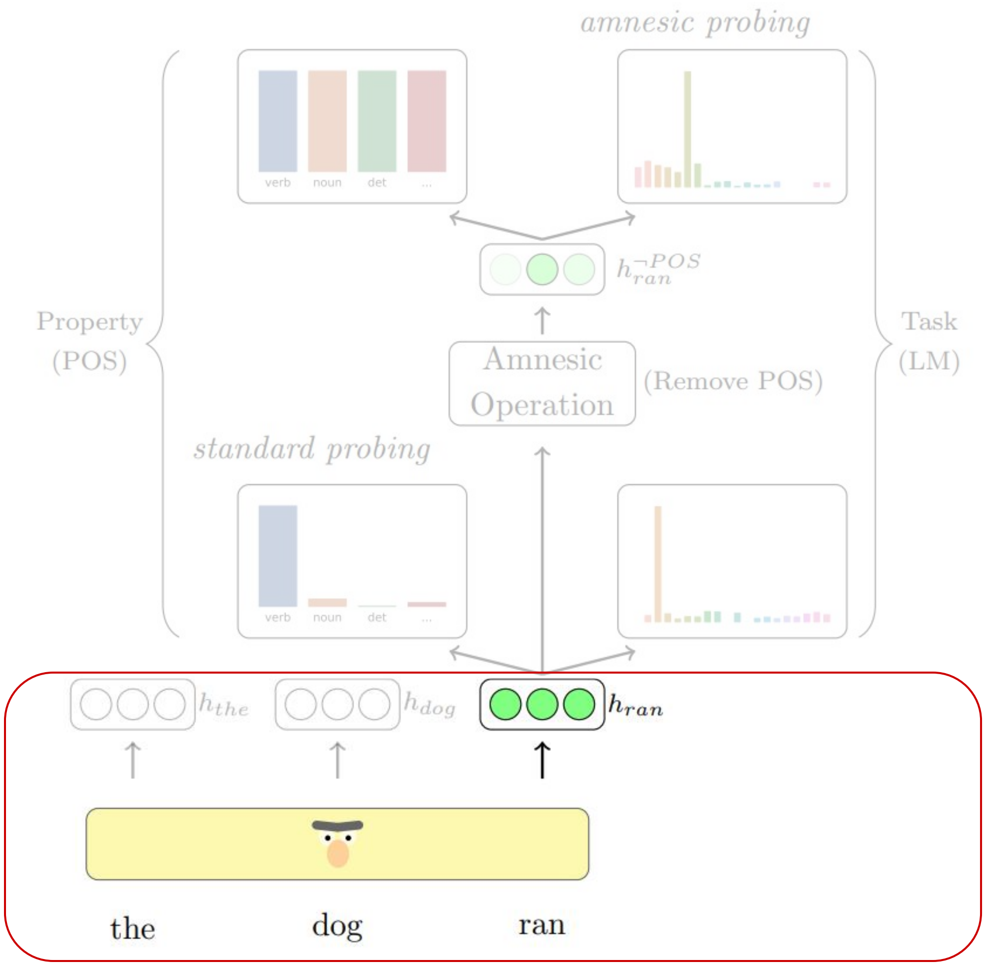
Amnesic Probing: Overview





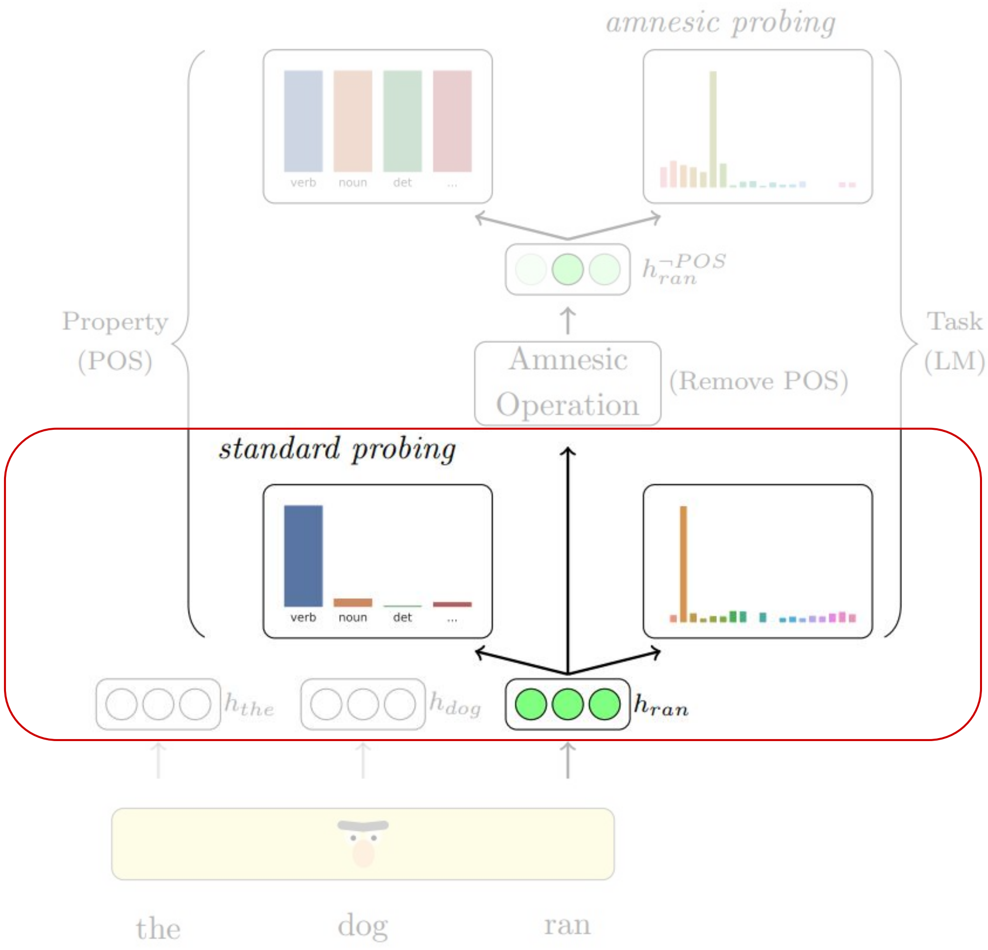
amnesic probing
verb
noun
det
...
$h_{ran}^{\neg POS}$
Property
(POS)
Amnesic
Operation
(Remove POS)
Task
(LM)
standard probing
verb
noun
det
...
h_{the}
h_{dog}
h_{ran}
the
dog
ran

1. Encode

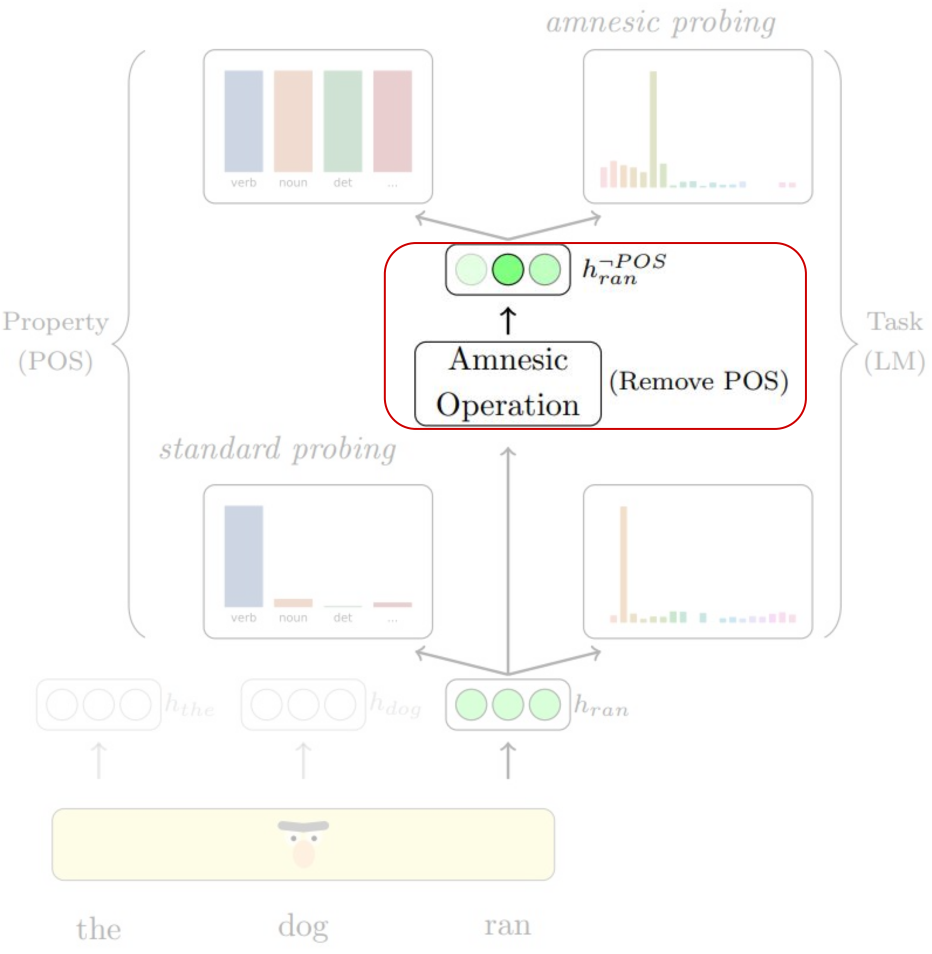


amnesic probing
verb
noun
det
...
$h_{ran}^{\neg POS}$
Property
(POS)
Amnesic
Operation
(Remove POS)
Task
(LM)
standard probing
verb
noun
det
...
h_{the}
h_{dog}
h_{ran}
the
dog
ran

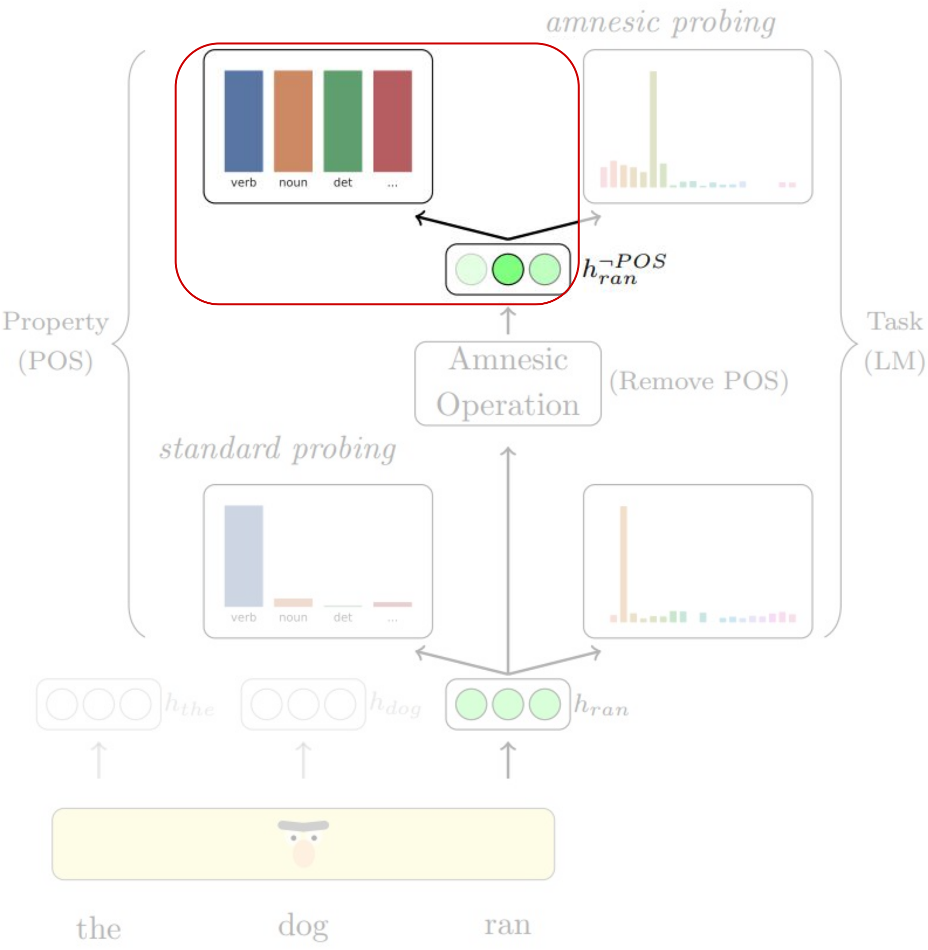
1. Encode
2. Probe



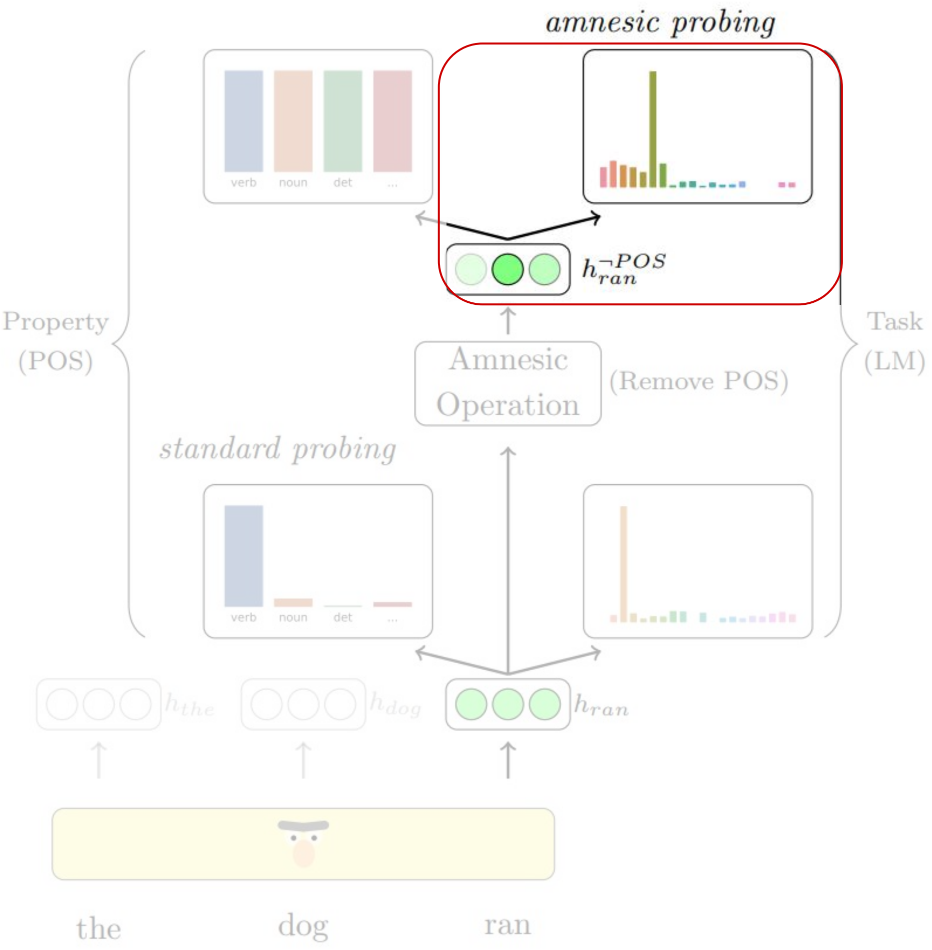
1. Encode
2. Probe
3. Amnesia



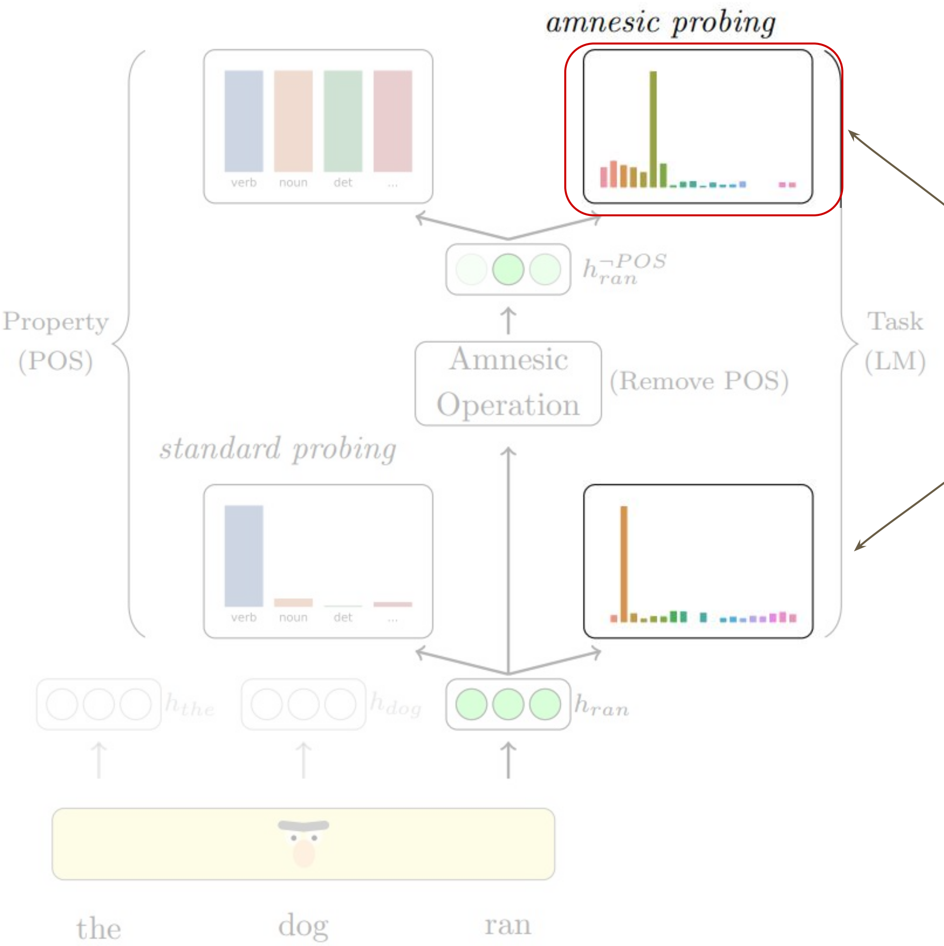
1. Encode
2. Probe
3. Amnesia
   3.1. Verify



1. Encode
2. Probe
3. Amnesia
    3.1. Verify
4. *Amnesic Probing*



1. Encode
2. Probe
3. Amnesia
   3.1. Verify
4. *Amnesic Probing*



Compare

The Amnesic Operation

Amnesic Probing: The Amnesia

One option: Adversarial Training

Adversarial Removal of Demographic Attributes from Text Data

Yanai Elazar[†] and **Yoav Goldberg**[†*]

[†]Computer Science Department, Bar-Ilan University, Israel

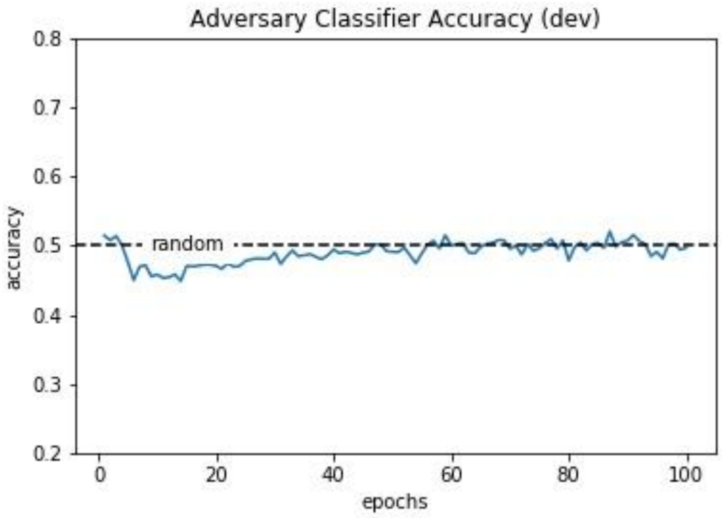
[*]Allen Institute for Artificial Intelligence

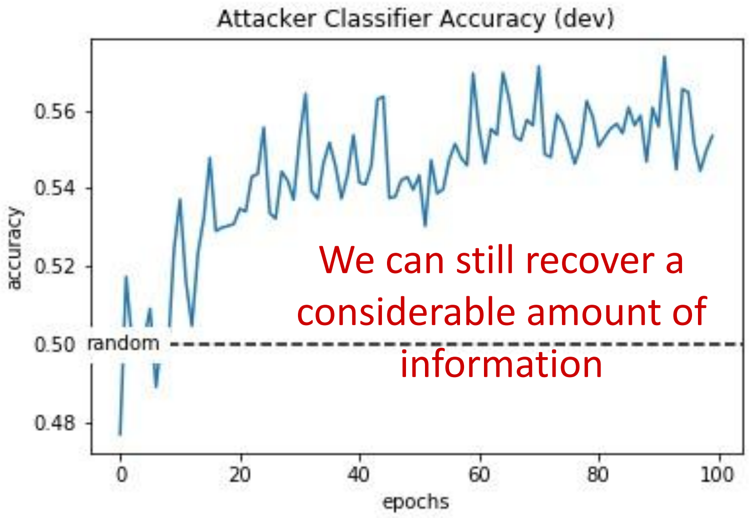
{yanaiela,yoav.goldberg}@gmail.com

EMNLP 2018

Amnesic Probing: The Amnesia

One option: Adversarial Training





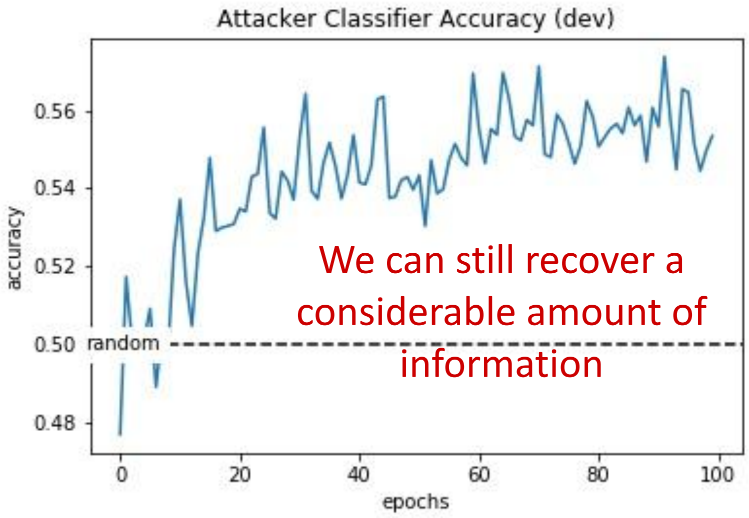
EMNLP 2018

Amnesic Probing: The Amnesia

One option: Adversarial Training

But also:

- Slow & unstable training
- Is it the same model afterwards?



EMNLP 2018

Amnesic Probing: The Amnesia

Null It Out: Guarding Protected Attributes by Iterative Nullspace Projection

Shauli Ravfogel[1,2] **Yanai Elazar**[1,2] **Hila Gonen**[1] **Michael Twiton**[3] **Yoav Goldberg**[1,2]
[1]Computer Science Department, Bar Ilan University
[2]Allen Institute for Artificial Intelligence
[3]Independent researcher

ACL 2020

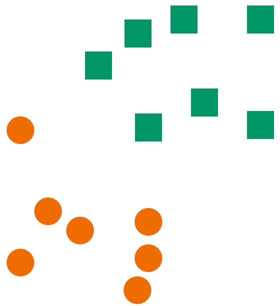
Amnesic Operation: Using INLP

- An algorithm for removing linear information from deep networks

Ravfogel et al., 2020

Amnesic Operation: Using INLP

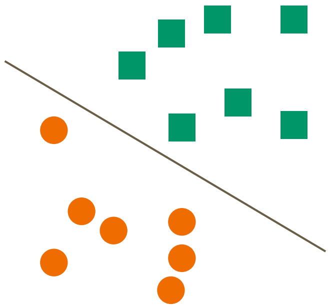
- An algorithm for removing linear information from deep networks
- Receives representations and labels, and returns a function



Ravfogel et al., 2020

Amnesic Operation: Using INLP

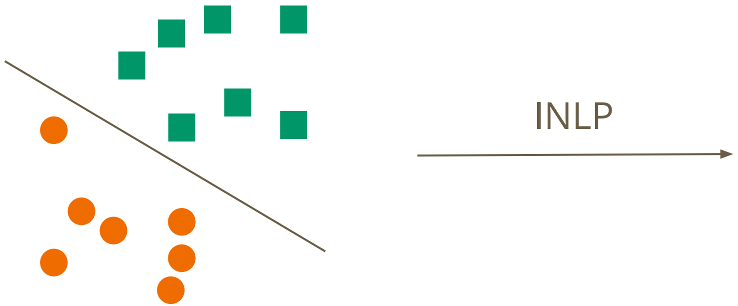
- An algorithm for removing linear information from deep networks
- Receives representations and labels, and returns a function



Ravfogel et al., 2020

Amnesic Operation: Using INLP

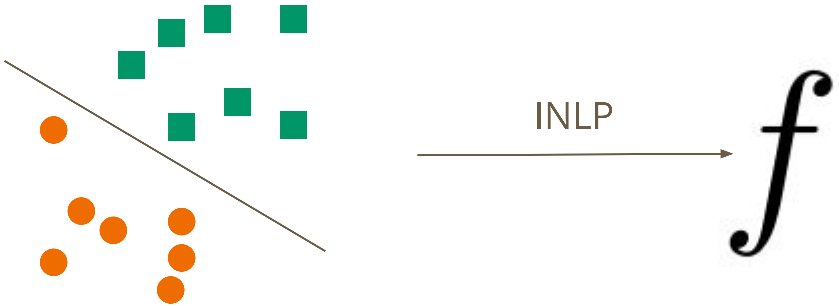
- An algorithm for removing linear information from deep networks
- Receives representations and labels, and returns a function



Ravfogel et al., 2020

Amnesic Operation: Using INLP

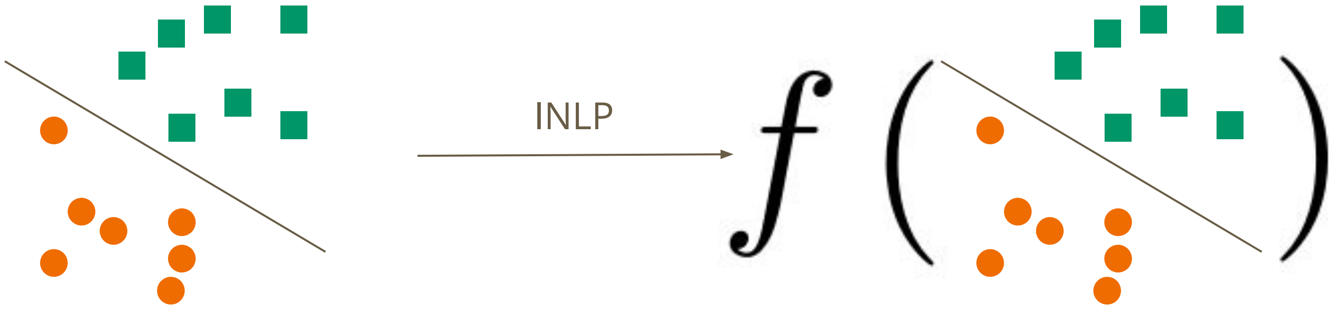
- An algorithm for removing linear information from deep networks
- Receives representations and labels, and returns a function



Ravfogel et al., 2020

Amnesic Operation: Using INLP

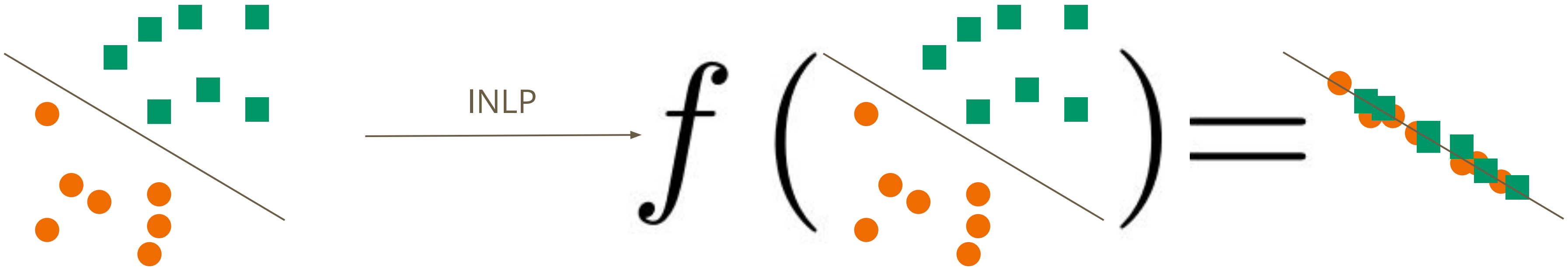
- An algorithm for removing linear information from deep networks
- Receives representations and labels, and returns a function
- When applied to vectors, any linear model cannot predict the labels



Ravfogel et al., 2020

Amnesic Operation: Using INLP

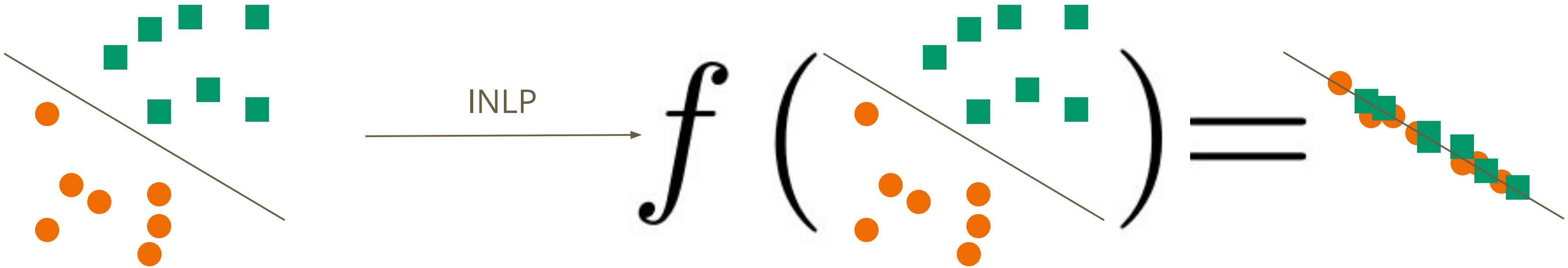
- An algorithm for removing linear information from deep networks
- Receives representations and labels, and returns a function
- When applied to vectors, any linear model cannot predict the labels



Ravfogel et al., 2020

Amnesic Operation: Using INLP

- An algorithm for removing linear information from deep networks
- Receives representations and labels, and returns a function
- When applied to vectors, any linear model cannot predict the labels



Ravfogel et al., 2020

(*) We use INLP in this work, but this is a component that can be replaced with a future (non-linear) alternative

Feder et al. 2021

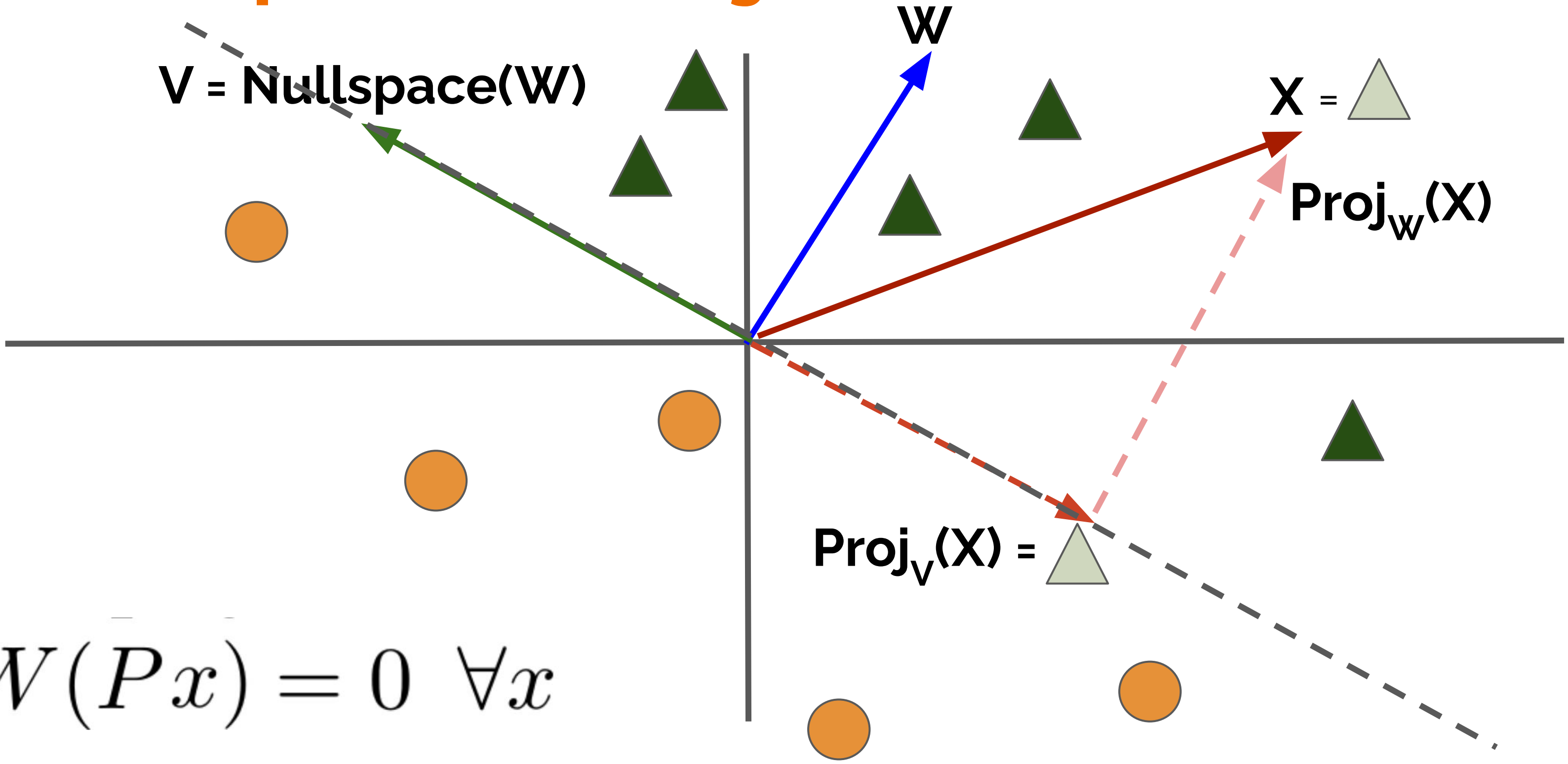
Amnesic Operation: Using INLP

INLP: Iterative Nullspace Projection

- Find a projection matrix P, which projects into the nullspace

$$N(W) = \{x | Wx = 0\}$$

Amnesic Operation: Using INLP



$$W(Px) = 0 \ \forall x$$

Amnesic Operation: Using INLP

- Each projection only removes a single direction
- Therefore the “iterative” part:
- We repeat this process until convergence

Amnesic Operation: Using INLP

- Debiasing applications (Ravfogel et al., 2020)

Check it out!

		BoW	FastText	BERT
Accuracy (profession)	Original	78.2	78.1	80.9
	+INLP	80.1	73.0	75.2
$GAP^{TPR,RMS}_{male}$	Original	0.203	0.184	0.184
	+INLP	0.124	0.089	0.095

Table 2: Fair classification on the Biographies corpus.

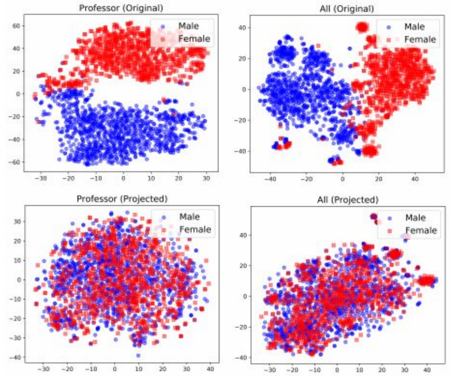


Figure 3: t-SNE projection of BERT representations for the profession "professor" (left) and for a random sample of all professions (right), before and after the projection.

Amnesic Probing: Setup

- Start with a trained model
- Encode and obtain the representations
- Choose properties/features of interest
- Remove them
- Measure the difference (behavioral!), via:
    - Accuracy (of predicting the “right” label)

Verifying that the Amnesic Operation Works

Amnesic Probing: Controls

- Did the amnesic operation remove too little?
- Did the amnesic operation remove too much?

Amnesic Probing: Controls

- Did the amnesic operation remove too little?
- Did the amnesic operation remove too much?

- Control over Information
  - Removing random features

Amnesic Probing: Controls

- Did the amnesic operation remove too little?
- Did the amnesic operation remove too much?

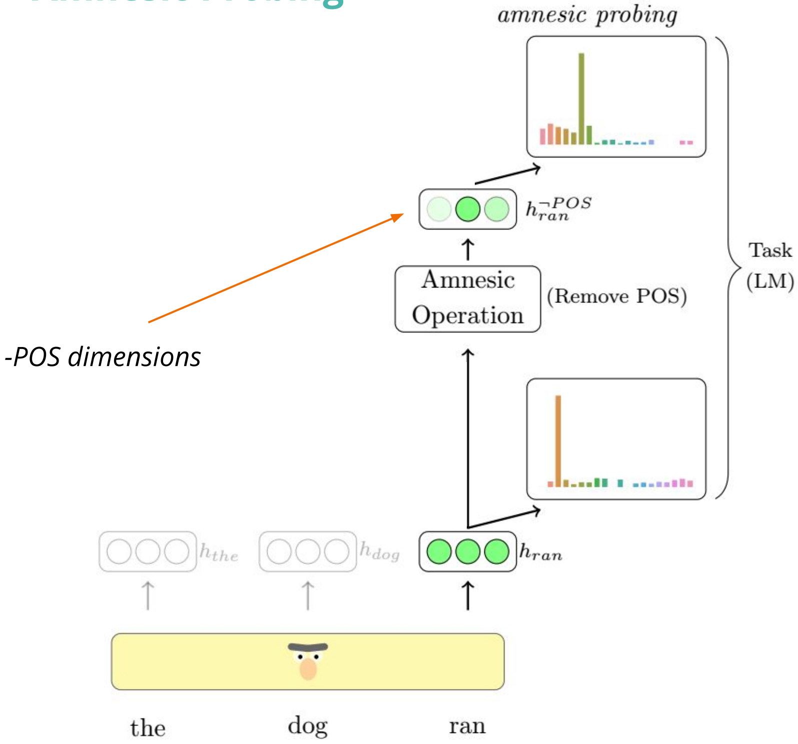
- Control over Information
  - Removing random features
- Control over Selectivity
  - Add back the “real” features, and retrain

Amnesic Probing: Controls

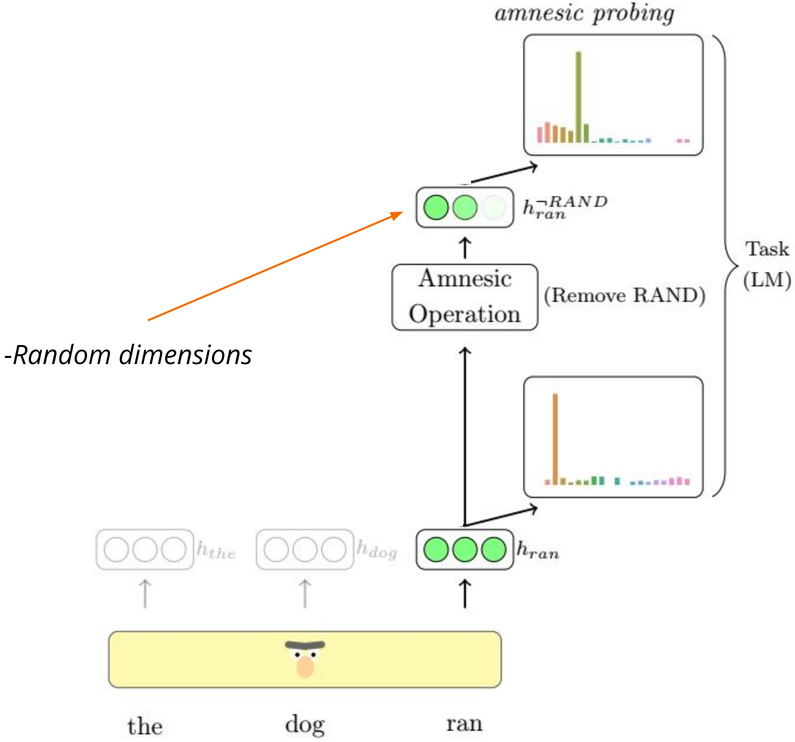
- Did the amnesic operation remove too little?
- Did the amnesic operation remove too much?

- Control over Information
  - Removing random features
- Control over Selectivity
  - Add back the “real” features, and retrain
- Hopefully we’ll be here

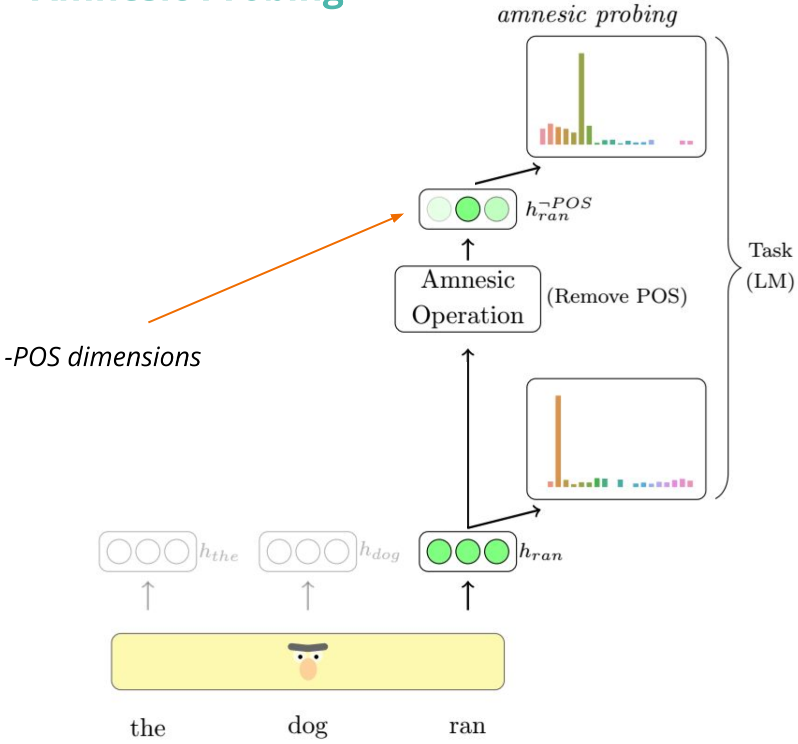
Amnesic Probing



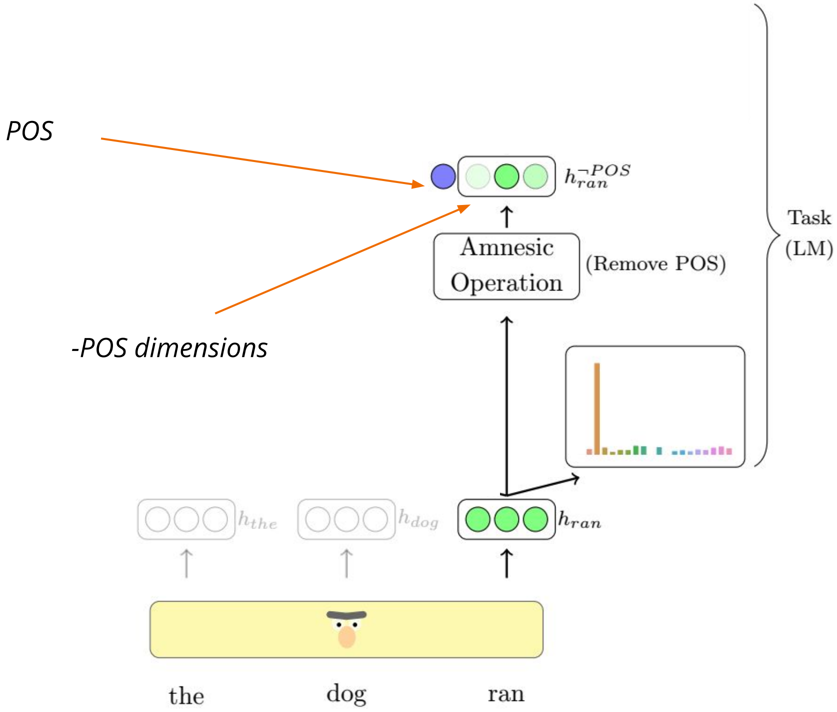
Control over Information



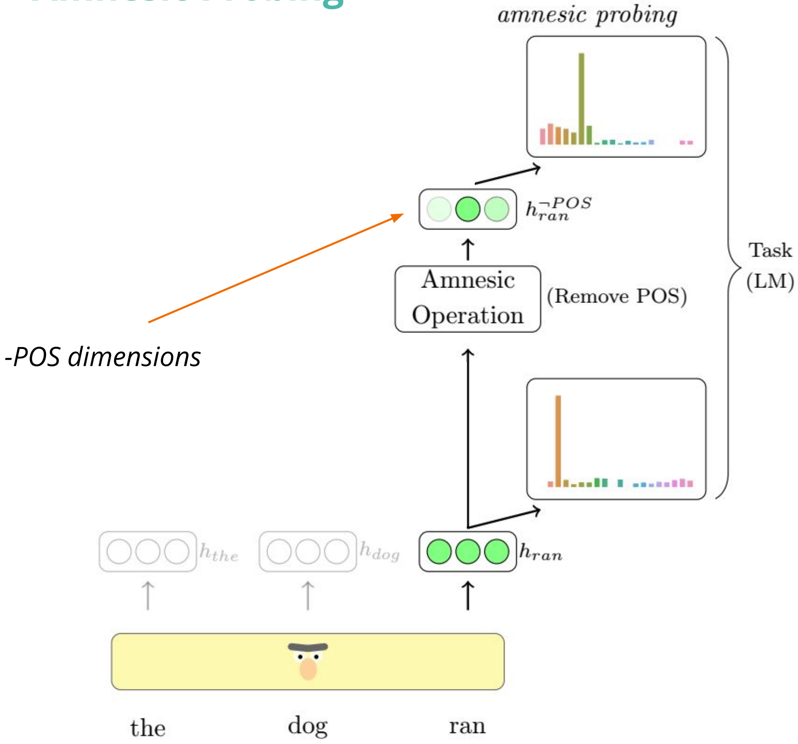
Amnesic Probing



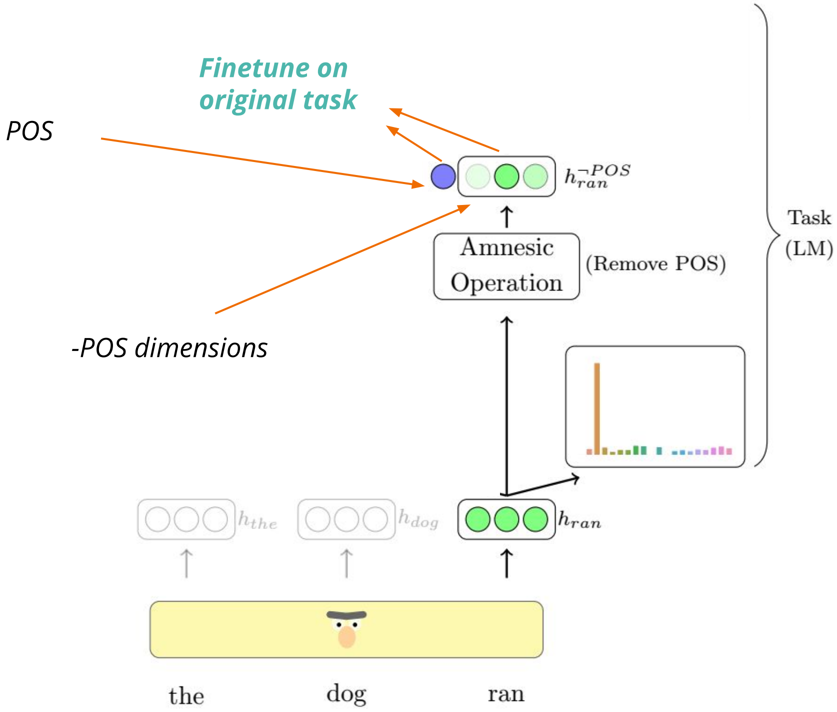
Control over Selectivity



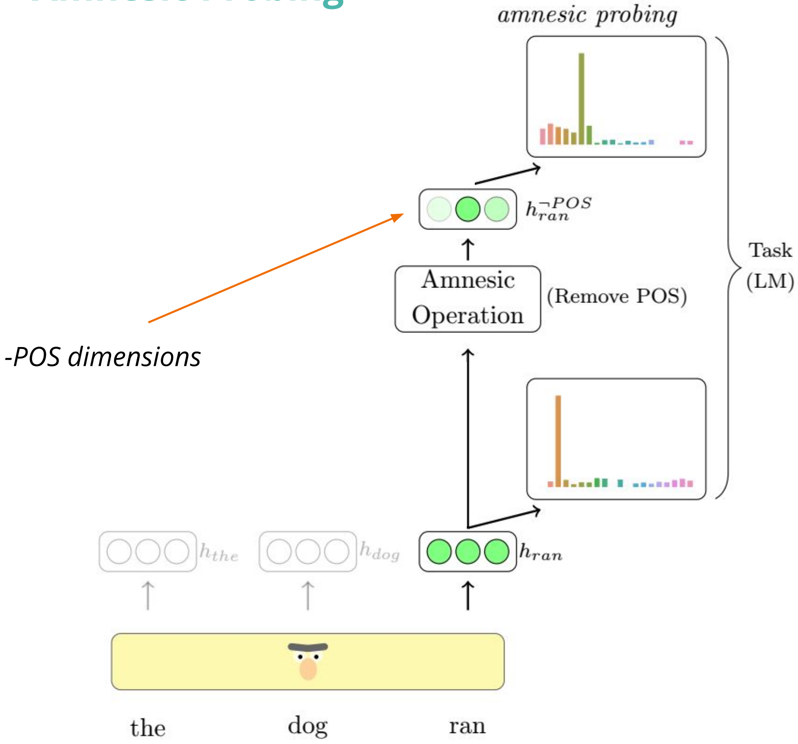
Amnesic Probing



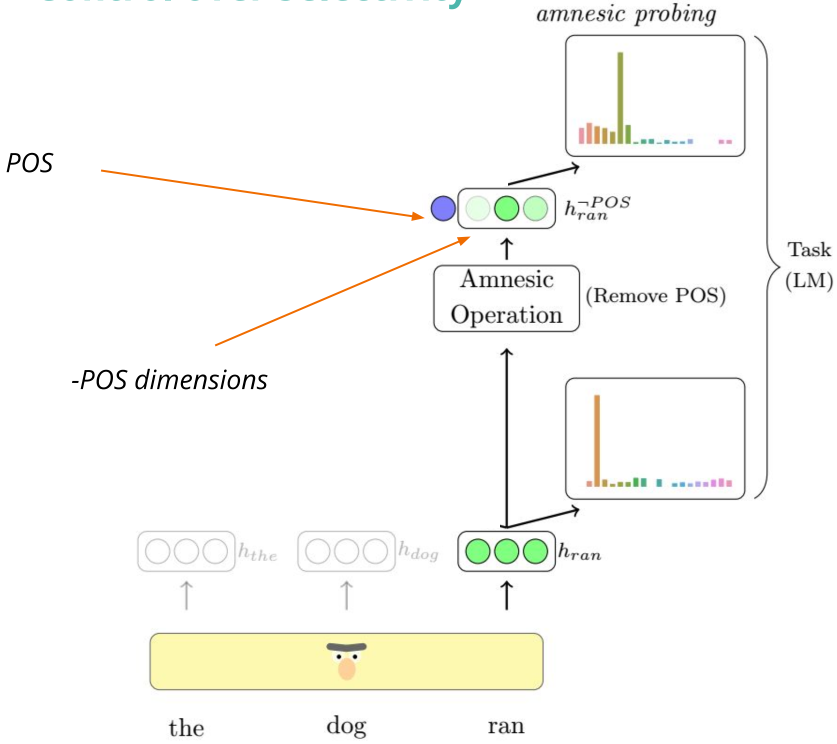
Control over Selectivity



Amnesic Probing



Control over Selectivity



Case Study: Pre-trained BERT

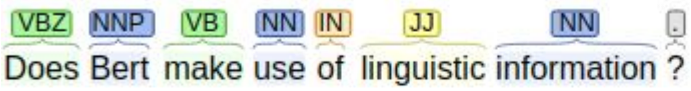
What linguistic properties are ~~encoded~~ used in BERT

Amnesic Probing: Setup

- The model: BERT-base

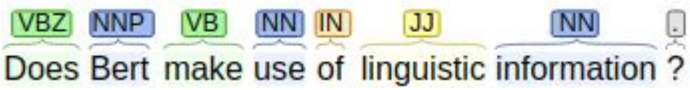
Amnesic Probing: Setup

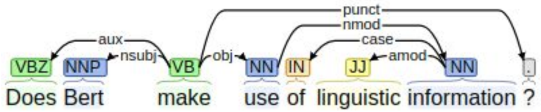
- The model: BERT-base
- Properties:
  - POS



Amnesic Probing: Setup

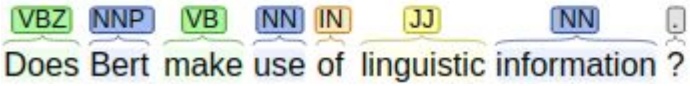
- The model: BERT-base
- Properties:
    - POS
    - Dependency edges

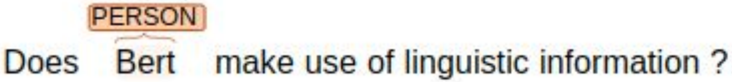


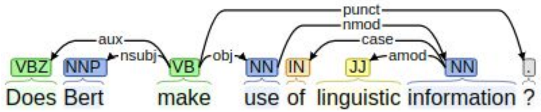


Amnesic Probing: Setup

- The model: BERT-base
- Properties:
  - POS
  - Dependency edges
  - NER

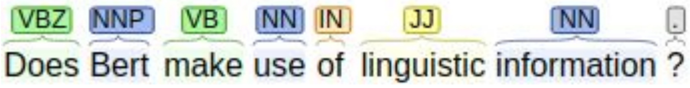


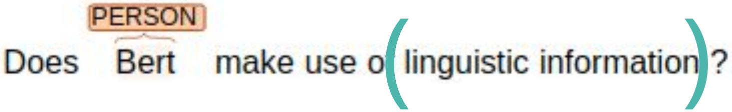


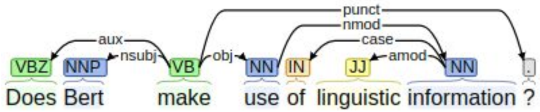


Amnesic Probing: Setup

- The model: BERT-base
- Properties:
  - POS
  - Dependency edges
  - NER
  - Constituency boundaries



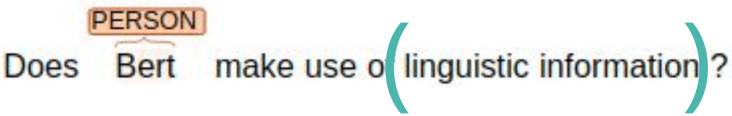


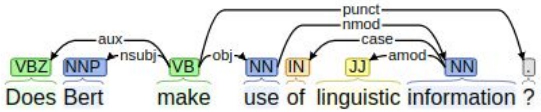


Amnesic Probing: Setup

- The model: BERT-base
- Properties:
  - POS
  - D...

Does BERT make use of *POS, Dep-edge, NER* and *Const-boundaries* when predicting words?





Amnesic Probing: Results

		dep	*f-pos*	*c-pos*	*ner*	*phrase start*	*phrase end*
Properties	N. dir	738	585	264	133	36	22
	N. classes	41	45	12	19	2	2
	Majority	11.44	13.22	31.76	86.09	59.25	58.51
Probing	Vanilla	76.00	89.50	92.34	93.53	85.12	83.09
LM-Acc	Vanilla	94.12	94.12	94.12	94.00	94.00	94.00
	Rand	12.31	56.47	89.65	92.56	93.75	93.86
	Selectivity	73.78	92.68	97.26	96.06	96.96	96.93
	Amnesic	7.05	12.31	61.92	83.14	94.21	94.32
LM-D_{KL}	Rand	8.11	4.61	0.36	0.08	0.01	0.01
	Amnesic	8.53	7.63	3.21	1.24	0.01	0.01

Amnesic Probing: Results

Linguistic Properties

		dep	*f-pos*	*c-pos*	*ner*	*phrase start*	*phrase end*
Properties	N. dir	738	585	264	133	36	22
	N. classes	41	45	12	19	2	2
	Majority	11.44	13.22	31.76	86.09	59.25	58.51
Probing	Vanilla	76.00	89.50	92.34	93.53	85.12	83.09
LM-Acc	Vanilla	94.12	94.12	94.12	94.00	94.00	94.00
	Rand	12.31	56.47	89.65	92.56	93.75	93.86
	Selectivity	73.78	92.68	97.26	96.06	96.96	96.93
	Amnesic	7.05	12.31	61.92	83.14	94.21	94.32
LM-D_{KL}	Rand	8.11	4.61	0.36	0.08	0.01	0.01
	Amnesic	8.53	7.63	3.21	1.24	0.01	0.01

Amnesic Probing: Results

Standard Probing

		dep	*f-pos*	*c-pos*	*ner*	*phrase start*	*phrase end*
Properties	N. dir	738	585	264	133	36	22
	N. classes	41	45	12	19	2	2
	Majority	11.44	13.22	31.76	86.09	59.25	58.51
Probing	**Vanilla**	**76.00**	**89.50**	**92.34**	**93.53**	**85.12**	**83.09**
LM-Acc	Vanilla	94.12	94.12	94.12	94.00	94.00	94.00
	Rand	12.31	56.47	89.65	92.56	93.75	93.86
	Selectivity	73.78	92.68	97.26	96.06	96.96	96.93
	Amnesic	7.05	12.31	61.92	83.14	94.21	94.32
LM-D_{KL}	Rand	8.11	4.61	0.36	0.08	0.01	0.01
	Amnesic	8.53	7.63	3.21	1.24	0.01	0.01

Amnesic Probing: Results

LM Accuracy Results

		dep	*f-pos*	*c-pos*	*ner*	*phrase start*	*phrase end*
Properties	N. dir	738	585	264	133	36	22
	N. classes	41	45	12	19	2	2
	Majority	11.44	13.22	31.76	86.09	59.25	58.51
Probing	Vanilla	76.00	89.50	92.34	93.53	85.12	83.09
LM-Acc	Vanilla	94.12	94.12	94.12	94.00	94.00	94.00
	Rand	12.31	56.47	89.65	92.56	93.75	93.86
	Selectivity	73.78	92.68	97.26	96.06	96.96	96.93
	Amnesic	7.05	12.31	61.92	83.14	94.21	94.32
LM-D_{KL}	Rand	8.11	4.61	0.36	0.08	0.01	0.01
	Amnesic	8.53	7.63	3.21	1.24	0.01	0.01

Amnesic Probing: Results

Amnesic Comparison

		dep	*f-pos*	*c-pos*	*ner*	*phrase start*	*phrase end*
Properties	N. dir	738	585	264	133	36	22
	N. classes	41	45	12	19	2	2
	Majority	11.44	13.22	31.76	86.09	59.25	58.51
Probing	Vanilla	76.00	89.50	92.34	93.53	85.12	83.09
LM-Acc	**Vanilla**	**94.12**	**94.12**	**94.12**	**94.00**	**94.00**	**94.00**
	Rand	**12.31**	**56.47**	**89.65**	**92.56**	**93.75**	**93.86**
	Selectivity	**73.78**	**92.68**	**97.26**	**96.06**	**96.96**	**96.93**
	Amnesic	**7.05**	**12.31**	**61.92**	**83.14**	**94.21**	**94.32**
LM-D_{KL}	Rand	8.11	4.61	0.36	0.08	0.01	0.01
	Amnesic	8.53	7.63	3.21	1.24	0.01	0.01

Amnesic Probing: Results

Amnesic Comparison

		dep	f-pos	c-pos	ner	phrase start	phrase end
Properties	N. dir	738	585	264	133	36	22
	N. classes	41	45	12	19	2	2
	Majority	11.44	13.22	31.76	86.09	59.25	58.51
Probing	Vanilla	76.00	89.50	92.34	93.53	85.12	83.09
LM-Acc	**Vanilla**	94.12	94.12	94.12	94.00	94.00	94.00
	Rand	12.31	56.47	89.65	92.56	93.75	93.86
	Selectivity	73.78	92.68	97.26	96.06	96.96	96.93
	Amnesic	7.05	12.31	61.92	83.14	94.21	94.32
LM-D_{KL}	Rand	8.11	4.61	0.36	0.08	0.01	0.01
	Amnesic	8.53	7.63	3.21	1.24	0.01	0.01

Amnesic Probing: Results

Comparison to Control:
Information

		dep	*f-pos*	*c-pos*	*ner*	*phrase start*	*phrase end*
Properties	N. dir	738	585	264	133	36	22
	N. classes	41	45	12	19	2	2
	Majority	11.44	13.22	31.76	86.09	59.25	58.51
Probing	Vanilla	76.00	89.50	92.34	93.53	85.12	83.09
LM-Acc	Vanilla	94.12	94.12	94.12	94.00	94.00	94.00
	Rand	12.31	56.47	89.65	92.56	93.75	93.86
	Selectivity	73.78	92.68	97.26	96.06	96.96	96.93
	Amnesic	7.05	12.31	61.92	83.14	94.21	94.32
LM-D_{KL}	Rand	8.11	4.61	0.36	0.08	0.01	0.01
	Amnesic	8.53	7.63	3.21	1.24	0.01	0.01

Amnesic Probing: Results

Comparison to Control:
Information

		dep	*f-pos*	*c-pos*	*ner*	*phrase start*	*phrase end*
Properties	N. dir	738	585	264	133	36	22
	N. classes	41	45	12	19	2	2
	Majority	11.44	13.22	31.76	86.09	59.25	58.51
Probing	Vanilla	76.00	89.50	92.34	93.53	85.12	83.09
LM-Acc	Vanilla	94.12	94.12	94.12	94.00	94.00	94.00
	Rand	12.31	56.47	89.65	92.56	93.75	93.86
	Selectivity	73.78	92.68	97.26	96.06	96.96	96.93
	Amnesic	7.05	12.31	61.92	83.14	94.21	94.32
LM-D_{KL}	Rand	8.11	4.61	0.36	0.08	0.01	0.01
	Amnesic	8.53	7.63	3.21	1.24	0.01	0.01

Amnesic Probing: Results

Comparison to Control:
Selectivity

		dep	*f-pos*	*c-pos*	*ner*	*phrase start*	*phrase end*
Properties	N. dir	738	585	264	133	36	22
	N. classes	41	45	12	19	2	2
	Majority	11.44	13.22	31.76	86.09	59.25	58.51
Probing	Vanilla	76.00	89.50	92.34	93.53	85.12	83.09
LM-Acc	**Vanilla**	**94.12**	**94.12**	**94.12**	**94.00**	**94.00**	**94.00**
	Rand	**12.31**	**56.47**	**89.65**	**92.56**	**93.75**	**93.86**
	Selectivity	**73.78**	**92.68**	**97.26**	**96.06**	**96.96**	**96.93**
	Amnesic	**7.05**	**12.31**	**61.92**	**83.14**	**94.21**	**94.32**
LM-D_{KL}	Rand	8.11	4.61	0.36	0.08	0.01	0.01
	Amnesic	8.53	7.63	3.21	1.24	0.01	0.01

Amnesic Probing: Results

Comparison to Control:
Selectivity

Doesn't Recover

		dep	*f-pos*	*c-pos*	*ner*	*phrase start*	*phrase end*
Properties	N. dir	738	585	264	133	36	22
	N. classes	41	45	12	19	2	2
	Majority	11.44	13.22	31.76	86.09	59.25	58.51
Probing	Vanilla	76.00	89.50	92.34	93.53	85.12	83.09
LM-Acc	Vanilla	94.12	94.12	94.12	94.00	94.00	94.00
	Rand	12.31	56.47	89.65	92.56	93.75	93.86
	Selectivity	73.78	92.68	97.26	96.06	96.96	96.93
	Amnesic	7.05	12.31	61.92	83.14	94.21	94.32
LM-D_{KL}	Rand	8.11	4.61	0.36	0.08	0.01	0.01
	Amnesic	8.53	7.63	3.21	1.24	0.01	0.01

Amnesic Probing: Results

Comparison to Control: ***Selectivity***

		dep	f-pos	c-pos	ner	phrase start	phrase end
Properties	N. dir	738	585	264	133	36	22
	N. classes	41	45	12	19	2	2
	Majority	11.44	13.22	31.76	86.09	59.25	58.51
Probing	Vanilla	76.00	89.50	92.34	93.53	85.12	83.09
LM-Acc	Vanilla	94.12	94.12	94.12	94.00	94.00	94.00
	Rand	12.31	56.47	89.65	92.56	93.75	93.86
	Selectivity	73.78	92.68	97.26	96.06	96.96	96.93
	Amnesic	7.05	12.31	61.92	83.14	94.21	94.32
LM-D_{KL}	Rand	8.11	4.61	0.36	0.08	0.01	0.01
	Amnesic	8.53	7.63	3.21	1.24	0.01	0.01

Does Recover

Amnesic Probing: Results

Phrase markers ***are not*** *being used*

Conclusions from all this:

				c-pos	ner	phrase start	phrase end
				264	133	36	22
				12	19	2	2
				31.76	86.09	59.25	58.51
Probing	Vanilla	76.00	89.50	92.34	93.53	85.12	83.09
LM-Acc	Vanilla	94.12	94.12	94.12	94.00	94.00	94.00
	Rand	12.31	56.47	89.65	92.56	93.75	93.86
	Selectivity	73.78	92.68	97.26	96.06	96.96	96.93
	Amnesic	7.05	12.31	61.92	83.14	94.21	94.32
LM-D_{KL}	Rand	8.11	4.61	0.36	0.08	0.01	0.01
	Amnesic	8.53	7.63	3.21	1.24	0.01	0.01

POS and NER are being ***used*** *by the model*

Amnesic Probing: Results

		dep	*f-pos*	*c-pos*	*ner*	*phrase start*	*phrase end*
Properties	N. dir	738	585	264	133	36	22
	N. classes	41	45	12	19	2	2
	Majority	11.44	13.22	31.76	86.09	59.25	58.51
Probing	Vanilla	76.00	89.50	92.34	93.53	85.12	83.09
LM-Acc	Vanilla	94.12	94.12	94.12	94.00	94.00	94.00
	Rand	12.31	56.47	89.65	92.56	93.75	93.86
	Selectivity	73.78	92.68	97.26	96.06	96.96	96.93
	Amnesic	7.05	12.31	61.92	83.14	94.21	94.32
LM-D_{KL}	Rand	8.11	4.61	0.36	0.08	0.01	0.01
	Amnesic	8.53	7.63	3.21	1.24	0.01	0.01

DKL Results

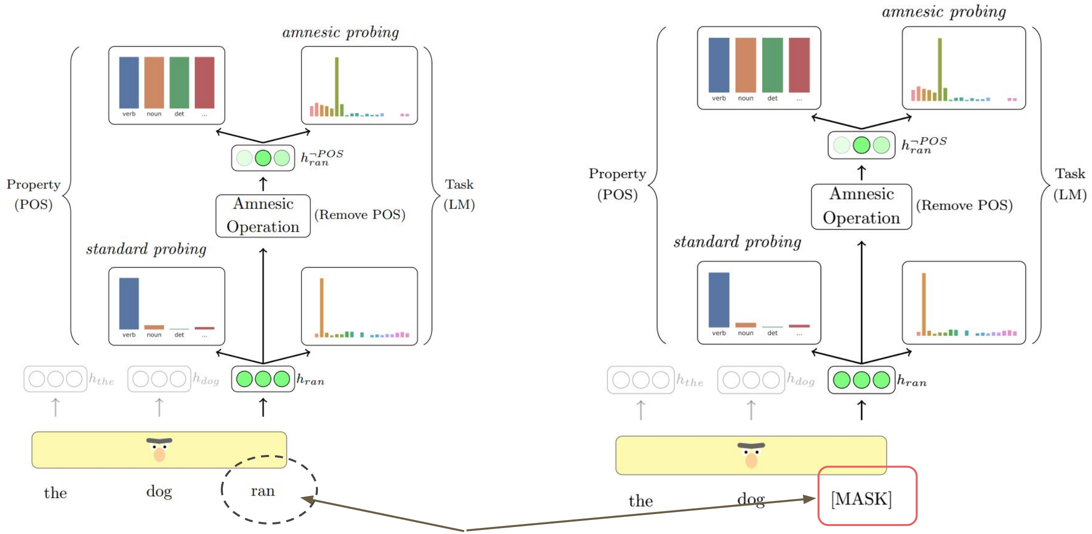
Amnesic Probing: Results

		dep	*f-pos*	*c-pos*	*ner*	*phrase start*	*phrase end*
Properties	N. dir	738	585	264	133	36	22
	N. classes	41	45	12	19	2	2
	Majority	11.44	13.22	31.76	86.09	59.25	58.51
Probing	Vanilla	76.00	89.50	92.34	93.53	85.12	83.09
LM-Acc	Vanilla	94.12	94.12	94.12	94.00	94.00	94.00
	Rand	12.31	56.47	89.65	92.56	93.75	93.86
	Selectivity	73.78	92.68	97.26	96.06	96.96	96.93
	Amnesic	7.05	12.31	61.92	83.14	94.21	94.32
LM-D_{KL}	**Rand**	8.11	4.61	0.36	0.08	0.01	0.01
	Amnesic	8.53	7.63	3.21	1.24	0.01	0.01

DKL Results

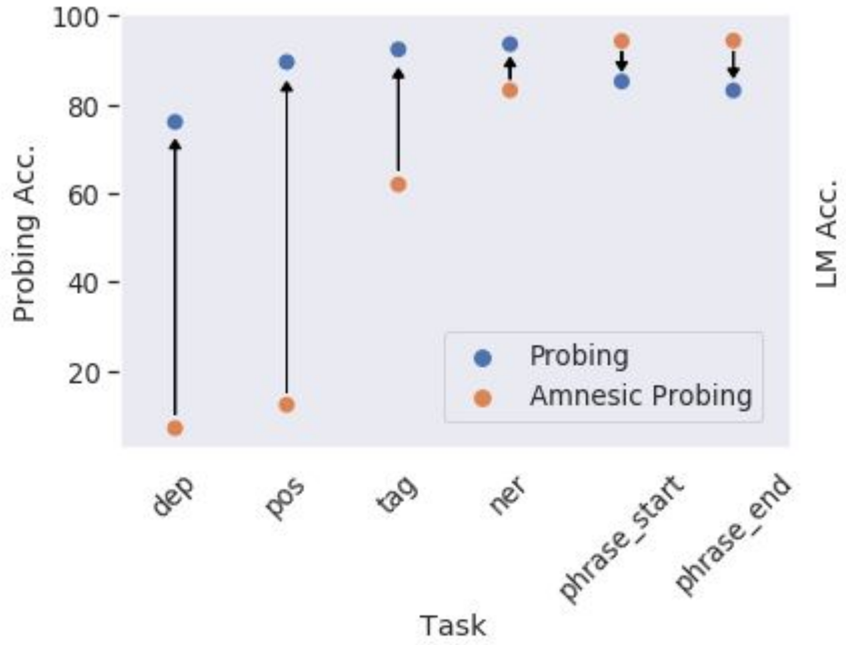
Amnesic Probing: Results

- We perform the same experiments on another setup, where the words are masked
  - (Similar results, will elaborate if time permits)



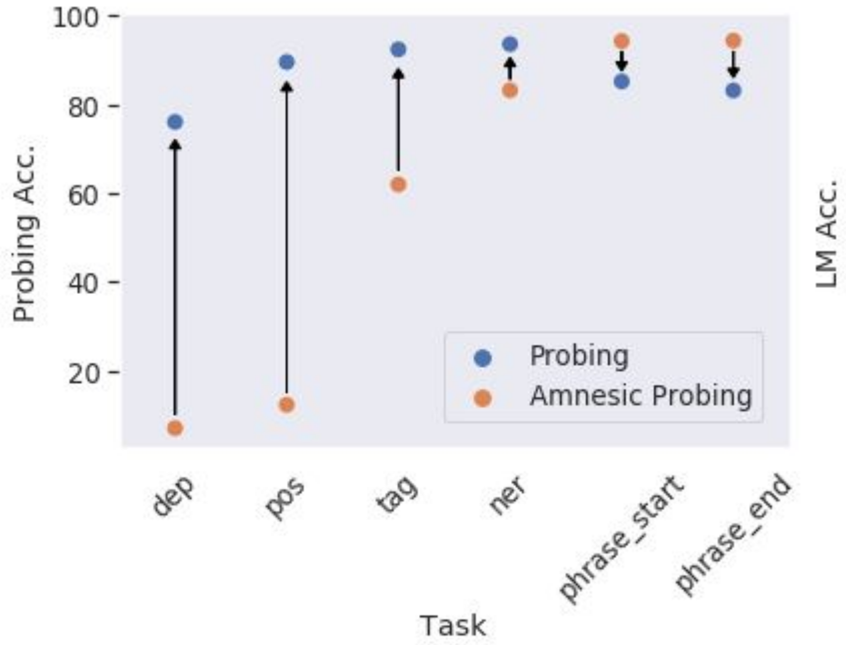
Amnesic Probing vs. Standard Probing

- We plot the probing extractability performance vs. *amnesic probing*
- We observe no correlation between the two metrics



Amnesic Probing vs. Standard Probing

- We plot the probing extractability performance vs. *amnesic probing*
- We observe no correlation between the two metrics
- **Can't make behavioural conclusions from standard probing results**



Ravichander et al., 2020, Tamkin et al., 2020

Amnesic Probing: Diving In

Amnesic Probing Fine Grained

Small changes

- How individuals POS are affected by the removal of POS information?
- Open vs. Closed vocabulary

c-pos	Vanilla	Rand	Amnesic	Δ
verb	46.72	44.85	34.99	11.73
noun	42.91	38.94	34.26	8.65
adposition	73.80	72.21	37.86	35.93
determiner	82.29	83.53	16.64	65.66
numeral	40.32	40.19	33.41	6.91
punctuation	80.71	81.02	47.03	33.68
particle	96.40	95.71	18.74	77.66
conjunction	78.01	72.94	4.28	73.73
adverb	39.84	34.11	23.71	16.14
pronoun	70.29	61.93	33.23	37.06
adjective	46.41	42.63	34.56	11.85
other	70.59	76.47	52.94	17.65

Large changes

Amnesic Probing: Inside The Model

The Inner Layers

- Until now, querying the last layer
    - INLP removes linear information, last layer is only multiplied by a matrix
- We perform the same analysis on the Inner layers
- Standard Probe (after the amnesic operation)
- Behavioral Probe

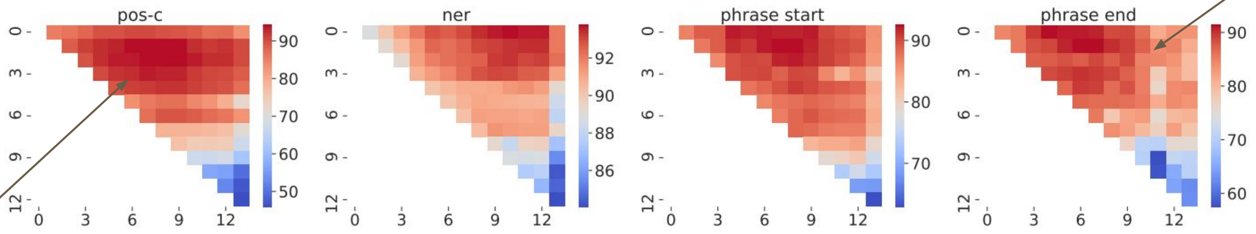
The Inner Layers: Probing

Probe scores

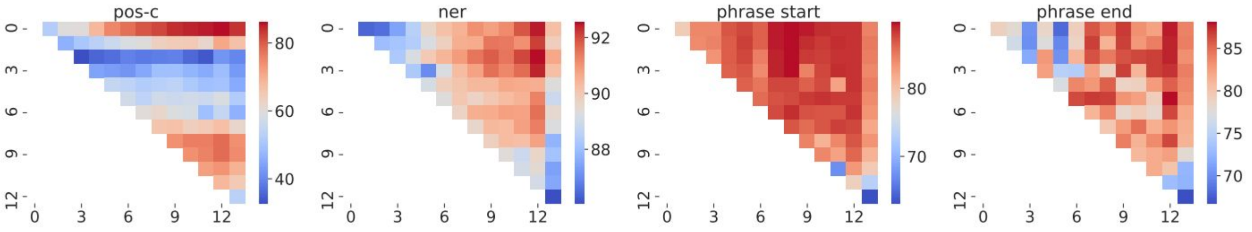
- Removing information from layer *i*, and probing in layer *j*

Remove from layer i

Probe layer j



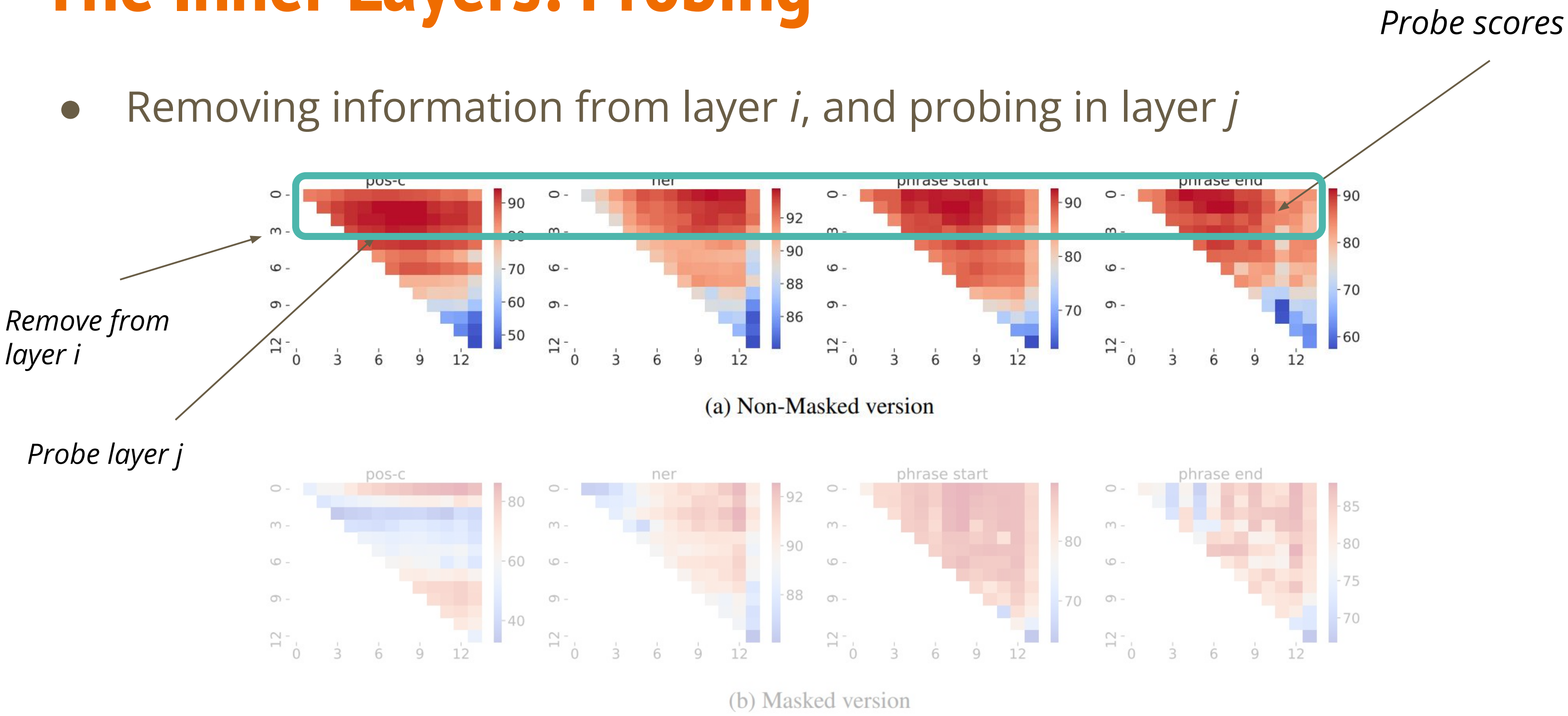
(a) Non-Masked version



(b) Masked version

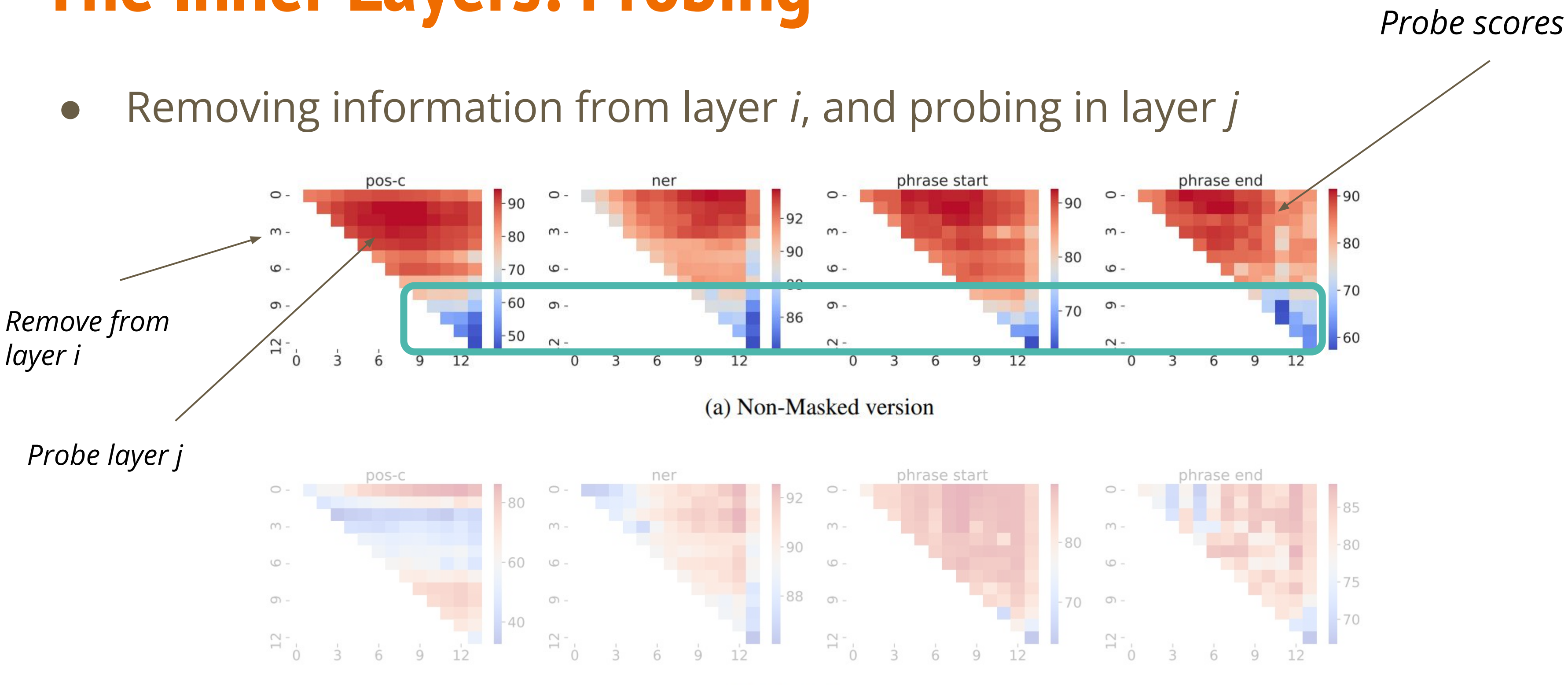
The Inner Layers: Probing

- Removing information from layer *i*, and probing in layer *j*



The Inner Layers: Probing

- Removing information from layer *i*, and probing in layer *j*



(a) Non-Masked version

(b) Masked version

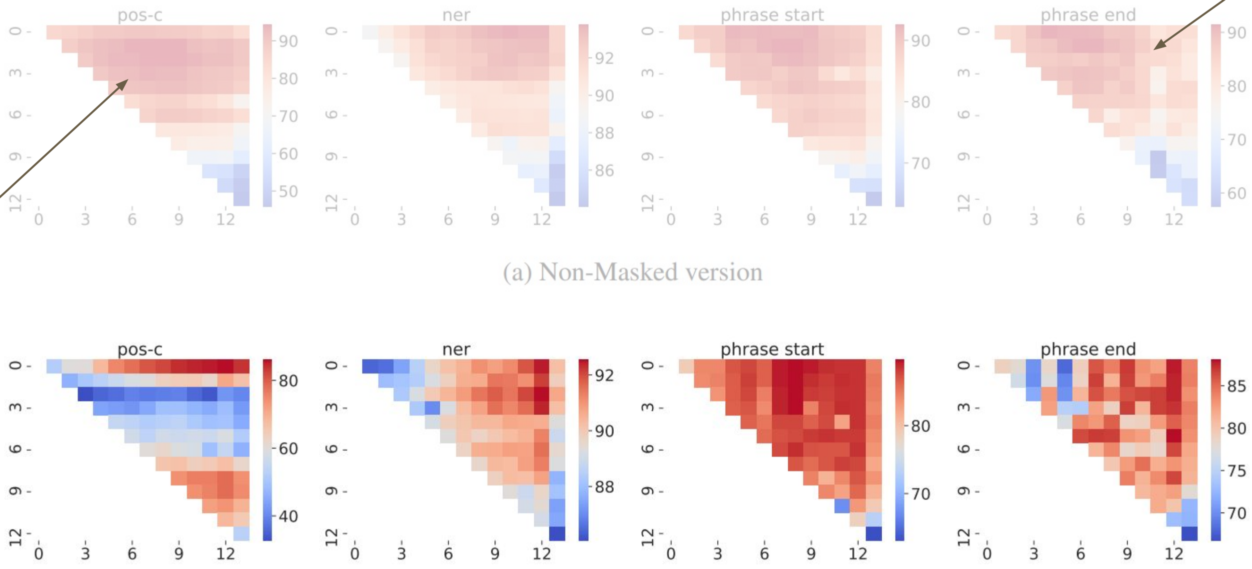
The Inner Layers: Probing

- Removing information from layer *i*, and probing in layer *j*

Probe scores

Remove from layer i

Probe layer j



(a) Non-Masked version

(b) Masked version

The Inner Layers: Probing

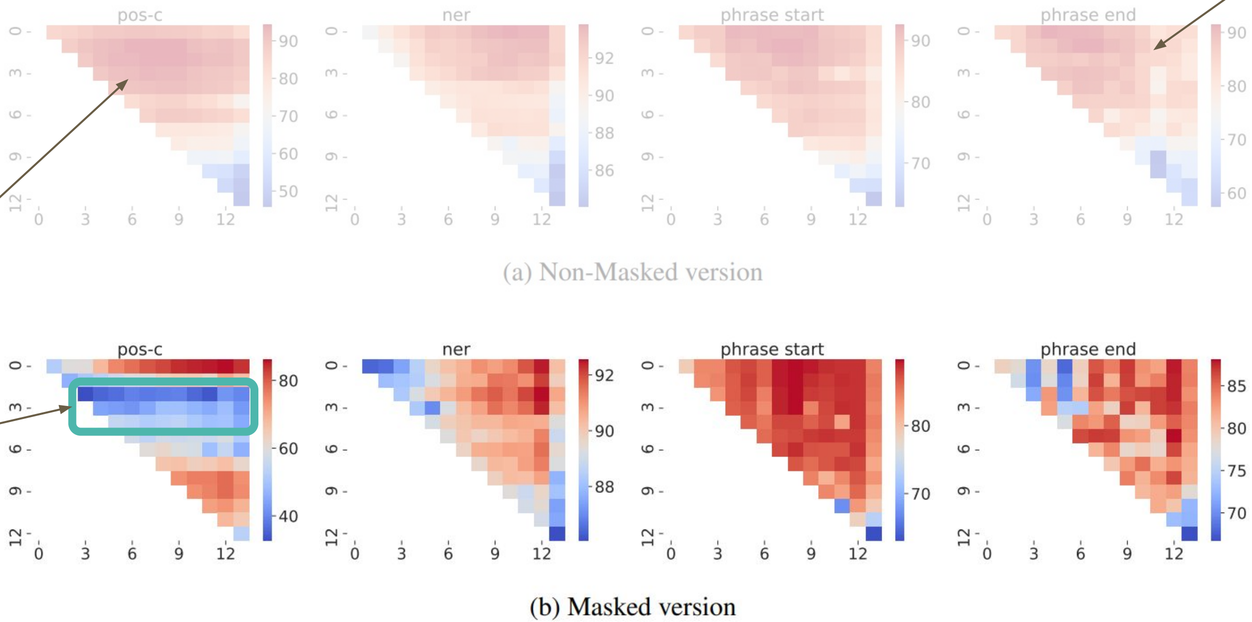
- Removing information from layer *i*, and probing in layer *j*

Probe scores

Remove from layer i

Probe layer j

Irreversible removal



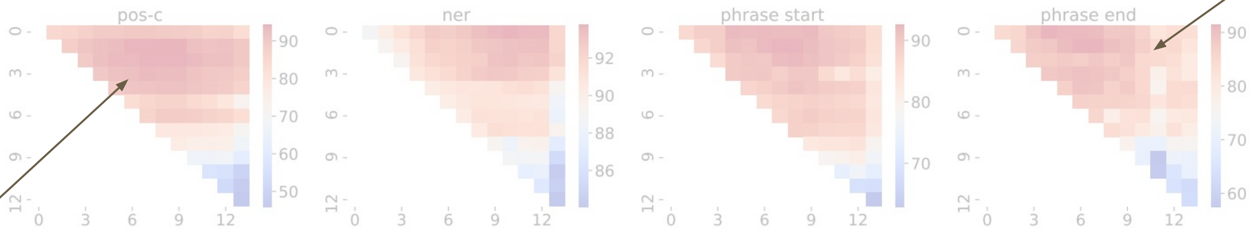
The Inner Layers: Probing

- Removing information from layer *i*, and probing in layer *j*

Probe scores

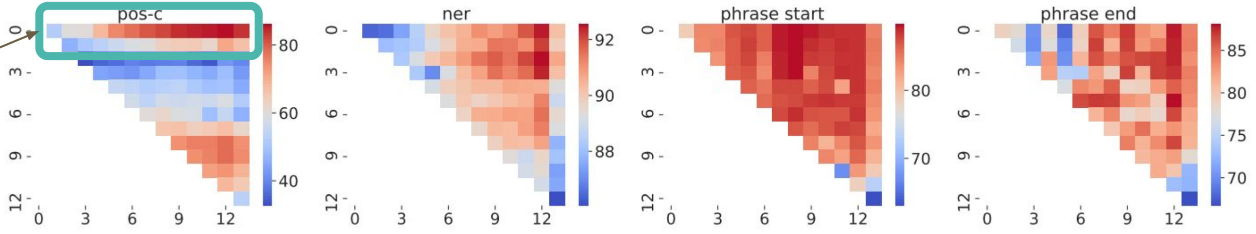
Remove from layer i

Probe layer j



(a) Non-Masked version

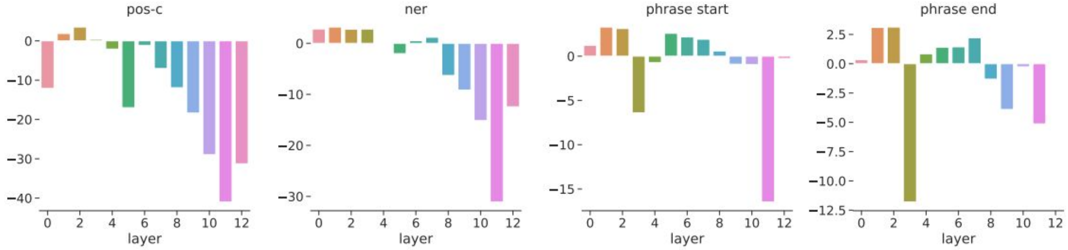
Reversible removal



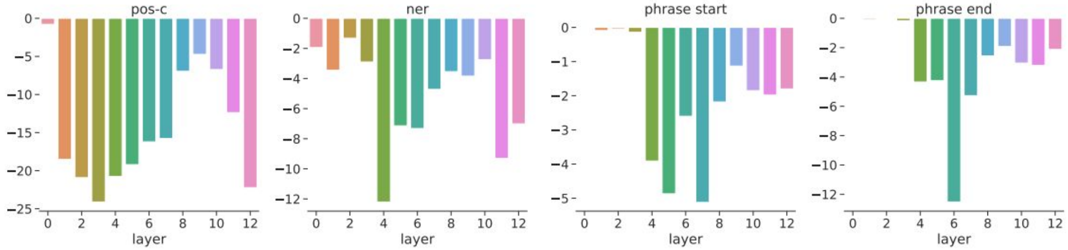
(b) Masked version

The Inner Layers: Amnesic Probing

- Removing information from layer *i*, and inspecting the model's predictions



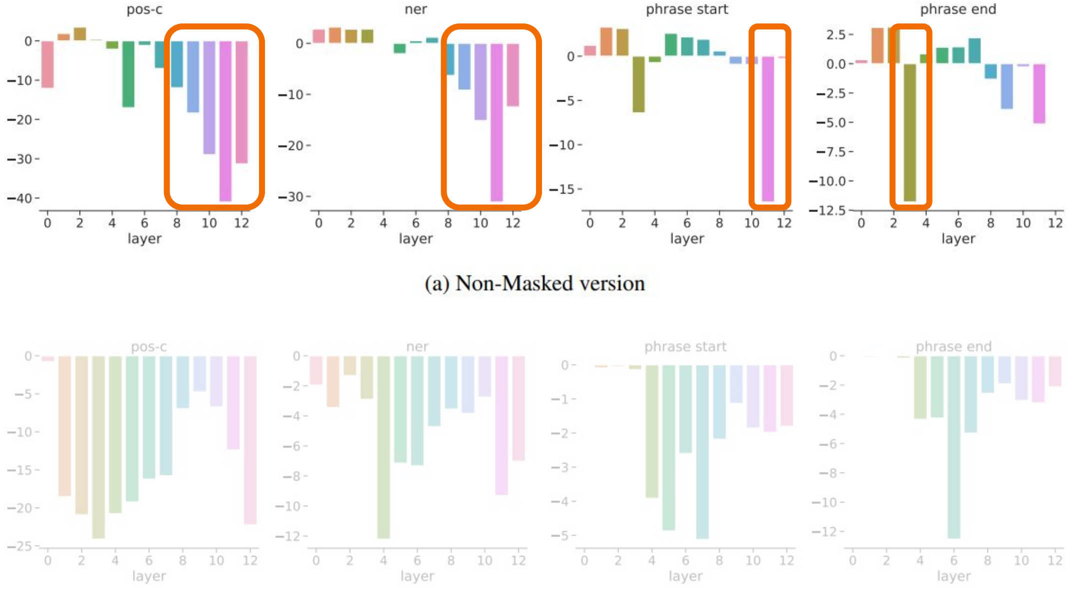
(a) Non-Masked version



(b) Masked version

The Inner Layers: Amnesic Probing

- Removing information from layer *i*, and inspecting the model's predictions

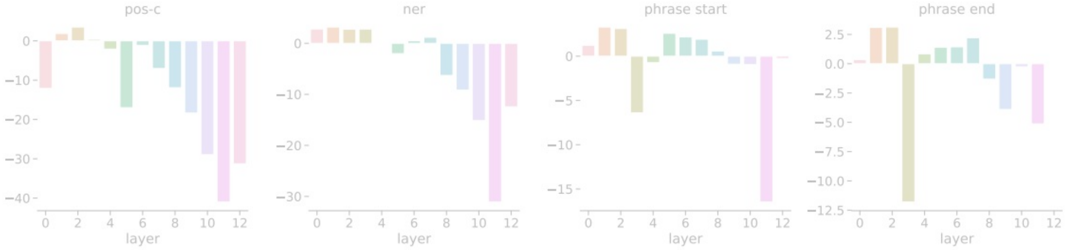


(a) Non-Masked version

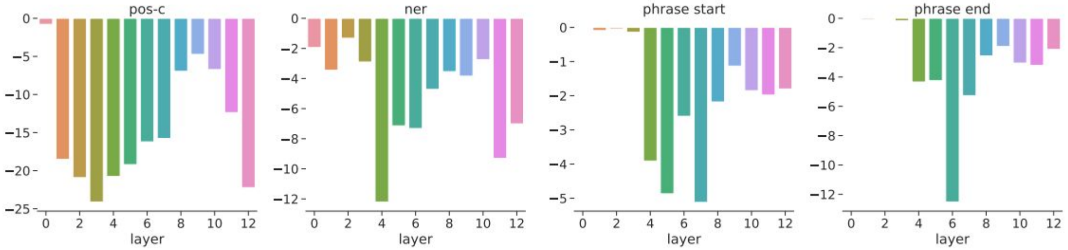
(b) Masked version

The Inner Layers: Amnesic Probing

- Removing information from layer *i*, and inspecting the model's predictions



(a) Non-Masked version

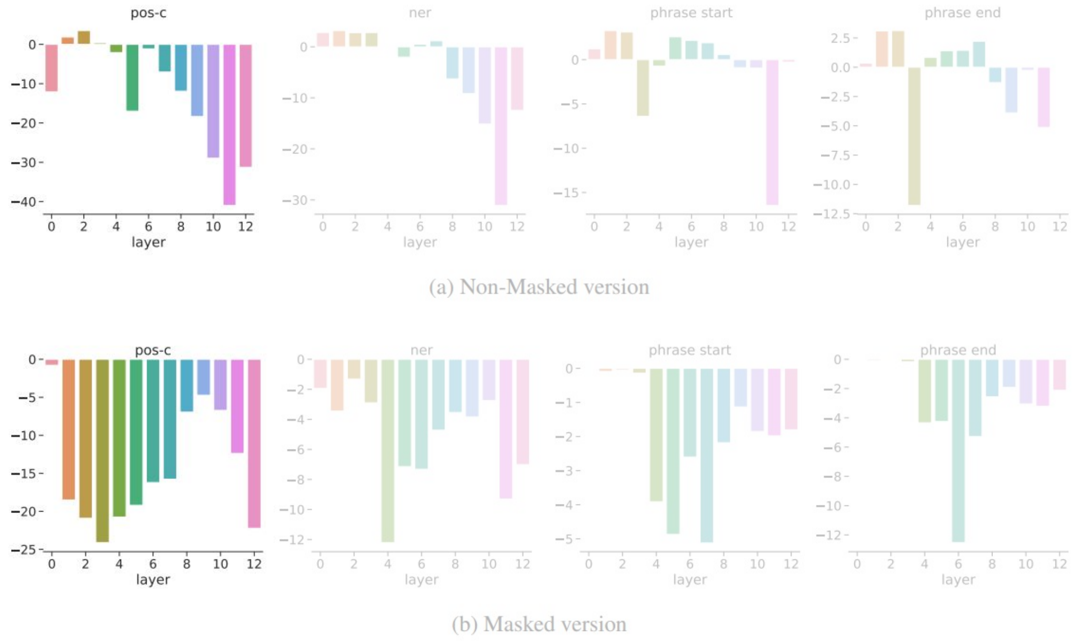


(b) Masked version

The Inner Layers: Amnesic Probing

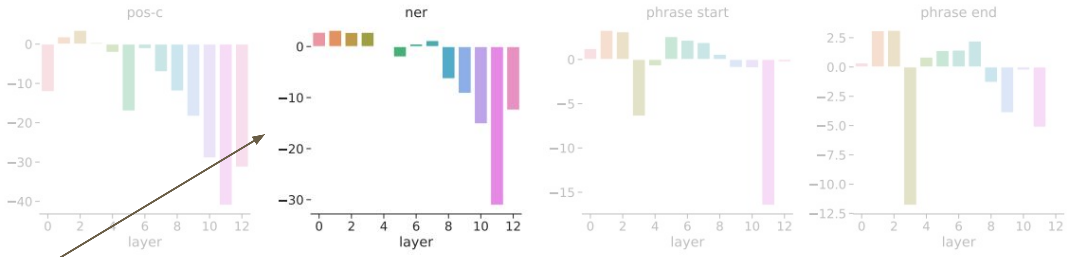
- Removing information from layer *i*, and inspecting the model's predictions

Masked vs Non-Masked



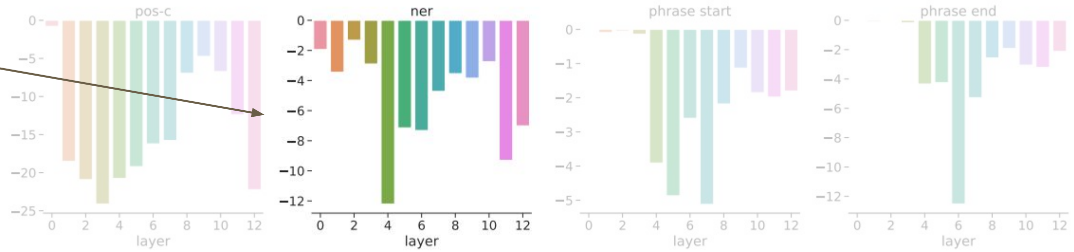
The Inner Layers: Amnesic Probing

- Removing information from layer *i*, and inspecting the model's predictions



(a) Non-Masked version

Masked vs Non-Masked

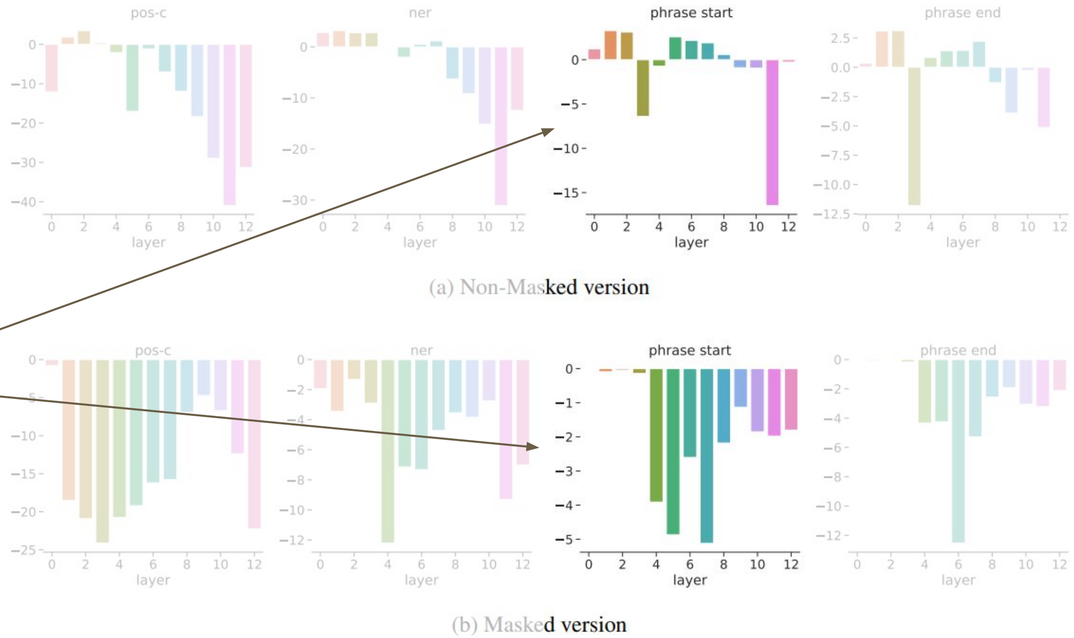


(b) Masked version

The Inner Layers: Amnesic Probing

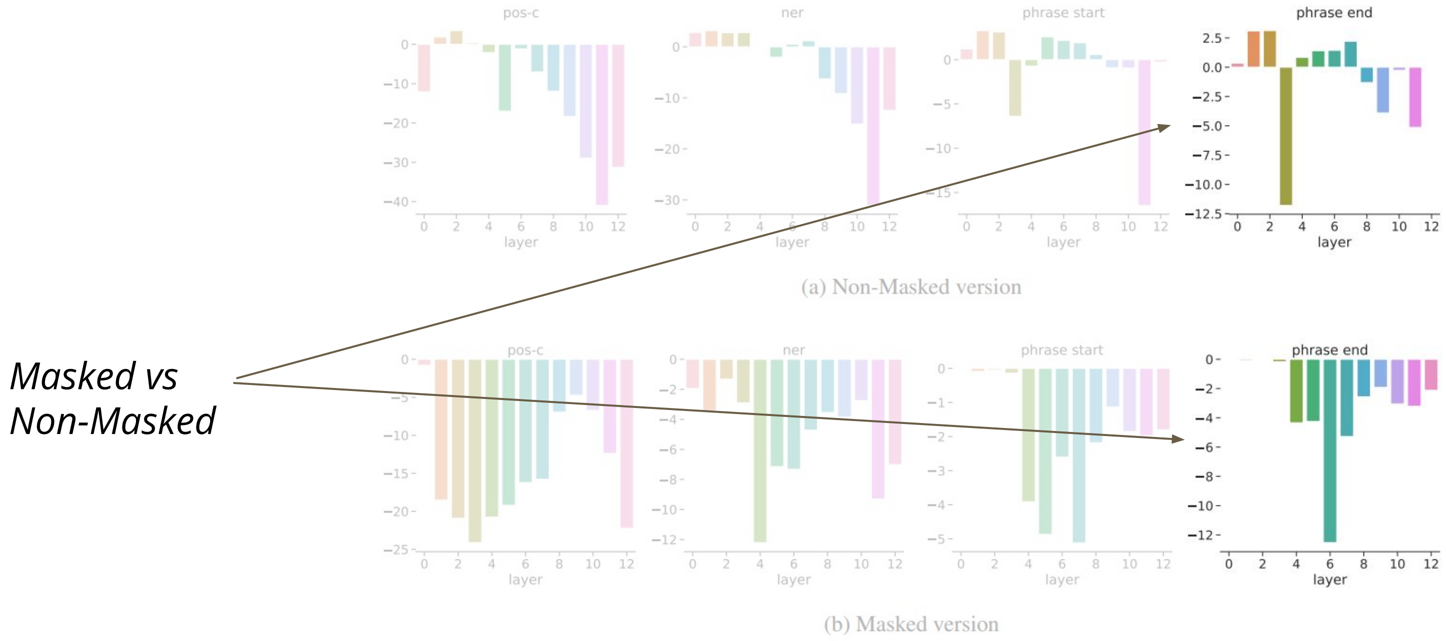
- Removing information from layer *i*, and inspecting the model's predictions

Masked vs Non-Masked



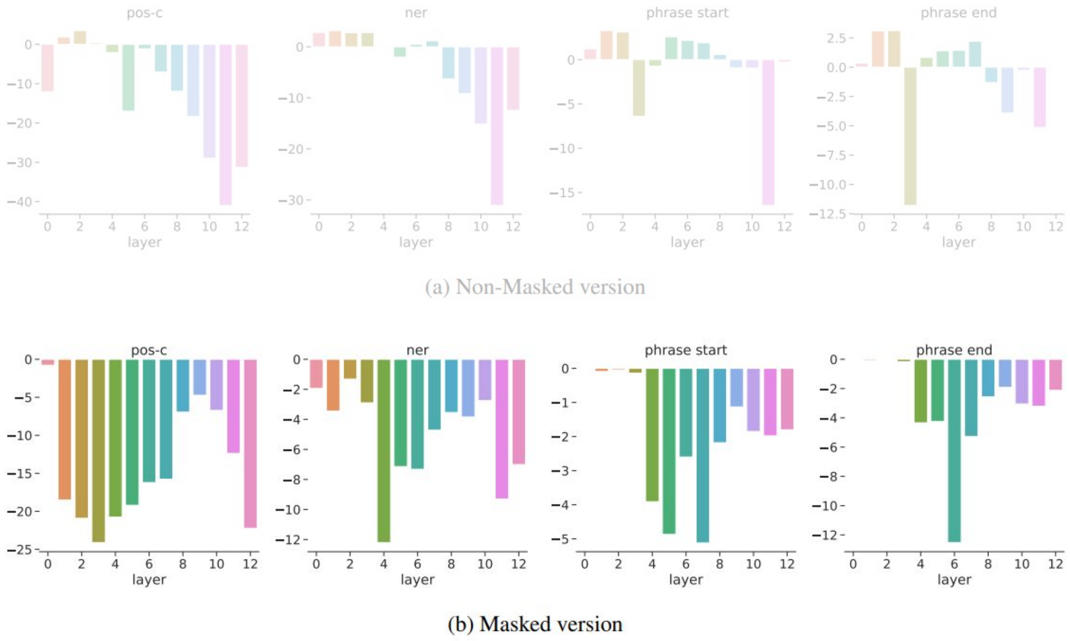
The Inner Layers: Amnesic Probing

- Removing information from layer *i*, and inspecting the model's predictions



The Inner Layers: Amnesic Probing

- Removing information from layer *i*, and inspecting the model's predictions



Strong impact in the first few layers!!

To conclude

- Probing answers the question of “what/how properties are encoded?”
- We are often interested in a **different** question: “what is being used?”
- We propose to ask the causal question and **offer a method** to answer it: ***Amnesic Probing***
- **We encourage you to use it!**

Going Forward

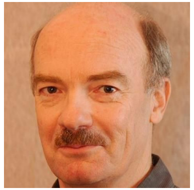
- What **does** it mean that some information is extractable?
- ... or, why is it there from the first place?
- Algorithms that remove also non-linear information

Part II

Measuring and Improving Consistency in Pretrained Language Models

Yanai Elazar[1,2] Nora Kassner[3] Shauli Ravfogel[1,2] Abhilasha Ravichander[4]
Eduard Hovy[4] Hinrich Schütze[3] Yoav Goldberg[1,2]
[1]Computer Science Department, Bar Ilan University
[2]Allen Institute for Artificial Intelligence
[3]Center for Information and Language Processing (CIS), LMU Munich
[4]Language Technologies Institute, Carnegie Mellon University



TACL 2021

Model’s Failure Mode

How many birds?	**A:** 1
Is there 1 bird?	**A:** no
Are there 2 birds?	**A:** yes
Are there any birds?	**A:** no

Ribeiro et al., 2019

Model's Failure Mode

Context: 826 Doctor Who instalments have been televised since 1963 … Starting with the 2009 special "Planet of the Dead", the series was filmed in 1080i for HDTV …

Q1: In what year did Doctor Who begin being shown in HDTV? **A**: 2009

Gan and Ng, 2019

Model's Failure Mode

Context: 826 Doctor Who instalments have been televised since 1963 ... Starting with the 2009 special "Planet of the Dead", the series was filmed in 1080i for HDTV ...

Q1: In what year did Doctor Who begin being shown in HDTV? **A**: 2009

Inconsistent

Q2: Since what year has Doctor Who been televised in HDTV? **A**: 1963

Gan and Ng, 2019

Model's Failure Mode

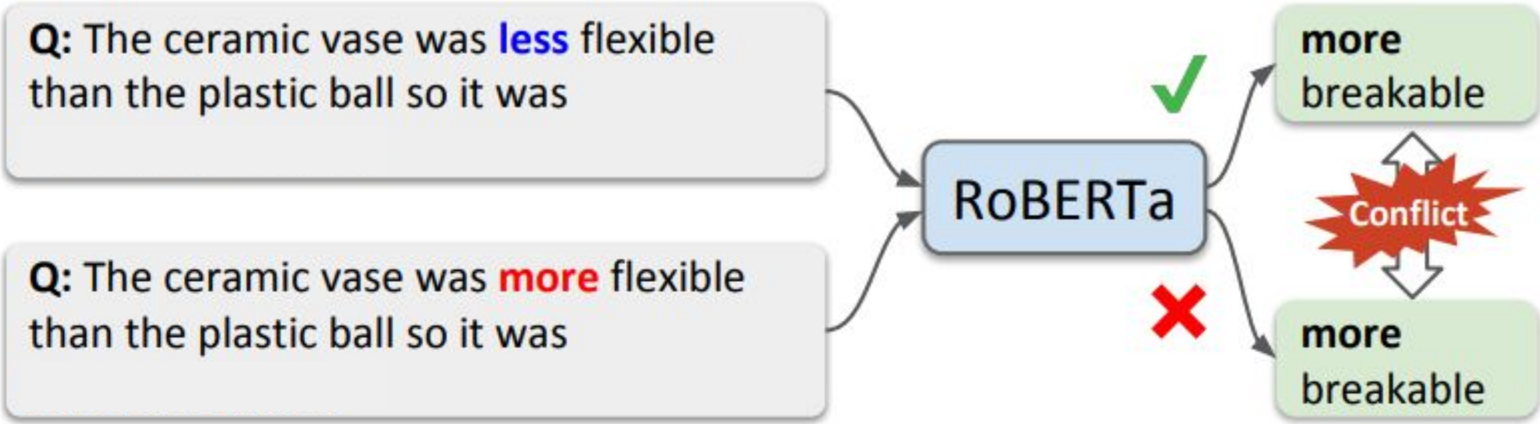
Kublai originally named his eldest son, Zhenjin, as the Crown Prince, but he died before Kublai in 1285.

(c) Excerpt from an input paragraph, **SQuAD dataset**.

Q: When did Zhenjin die?	**A:** 1285
Q: Who died in 1285?	**A:** Kublai

Ribeiro et al., 2019

Model's Failure Mode



Asai and Hajishirzi, 2020

Model's Failure Mode

Context	Match
A robin is a __	*bird*
A robin is not a __	*bird*

Ettinger, 2020

Model’s Failure Mode

Entail

P: John is on a train to Berlin.

H: John is traveling to Berlin.

Contradict

Z: John is having lunch in Berlin.

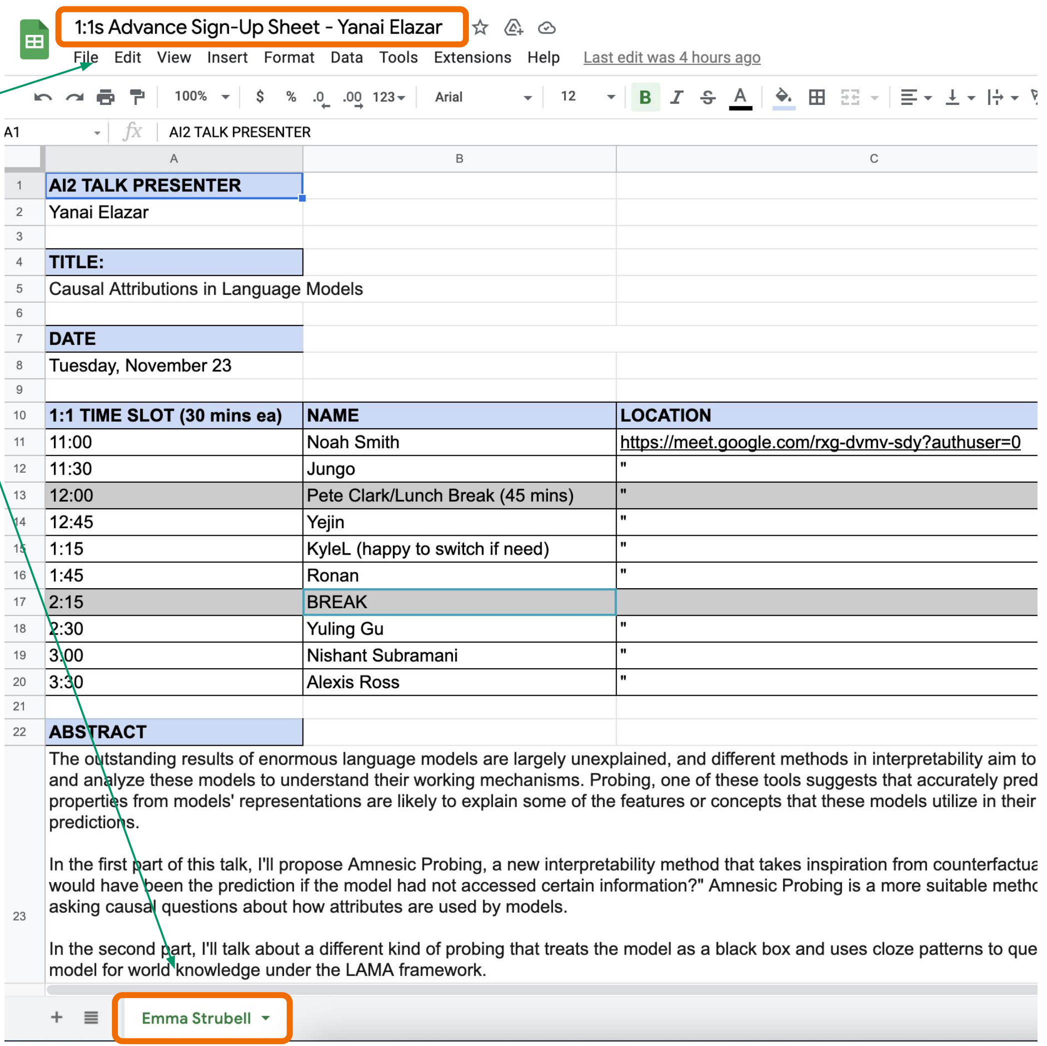
~~Entail~~

Li et al., 2020

Consistency in Models

- End-task models suffer from inconsistency
- Today’s standard pipeline is: Pretrain -> Finetune
- **In this work**: we show that *Inconsistency starts in the PLM itself*

Consistency in Humans



1:1s Advance Sign-Up Sheet - Yanai Elazar

File Edit View Insert Format Data Tools Extensions Help Last edit was 4 hours ago

A1 AI2 TALK PRESENTER

	A	B	C
1	**AI2 TALK PRESENTER**		
2	Yanai Elazar		
3			
4	**TITLE:**		
5	Causal Attributions in Language Models		
6			
7	**DATE**		
8	Tuesday, November 23		
9			
10	**1:1 TIME SLOT (30 mins ea)**	**NAME**	**LOCATION**
11	11:00	Noah Smith	https://meet.google.com/rxg-dvmv-sdy?authuser=0
12	11:30	Jungo	"
13	12:00	Pete Clark/Lunch Break (45 mins)	"
14	12:45	Yejin	"
15	1:15	KyleL (happy to switch if need)	"
16	1:45	Ronan	"
17	2:15	BREAK	
18	2:30	Yuling Gu	"
19	3:00	Nishant Subramani	"
20	3:30	Alexis Ross	"
21			
22	**ABSTRACT**		

23 The outstanding results of enormous language models are largely unexplained, and different methods in interpretability aim to and analyze these models to understand their working mechanisms. Probing, one of these tools suggests that accurately pred properties from models' representations are likely to explain some of the features or concepts that these models utilize in their predictions.

In the first part of this talk, I'll propose Amnesic Probing, a new interpretability method that takes inspiration from counterfactua would have been the prediction if the model had not accessed certain information?" Amnesic Probing is a more suitable metho asking causal questions about how attributes are used by models.

In the second part, I'll talk about a different kind of probing that treats the model as a black box and uses cloze patterns to que model for world knowledge under the LAMA framework.

Emma Strubell

Consistency in Humans

Yanai Elazar
Bar Ilan University AI2
yanaiela@gmail.com

Yanai Elazar[1,3] **Yoav Goldberg**[1,3]
[1]The Allen Institute for AI

er[1,2] **Yanai Elazar**[3,4] **Benoît Sagot**[1]
’aris, France [2]Sorbonne Université, Paris
nputer Science Department, Bar Ilan Unive
[4]Allen Institute for Artificial Intelligence

Consistency in Models: This Part

1. Language Models as Knowledge Bases
2. Why is consistency crucial?

background

Consistency in Models: This Part

1. Language Models as Knowledge Bases
2. Why is consistency crucial?

background

3. ParaRel 🤘: a new resource that enables us to measure consistency

Consistency in Models: This Part

1. Language Models as Knowledge Bases
2. Why is consistency crucial?

background

3. ParaRel🤘: a new resource that enables us to measure consistency
4. A framework for measuring (In)Consistency in Language Models
    - In the context of factual knowledge

Consistency in Models: This Part

1. Language Models as Knowledge Bases
2. Why is consistency crucial?

background

3. ParaRel 🤘: a new resource that enables us to measure consistency
4. A framework for measuring (In)Consistency in Language Models
    - In the context of factual knowledge

5. A proposal to improve consistency in LMs.

Consistency in Models: This Part

1. Language Models as Knowledge Bases
2. Why is consistency crucial?

background

3. ParaRel🤘: a new resource that enables us to measure consistency
4. A framework for measuring (In)Consistency in Language Models
    - In the context of factual knowledge

5. A proposal to improve consistency in LMs.

novelty

Setup: LMs as Knowledge Bases

Language Models as Knowledge Bases?

Fabio Petroni[1] **Tim Rocktäschel**[1,2] **Patrick Lewis**[1,2] **Anton Bakhtin**[1]
Yuxiang Wu[1,2] **Alexander H. Miller**[1] **Sebastian Riedel**[1,2]
[1]Facebook AI Research
[2]University College London
{fabiopetroni, rockt, plewis, yolo, yuxiangwu, ahm, sriedel}@fb.com

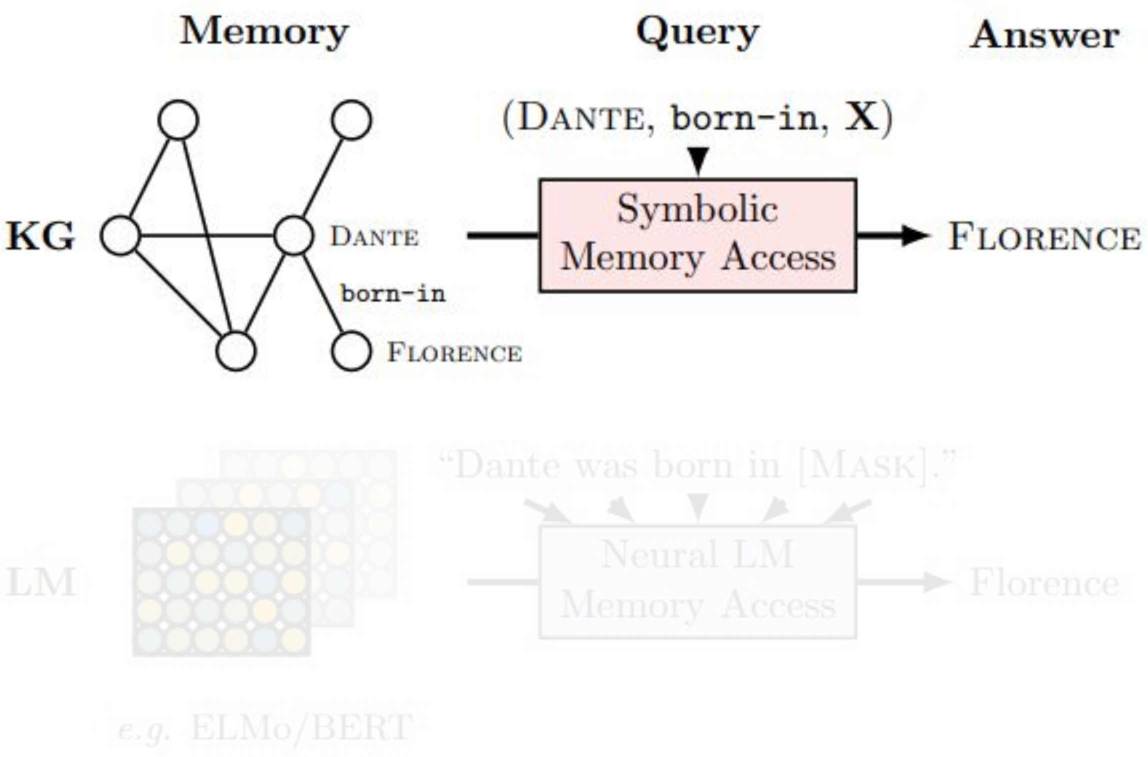


Figure 1: Querying knowledge bases (KB) and language models (LM) for factual knowledge.

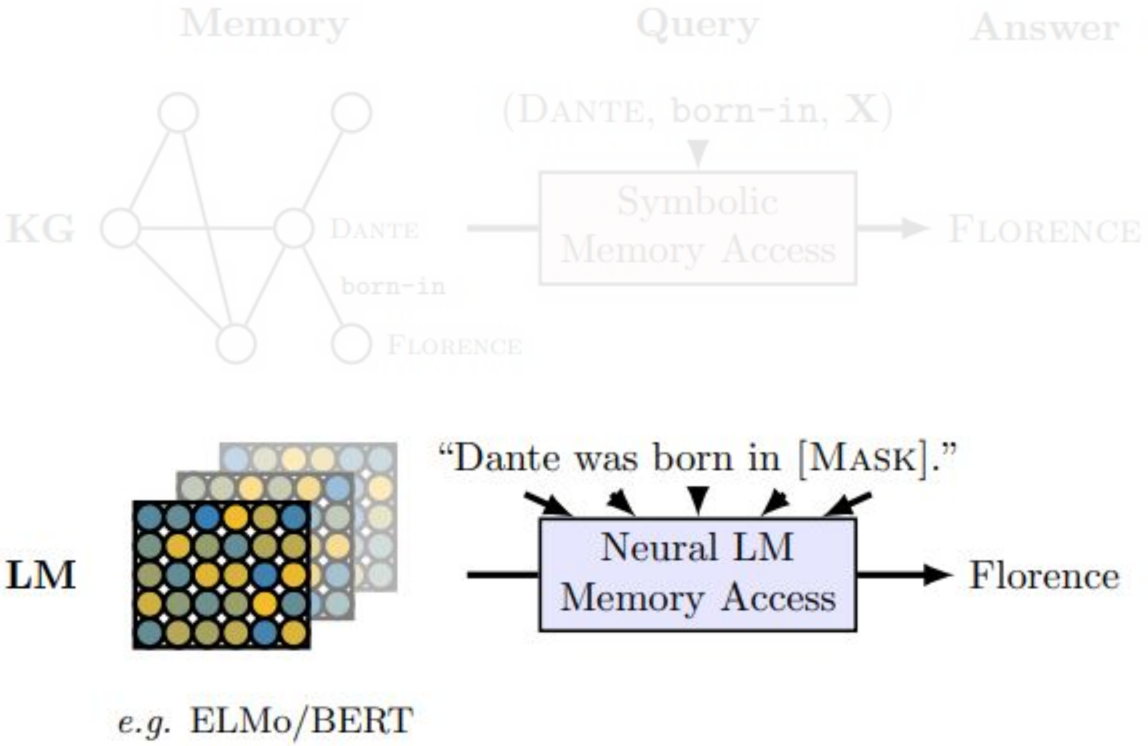


Figure 1: Querying knowledge bases (KB) and language models (LM) for factual knowledge.

Using Patterns to Query LMs

- Born-In: “[X] was born in [Y] .”
    - *Barack Obama was born in [MASK].*
- Broadcasting Channel: “[X] was originally aired on [Y] .”
    - *Lost was originally aired on [MASK].*
- …

Language Models as KBs - Setup

- The data is of the form <subject, relation, object>
  - E.g. <“Barack Obama”, “born-in”, “Hawaii”>
- To query an LM, we write a ‘pattern’ that expresses a relation
  - E.g. “*[X]* was born in *[Y]*”
- Given the subject and relation, the task is to predict the object
  - E.g. <“Barack Obama”, born-in> -> “Hawaii”
  - In Petroni et al., 2019, used 1 pattern for every relation

Corpus	Relation	Statistics		Baselines		KB		LM					
		#Facts	#Rel	Freq	DrQA	RE_n	RE_o	Fs	Txl	Eb	E5B	Bb	Bl
Google-RE	`birth-place`	2937	1	4.6	-	3.5	13.8	4.4	2.7	5.5	7.5	14.9	**16.1**
	`birth-date`	1825	1	1.9	-	0.0	**1.9**	0.3	1.1	0.1	0.1	1.5	1.4
	`death-place`	765	1	6.8	-	0.1	7.2	3.0	0.9	0.3	1.3	13.1	**14.0**
	Total	5527	3	4.4	-	1.2	7.6	2.6	1.6	2.0	3.0	9.8	**10.5**
T-REx	1-1	937	2	1.78	-	0.6	10.0	17.0	36.5	10.1	13.1	68.0	**74.5**
	N-1	20006	23	23.85	-	5.4	**33.8**	6.1	18.0	3.6	6.5	32.4	34.2
	N-*M*	13096	16	21.95	-	7.7	**36.7**	12.0	16.5	5.7	7.4	24.7	24.3
	Total	34039	41	22.03	-	6.1	**33.8**	8.9	18.3	4.7	7.1	31.1	32.3
ConceptNet	Total	11458	16	4.8	-	-	-	3.6	5.7	6.1	6.2	15.6	**19.2**
SQuAD	Total	305	-	-	**37.5**	-	-	3.6	3.9	1.6	4.3	14.1	17.4

Language Models as KBs

- LMs were trained on large sources of knowledge (e.g. Wikipedia)
- Can capture (memorize) some of these facts as part of the pretraining objective

Pretraining a Language Model

Background

Early life of Barack Obama

Main articles: Early life and career of Barack Obama and Ann Dunham

People who express doubts about Obama's eligibility or reject details about his early life are often informally called "birthers", a term that parallels[23] the nickname "truthers" for adherents of 9/11 conspiracy theories.[24][25] These conspiracy theorists reject at least some of the following facts about his early life:

Barack Obama was born on August 4, 1961, at Kapi'olani Maternity & Gynecological Hospital (now called Kapi'olani Medical Center for Women & Children) in Honolulu, Hawaii,[26][27][28][29] to Ann Dunham,[30] from Wichita, Kansas,[31] and her husband Barack Obama Sr., a Luo from Nyang'oma Kogelo, Nyanza Province (in what was then the Colony and Protectorate of Kenya), who was attending the University of Hawaii. Birth notices for Barack Obama were published in *The Honolulu Advertiser* on August 13 and the *Honolulu Star-Bulletin* on August 14, 1961.[26][31] Obama's father's immigration file also clearly states Barack Obama was born in Hawaii.[32] One of his high school teachers, who was acquainted with his mother at the time, remembered hearing about the day of his birth.[30]

Pretraining a Language Model

And it actually works!

LM predictions

#1 mask:Tel Aviv is located in **[MASK]**.

	bert_large_cased
0	Israel
1	Jerusalem
2	Palestine
3	Haifa
4	Egypt
5	Europe
6	Ukraine
7	Lebanon
8	Jordan
9	Germany

Pretraining a Language Model

Well, sometimes…

LM predictions

#1 mask:Barack Obama was born in **[MASK]**.

	bert_large_cased
0	Chicago
1	Philadelphia
2	Detroit
3	Houston
4	Atlanta
5	Georgia
6	Boston
7	Texas
8	Paris
9	Dallas

Pretraining a Language Model

Well, sometimes...

LM predictions

#1 mask:Barack Obama was born in **[MASK]**.

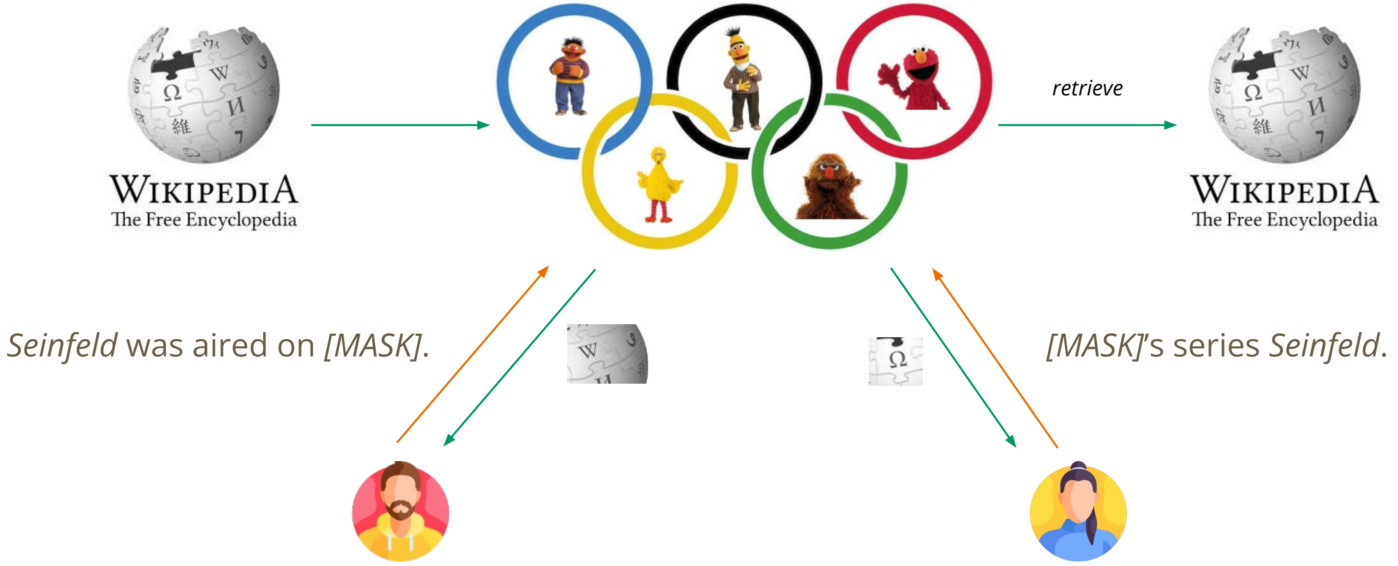
	bert_large_cased
0	Chicago
1	Philadelphia
2	Detroit
3	Houston
4	Atlanta
5	Georgia
6	Boston
7	Texas
8	Paris
9	Dallas

Teaser:
we'll get back to the reason behind this prediction in Part III

Language Models as KBs - Setup

- Restricting to MLM predictions: single token objects
- Restricting to the possible objects for a specific relation

Language Models as KBs



Language Models as KBs

So the real question is

Does It Generalize?

Language Models as KBs - Consistency?

We'd like that an LM would make the same prediction across paraphrases

E.g.:

"*Seinfeld* was aired on [Y]."

- ↔️ "*Seinfeld*, that was aired on [Y],"
- ↔️ "[Y]'s series *Seinfeld*,"

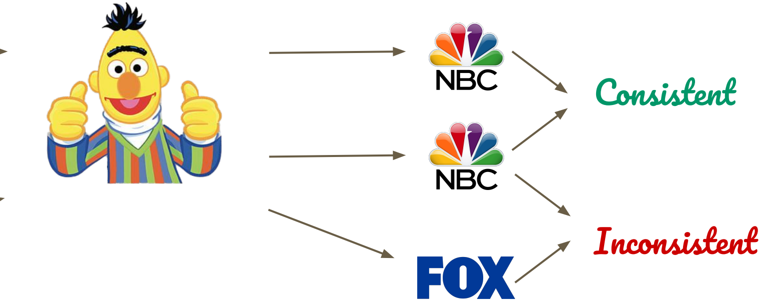
Language Models as KBs - Consistency?

We'd like that an LM would make the same prediction across paraphrases

E.g.:

"*Seinfeld* was aired on [Y]."

- ↔️ "*Seinfeld*, that was aired on [Y],"
- ↔️ "[Y]'s series *Seinfeld*,"



Measuring Consistency:

ParaRel 🤘

Language Models as KBs - ParaRel 🤘

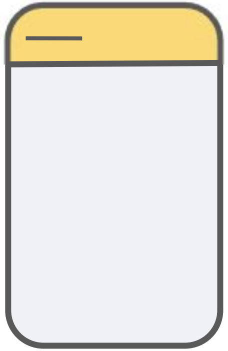
But where can we get these patterns?

We build a new resource:

ParaRel🤘 (**Para**phrase **Rel**ations)

ParaRel 🤘 - Creation

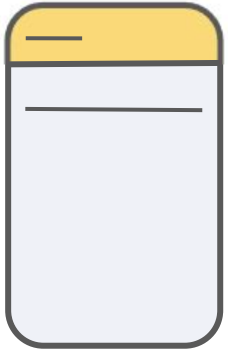
- For every relation, we manually build a set of patterns that are paraphrases of each other, in 4 steps:



ParaRel 🤘 - Creation

(a) A single pattern

- For every relation, we manually build a set of patterns that are paraphrases of each other, in 4 steps:
    a. Starting with the LAMA patterns (*Petroni et al., 2019*)



ParaRel 🤘 - Creation

(a) A single pattern

- For every relation, we manually build a set of patterns that are paraphrases of each other, in 4 steps:
  a. Starting with the LAMA patterns (*Petroni et al., 2019*)
  b. Augmenting with LPAQA patterns (*Jiang et al., 2020*)

(b) Multiple patterns, noisy

ParaRel 🤘 - Creation

- For every relation, we manually build a set of patterns that are paraphrases of each other, in 4 steps:
  a. Starting with the LAMA patterns (*Petroni et al., 2019*)
  b. Augmenting with LPAQA patterns (*Jiang et al., 2020*)
  c. Searching for patterns in wikipedia using SPIKE (*Shlain et al., 2020*)

(a) A single pattern

LAMA

(b) Multiple patterns, noisy

(c) Searching for syntactic patterns

SPIKE

ParaRel 🤘 - Creation

- For every relation, we manually build a set of patterns that are paraphrases of each other, in 4 steps:
  a. Starting with the LAMA patterns (*Petroni et al., 2019*)
  b. Augmenting with LPAQA patterns (*Jiang et al., 2020*)
  c. Searching for patterns in wikipedia using SPIKE (*Shlain et al., 2020*)
  d. Additional patterns using linguistic expertise

(a) A single pattern

(b) Multiple patterns, noisy

(c) Searching for syntactic patterns

(d) linguistic expertise, expanding previous patterns

ParaRel 🤘 - Summary

aired-on

[X] was aired on *[Y]*.
[X], that was aired on *[Y]*.
[Y]'s series *[X]*

employer

[X] used to work in *[Y]*.
[X] found employment in *[Y]*.
[X] took up work in *[Y]*.

# Relations	38
# Patterns	328
Min # patterns	2
Max # patterns	20
Avg # patterns	8.63

instrument

[X] plays *[Y]*.
[Y] player *[X]*.
[X] is a *[Y]* player.

twin-cities

[X] and *[Y]* are twin cities.
[Y] and *[X]* are twin cities.
[X] is a twin city of *[Y]*.

ParaRel 🤘 - Creation

- For every relation, we manually build a set of patterns that are paraphrases of each other, in 4 steps:
  - Starting with the single pattern from LAMA (Petroni et al., 2019)
  - Augmenting with automatically extracted patterns from LPAQA (Jiang et al., 2020)
  - Searching for patterns in wikipedia using SPIKE (Shlain et al., 2020)
  - Additional patterns using linguistic expertise of the authors

ParaRel 🤘 - Verification

- Was collected manually by the authors of this paper
- 2 additional authors verified the quality, while engaging in discussion to reach an agreement (discarding otherwise)
- Human Eval: Sampled 156 pairs, and ask NLP grad students to annotate. Reaching **95.5%** agreement (and later fixed the errors)

Setup & Evaluation

Consistency - Setup

Data Pairs (D) $\quad (D_1, r_1, P_1), \ldots, (D_i, r_i, P_i), \ldots, (D_n, r_n, P_n) \quad$ **Patterns** (P)

D_1
(*Lou Reed*, *Brooklyn*)
(*Masako Natsume*, *Tokyo*)
...

...

D_i
(*Seinfeld*, *NBC*)
(*Homeland*, *Showtime*)
...

...

r_i =*originally-aired-on*

Homeland originally aired on *[MASK]*
Homeland premiered on *[MASK]*
...
Seinfeld originally aired on *[MASK]*
Seinfeld premiered on *[MASK]*

P_1
(X was born in Y)
(X is native to Y)
...

...

P_i
(X originally aired in Y)
(X premiered on Y)
...

...

Consistency - Setup

Data Pairs (D) $(D_1, r_1, P_1), \ldots, (D_i, r_i, P_i), \ldots, (D_n, r_n, P_n)$ **Patterns (P)**

D_1
(*Lou Reed, Brooklyn*)
(*Masako Natsume, Tokyo*)
...

...

D_i
(*Seinfeld, NBC*)
(*Homeland, Showtime*)
...

...

r_i =*originally-aired-on*

Homeland originally aired on *[MASK]*
Homeland premiered on *[MASK]*
...
Seinfeld originally aired on *[MASK]*
Seinfeld premiered on *[MASK]*

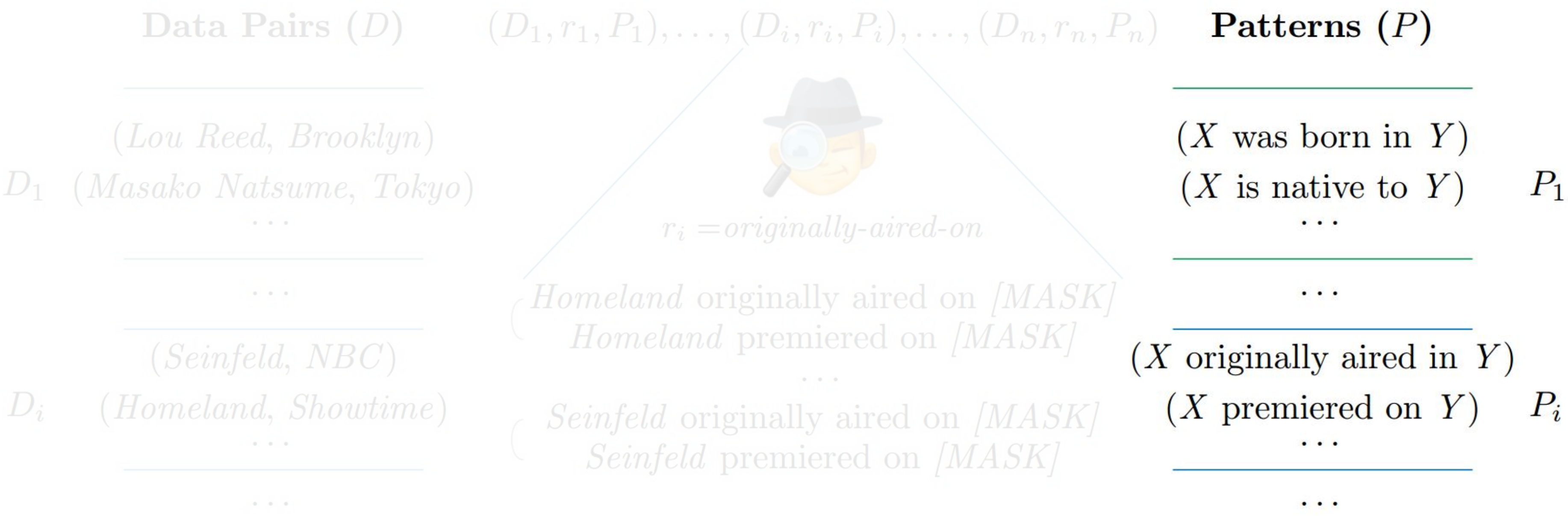
P_1
(X was born in Y)
(X is native to Y)
...

...

P_i
(X originally aired in Y)
(X premiered on Y)
...

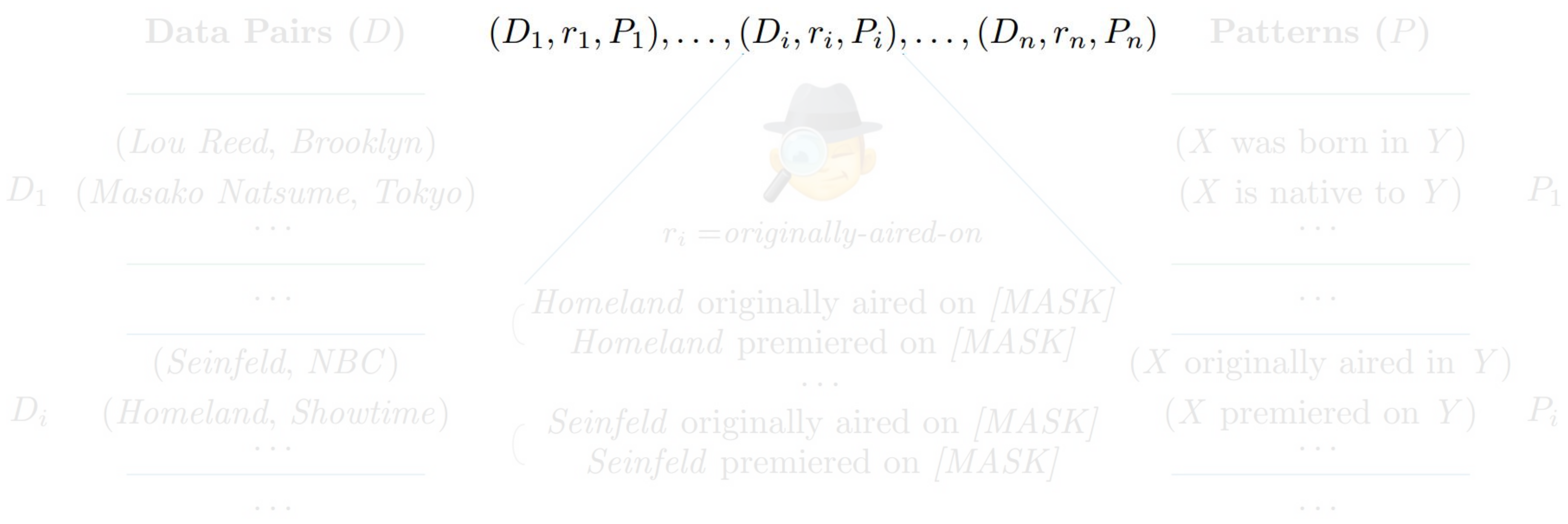
...

Consistency - Setup

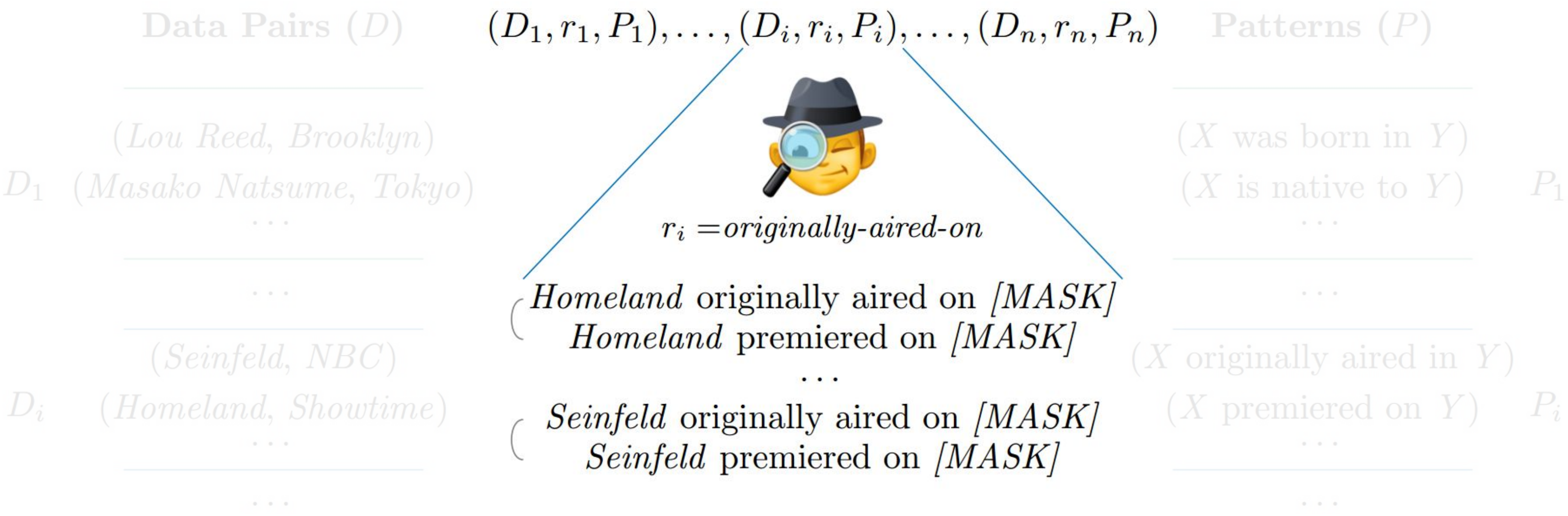


Consistency - Setup

$$(D_1, r_1, P_1), \ldots, (D_i, r_i, P_i), \ldots, (D_n, r_n, P_n)$$



Consistency - Setup



Consistency - Models

- BERT
- BERT Whole-Word-Masking
- RoBERTa
- ALBERT

And a Baseline:

- Most common object (consistent by definition)

Consistency - Evaluation

- **Accuracy**: Accurate prediction of the LAMA pattern
- **Consistency**: For each relation and tuple, compute all paraphrases pairs, and test if the predictions are equal: n(n-1)/2 pairs
- **Consistent-Acc**: Consistent and accurate prediction of all paraphrases

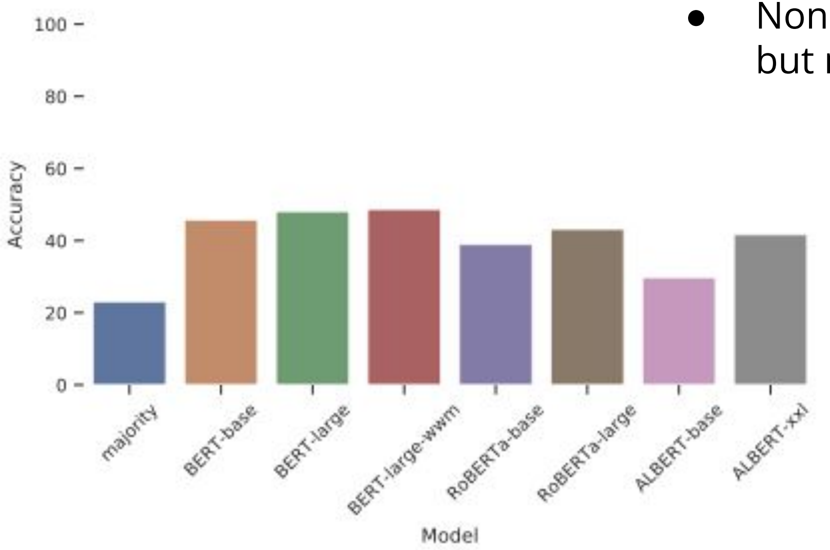
Results

Consistency - Results

Are LMs Consistent?

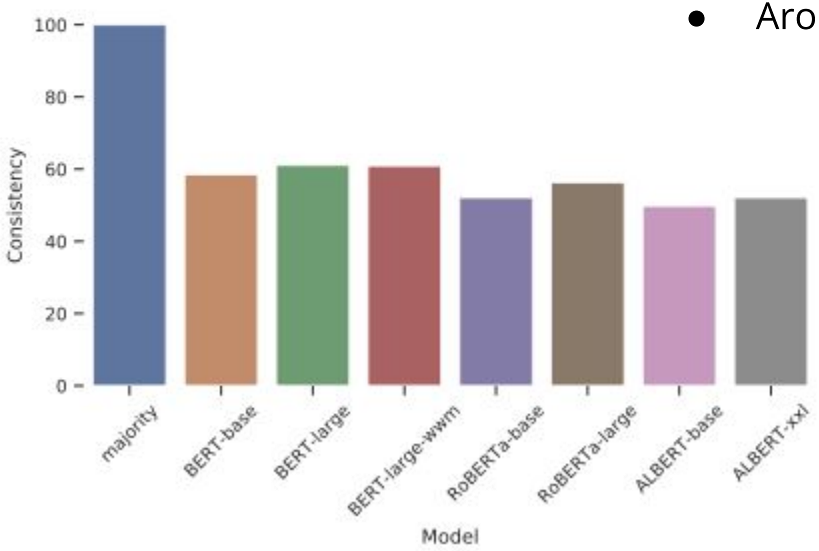
No!

Consistency - Results



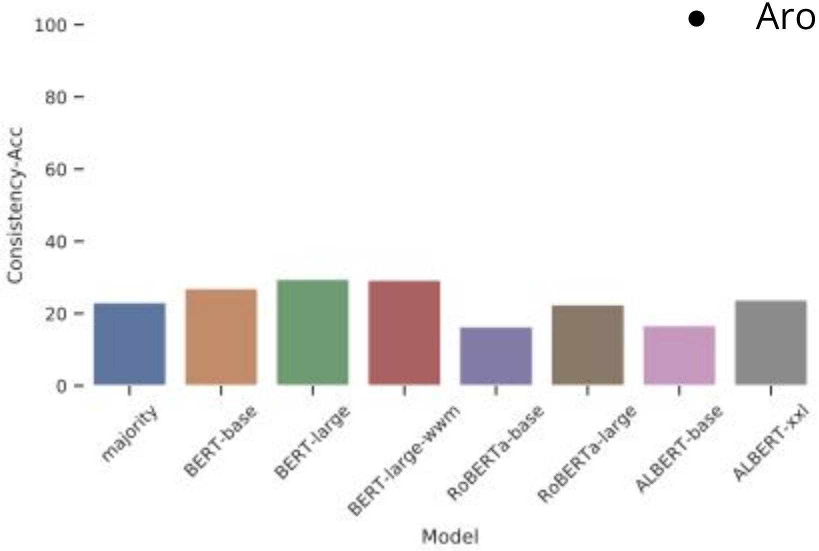
- LAMA accuracy performance
- Non-trivial retrieval abilities (~40%), but not good in any way

Consistency - Results



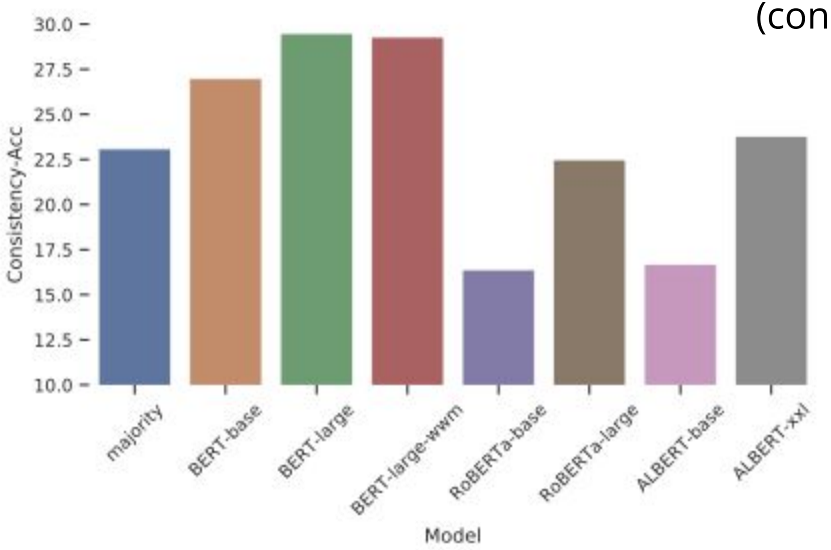
- Consistency results
- Around 50%, not consistent!

Consistency - Results



- Consistent-accurate results
- Around 20-30%, much worse!

Consistency - Results

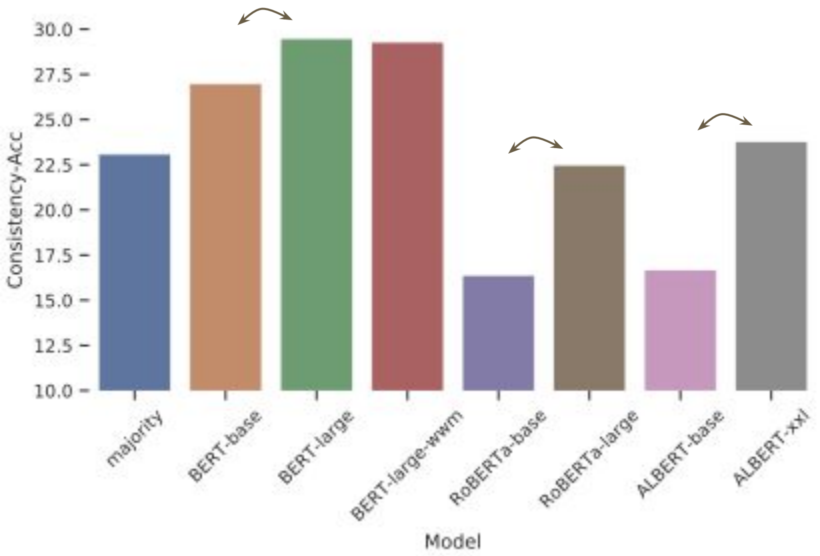


- Drill down
  (consistent-accurate results)

Consistency - Results

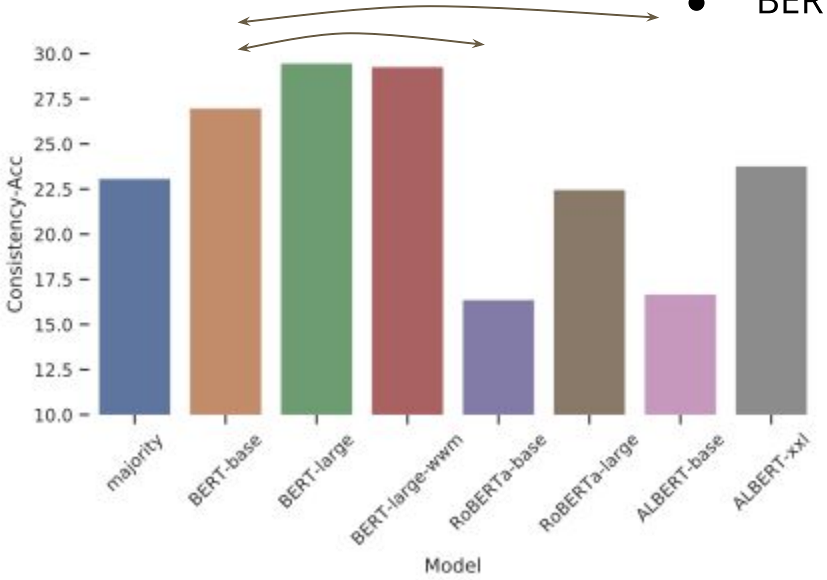
Interesting trends: *base vs. large*

- large > small



Consistency - Results

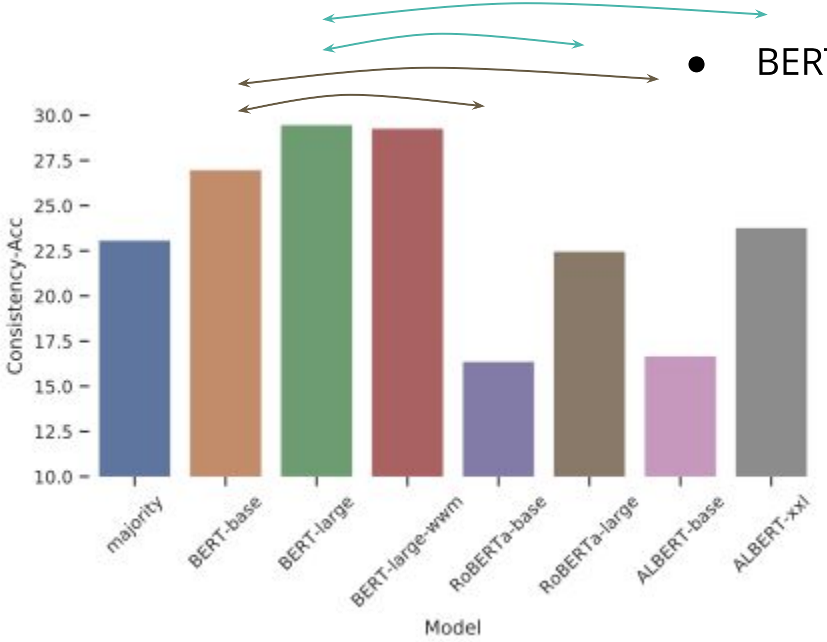
Interesting trends: BERT vs. others



- BERT > others

Consistency - Results

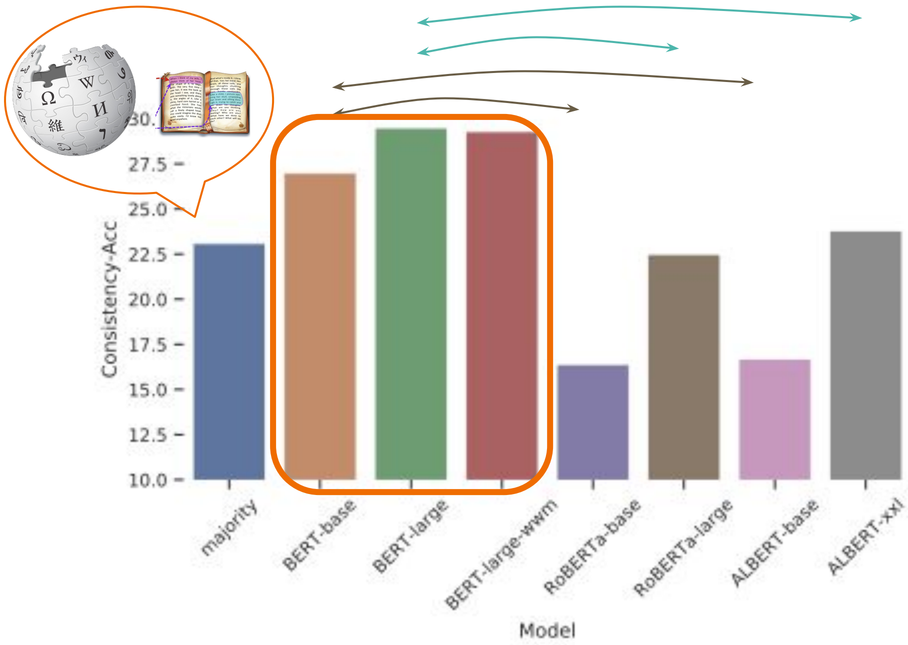
Interesting trends: BERT vs. others



- BERT > others

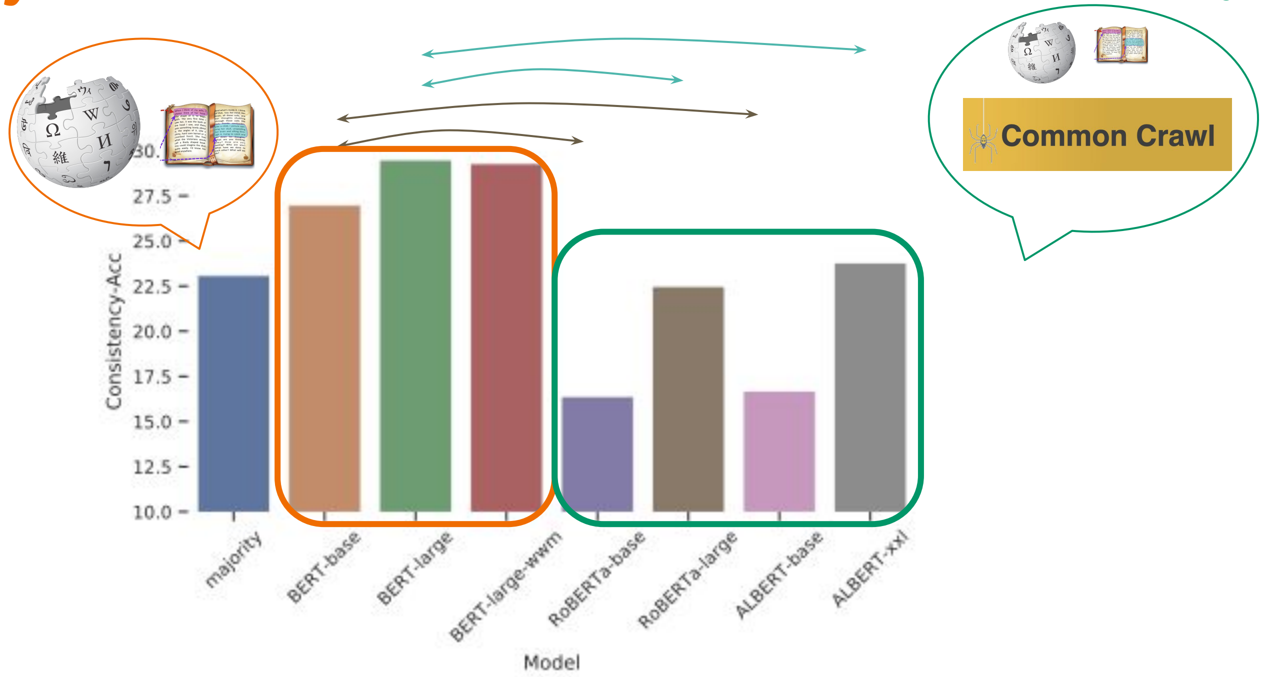
Consistency - Results

Interesting trends: BERT vs. others



Consistency - Results

Interesting trends: BERT vs. others



Consistency - Summary

We have shown that:

- Some relations are more consistent than others
- Some models are more consistent than others

But overall, **models are inconsistent!**

Much more analysis and experiments in the paper!!

Improving Consistency

Improving Consistency

More details in the paper!

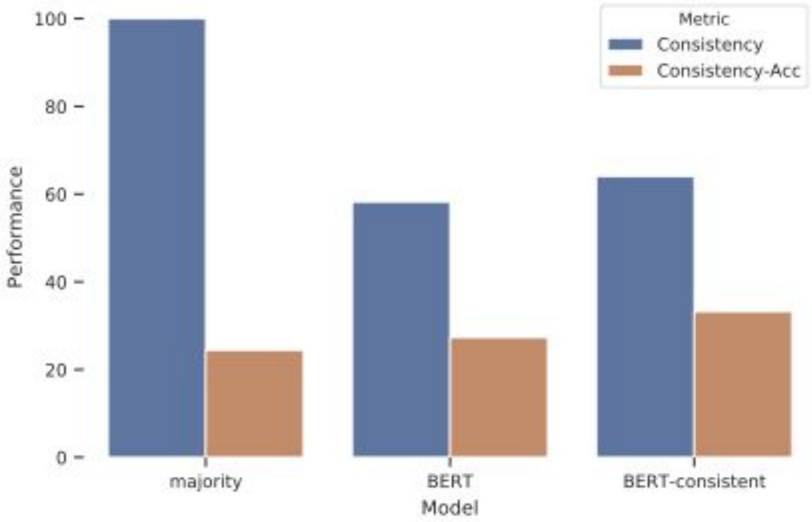
- Can we improve the consistency of PLMs?
- We want predictions from paraphrases to be equal
- We try to make the distributions alike

$$Q_n = softmax(f_\theta(P_n))$$

$$\mathcal{L}_c = \sum_{n=1}^{k} \sum_{m=n+1}^{k} D_{KL}(Q_n^{r_i} || Q_m^{r_i}) + D_{KL}(Q_m^{r_i} || Q_n^{r_i})$$

$$\mathcal{L} = \lambda \mathcal{L}_c + \mathcal{L}_{MLM}$$

Improved Consistency



Part III

Explaining The “Knowledge”

Work In Progress

Explaining Knowledge in PLMs

LM predictions

#1 mask:Barack Obama was born in **[MASK]**.

	bert_large_cased
0	Chicago
1	Philadelphia
2	Detroit
3	Houston
4	Atlanta
5	Georgia
6	Boston
7	Texas
8	Paris
9	Dallas

	roberta_large
0	Kenya
1	Hawaii
2	1961
3	1964
4	Chicago
5	Honolulu
6	Indonesia
7	1965
8	1969
9	1963

WHY???

Explaining Knowledge in PLMs

LM predictions

#1 mask:Barack Obama was born in **[MASK]**.

	bert_large_cased
0	Chicago
1	Philadelphia
2	Detroit
3	Houston
4	Atlanta
5	Georgia
6	Boston
7	Texas
8	Paris
9	Dallas

	roberta_large
0	Kenya
1	Hawaii
2	1961
3	1964
4	Chicago
5	Honolulu
6	Indonesia
7	1965
8	1969
9	1963

University of Chicago Law School and civil rights attorney

In 1991, Obama accepted a two-year position as Visiting Law and Government Fellow at the University of Chicago Law School to work on his first book.[120][121] He then taught constitutional law at the University of Chicago Law School for twelve years, first as a lecturer from 1992 to 1996, and then as a senior lecturer from 1996 to 2004.[122]

From April to October 1992, Obama directed Illinois's Project Vote, a voter registration campaign with ten staffers and seven hundred volunteer registrars; it achieved its goal of registering 150,000 of 400,000 unregistered African Americans in the state, leading *Crain's Chicago Business* to name Obama to its 1993 list of "40 under Forty" powers to be.[123]

He joined Davis, Miner, Barnhill & Galland, a 13-attorney law firm specializing in civil rights litigation and neighborhood economic development, where he was an associate for three years from 1993 to 1996, then of counsel from 1996 to 2004. In 1994, he was listed as one of the lawyers in *Buycks-Roberson v. Citibank Fed. Sav. Bank*, 94 C 4094 (N.D. Ill.).[124] This class action lawsuit was filed in 1994 with Selma Buycks-Roberson as lead plaintiff and alleged that Citibank Federal Savings Bank had engaged in practices forbidden under the Equal Credit Opportunity Act and the Fair Housing Act.[125] The case was settled out of court.[126] Final judgment was issued on May 13, 1998, with Citibank Federal Savings Bank agreeing to pay attorney fees.[127] His law license became inactive in 2007.[128][129]

From 1994 to 2002, Obama served on the boards of directors of the Woods Fund of Chicago—which in 1985 had been the first foundation to fund the Developing Communities Project—and of the Joyce Foundation.[58] He served on the board of directors of the Chicago Annenberg Challenge from 1995 to 2002, as founding president and chairman of the board of directors from 1995 to 1999.[58]

Explaining Knowledge in PLMs

Data as a source of explanations

Explaining Knowledge in PLMs

- Taking another look at the data: Wikipedia

LM predictions

#1 mask:Barack Obama was born in **[MASK]**.

	bert_large_cased
0	Chicago
1	Philadelphia
2	Detroit
3	Houston
4	Atlanta
5	Georgia
6	Boston
7	Texas
8	Paris
9	Dallas

Background

Early life of Barack Obama

Main articles: Early life and career of Barack Obama and Ann Dunham

People who express doubts about Obama's eligibility or reject details about his early life are often informally called "birthers", a term that parallels[23] the nickname "truthers" for adherents of 9/11 conspiracy theories.[24][25] These conspiracy theorists reject at least some of the following facts about his early life:

Barack Obama was born on August 4, 1961, at Kapi'olani Maternity & Gynecological Hospital (now called Kapi'olani Medical Center for Women & Children) in Honolulu, Hawaii,[26][27][28][29] to Ann Dunham,[30] from Wichita, Kansas,[31] and her husband Barack Obama Sr., a Luo from Nyang'oma Kogelo, Nyanza Province (in what was then the Colony and Protectorate of Kenya), who was attending the University of Hawaii. Birth notices for Barack Obama were published in *The Honolulu Advertiser* on August 13 and the *Honolulu Star-Bulletin* on August 14, 1961.[26][31] Obama's father's immigration file also clearly states Barack Obama was born in Hawaii.[32] One of his high school teachers, who was acquainted with his mother at the time, remembered hearing about the day of his birth.[30]

Explaining Knowledge in PLMs

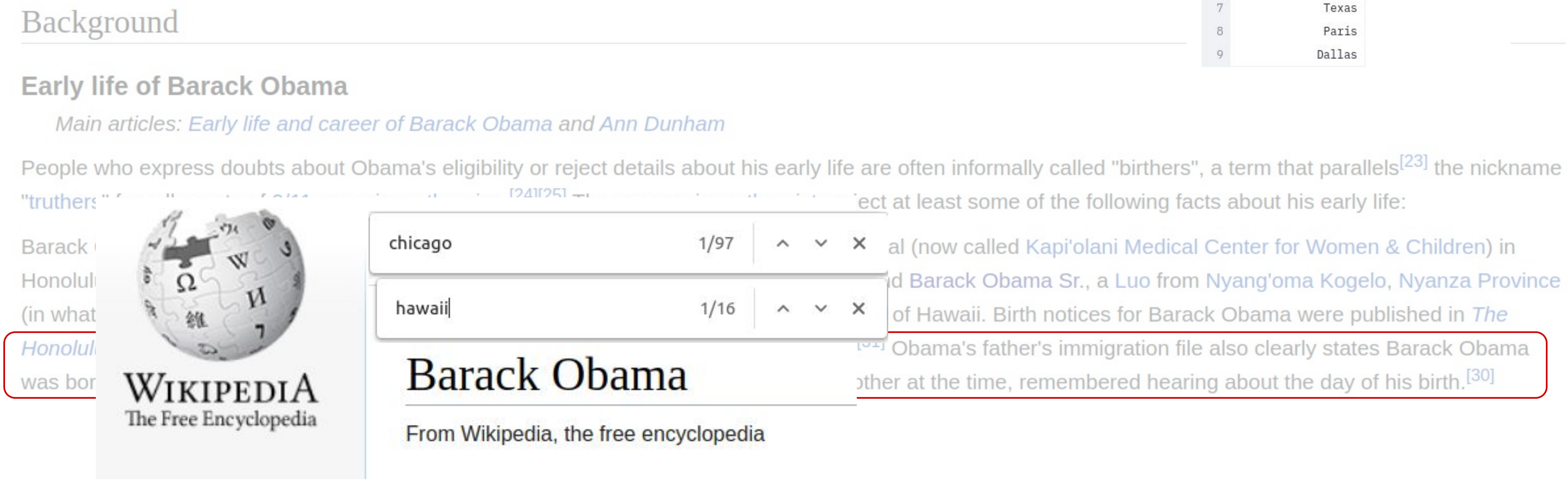
- Taking another look at the data: Wikipedia

LM predictions

#1 mask:Barack Obama was born in **[MASK]**.

	bert_large_cased
0	Chicago
1	Philadelphia
2	Detroit
3	Houston
4	Atlanta
5	Georgia
6	Boston
7	Texas
8	Paris
9	Dallas

WHY???



Explaining Knowledge in PLMs

- Maybe these models rely on co-occurrences?

What else?
(How to predict a word, given a cloze sentence such as:
“Barack Obama was born in [MASK].”)

Pitfalls of LMs as KBs

Explaining Knowledge in PLMs

- We inspect 3 pitfalls (or heuristics) in LMs with respect to knowledge extraction

Explaining Knowledge in PLMs

- We inspect 3 pitfalls (or heuristics) in LMs with respect to knowledge extraction
  - Model ignores the subject, and only uses the pattern to make prediction

- Example:
  - ~~Barack Obama~~ was born in [MASK]. *(born-in)*

Explaining Knowledge in PLMs

- We inspect 3 pitfalls (or heuristics) in LMs with respect to knowledge extraction
    - Model ignores the subject, and only uses the pattern to make prediction
    - Model ignores the pattern, and only uses the subject to make prediction

- Example:
    - ~~Barack Obama~~ was born in [MASK]. *(born-in)*
    - Barack Obama ~~was born in~~ [MASK]. *(~~born-in~~)*

Explaining Knowledge in PLMs

- We inspect 3 pitfalls (or heuristics) in LMs with respect to knowledge extraction
  - Model ignores the subject, and only uses the pattern to make prediction
  - Model ignores the pattern, and only uses the subject to make prediction
  - Model ignores the abstract relation, and uses the subject+pattern to make prediction

- Example:
  - ~~Barack Obama~~ was born in [MASK]. *(born-in)*
  - Barack Obama ~~was born in~~ [MASK]. *(~~born-in~~)*
  - Barack Obama was born in [MASK]. *(~~born-in~~)*

Explaining Knowledge in PLMs

- We inspect 3 pitfalls (or heuristics) in LMs with respect to knowledge extraction
  - Model ignores the subject, and only uses the pattern to make prediction
  - Model ignores the pattern, and only uses the subject to make prediction
  - Model ignores the abstract relation, and uses the subject+pattern to make prediction

- Example:
  - ~~Barack Obama~~ was born in [MASK]. *(born-in)* ← *Pattern's preference*
  - Barack Obama ~~was born in~~ [MASK]. *(~~born-in~~)* ← *Subj-obj cooccurrences*
  - Barack Obama was born in [MASK]. *(~~born-in~~)* ← *Memorization*

Explaining Knowledge in PLMs

- Example:
    - ~~Barack Obama~~ was born in [MASK]. *(born-in)*
    - Barack Obama ~~was born in~~ [MASK]. *(~~born-in~~)*
    - Barack Obama was born in [MASK]. *(~~born-in~~)*

- Tests using the model:
    - Default Behavior
        - was born in [MASK]
    - Entities association
        - Barack Obama died in [MASK]
    - Consistency
        - Barack Obama, born in [MASK].

Explaining Knowledge in PLMs

- Example:
  - ~~Barack Obama~~ was born in [MASK]. *(born-in)*
  - Barack Obama ~~was born in~~ [MASK]. *(~~born-in~~)*
  - Barack Obama was born in [MASK]. *(~~born-in~~)*

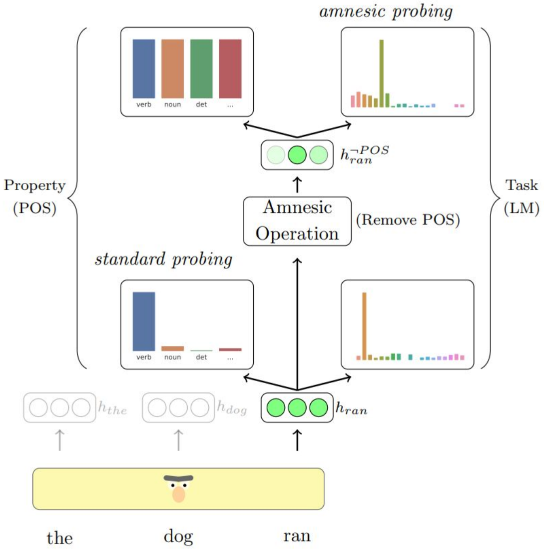
- Explaining the Model through the Data:
  - Occurrences of pattern+object (count in wiki: “was born in Hawaii”)
  - Entities Co-occurrence (count <“Barack Obama, Hawaii”>, <“Barack Obama, Chicago”>, …)
  - Memorization (count “Barack Obama was born in Hawaii”)

Explaining Knowledge in PLMs

- Entities Association:
    - Probability that BERT predicts the most co-occurred entity when the pattern describes a correct relation is 40%, compared to 35% when the relation doesn't hold
- Similar trends for the memorization
- But this is not a causal attribution!

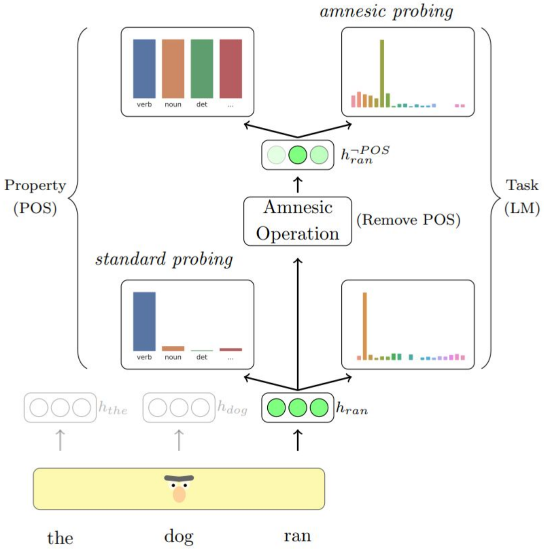
Causal Explanation through the Data

- Can't use *amnesic probing*
  - *Concepts aren't clear*



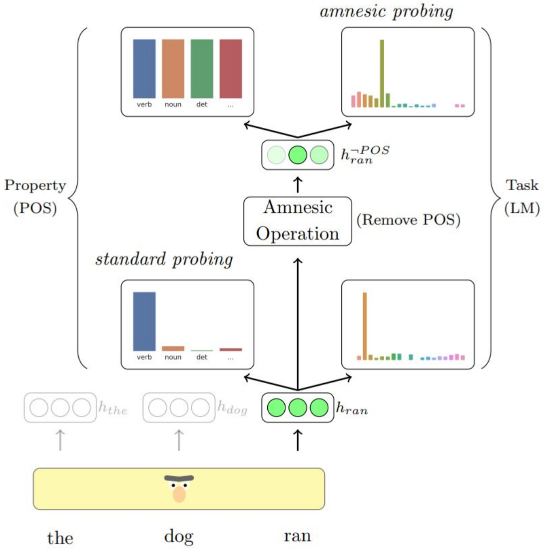
Causal Explanation through the Data

- Can't use *amnesic probing*
  - *Concepts aren't clear*
- Can't perform intervention on the data
  - Retraining BERT (on each combination) is expensive



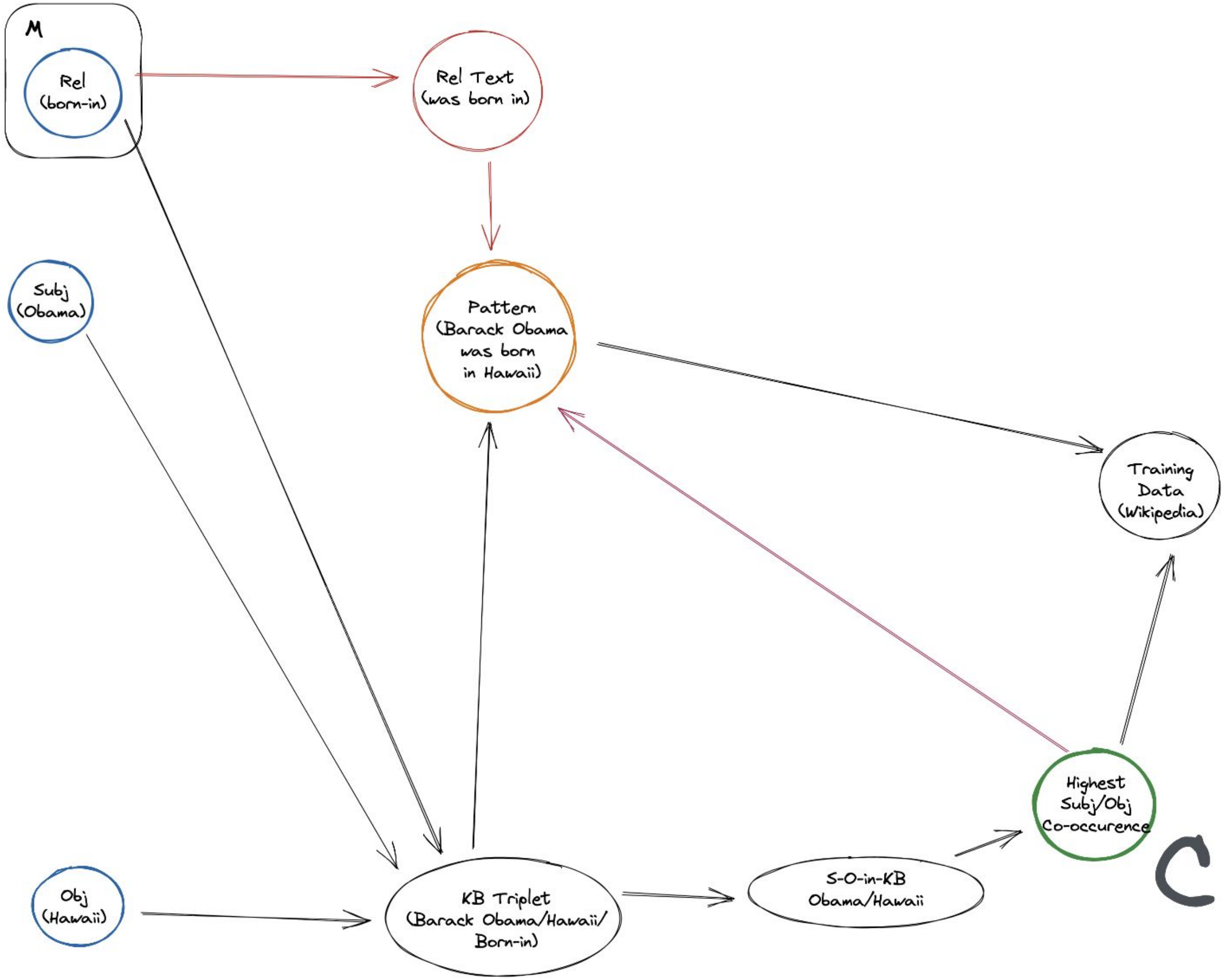
Causal Explanation through the Data

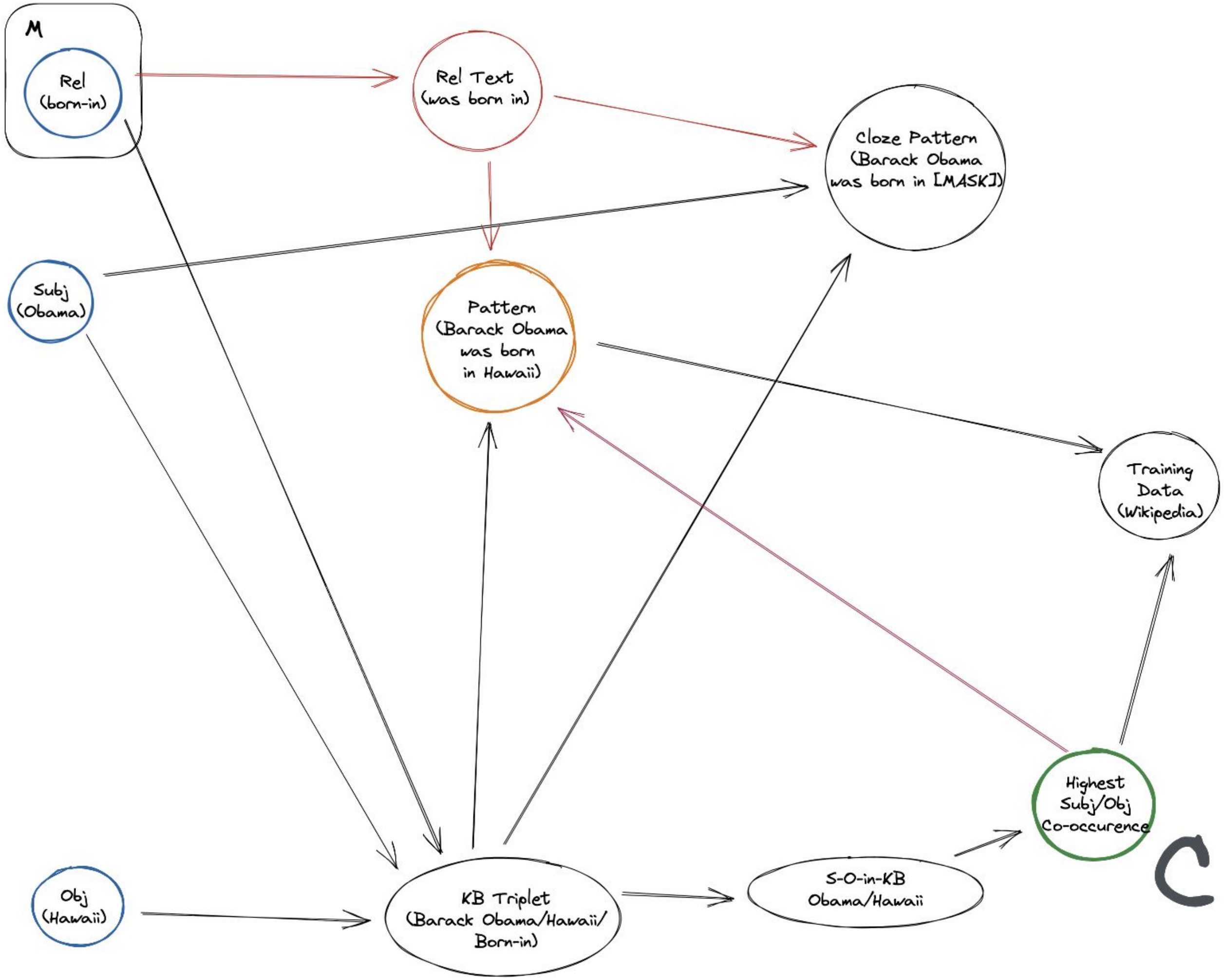
- Can't use *amnesic probing*
  - *Concepts aren't clear*
- Can't perform intervention on the data
  - Retraining BERT (on each combination) is expensive
- Solution: Measure ***Average Treatment Effect (ATE)*** using observational data
  - Assuming we can observe the measurable variables

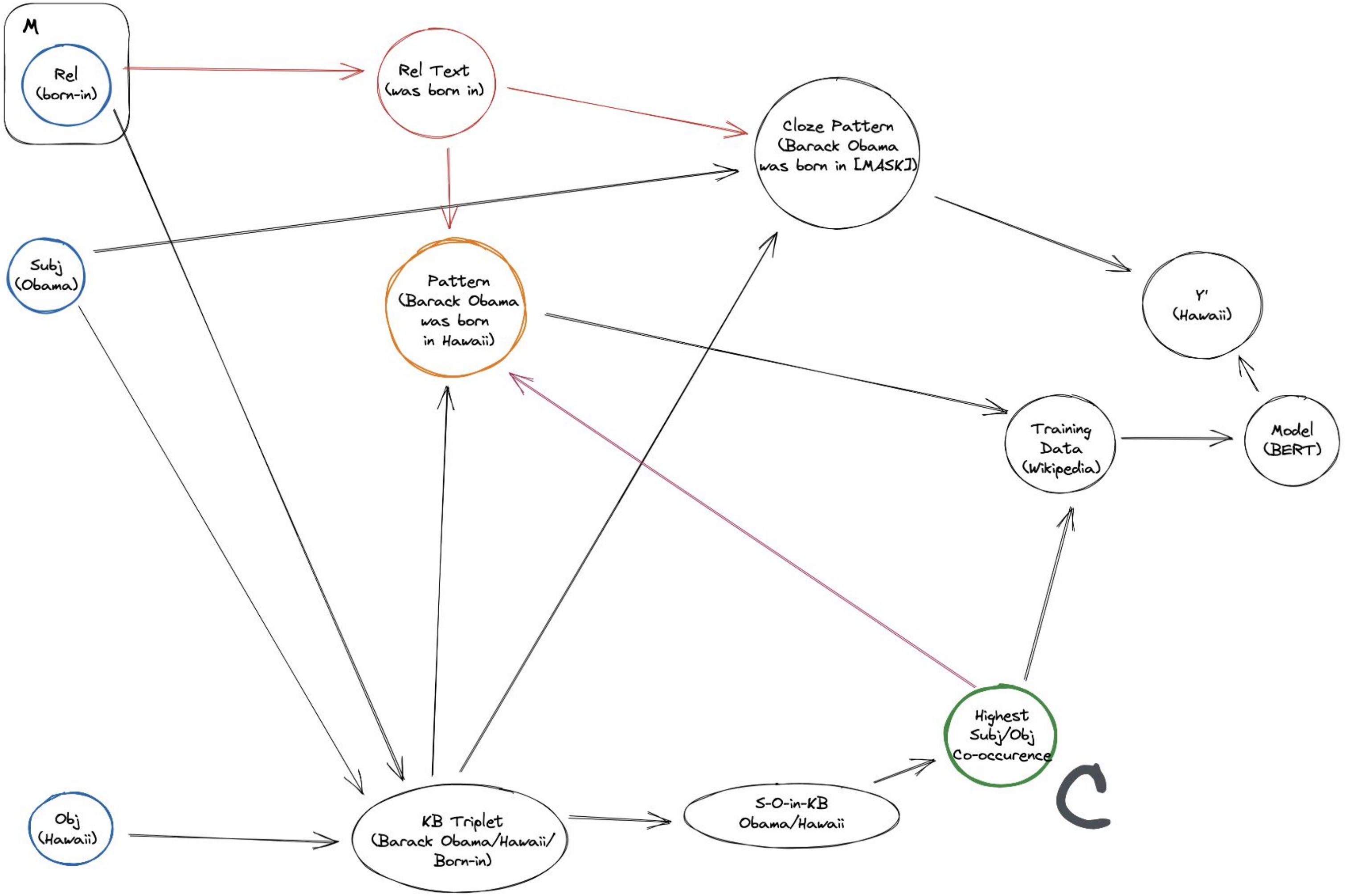


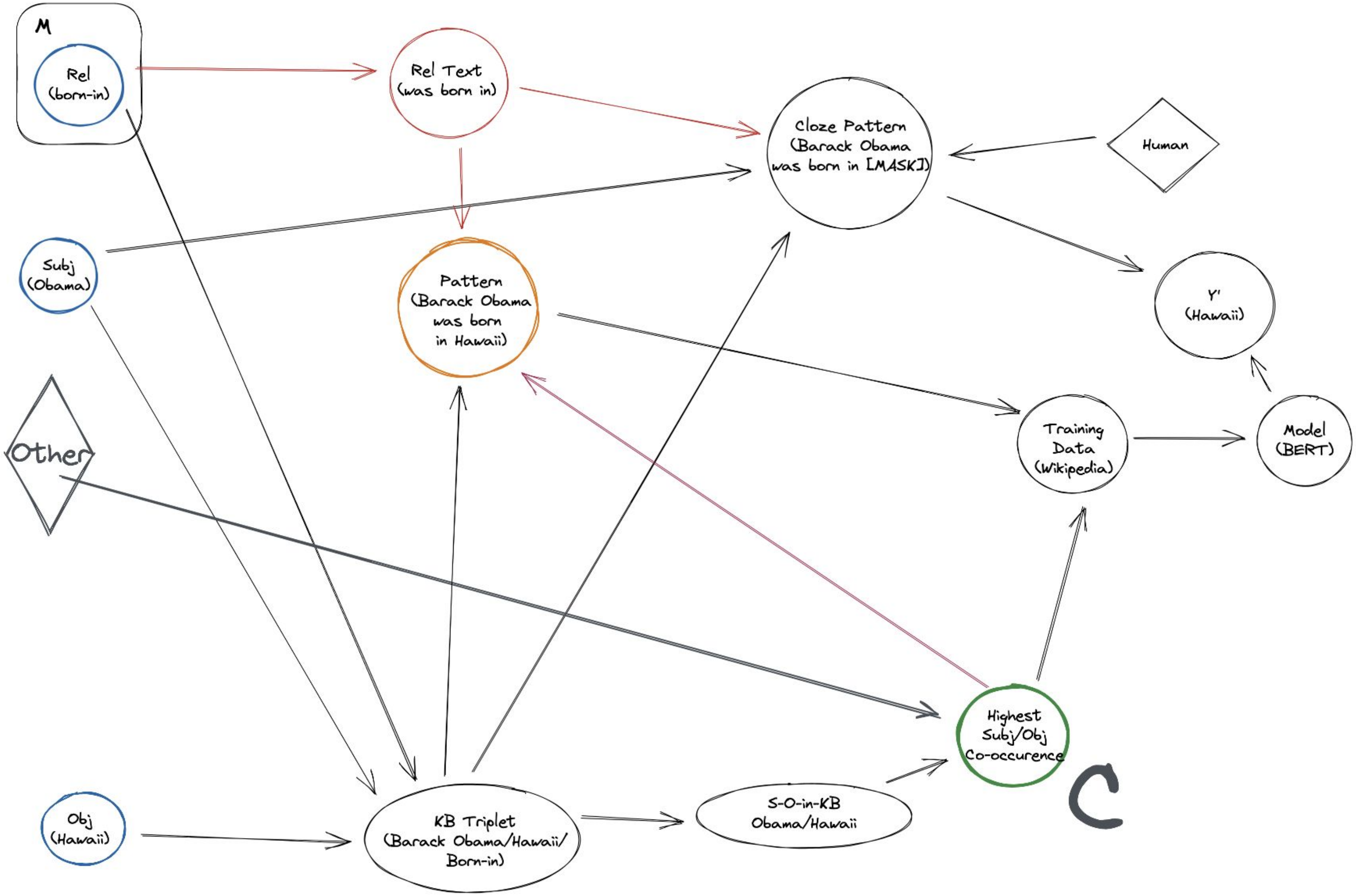
Causal Diagram

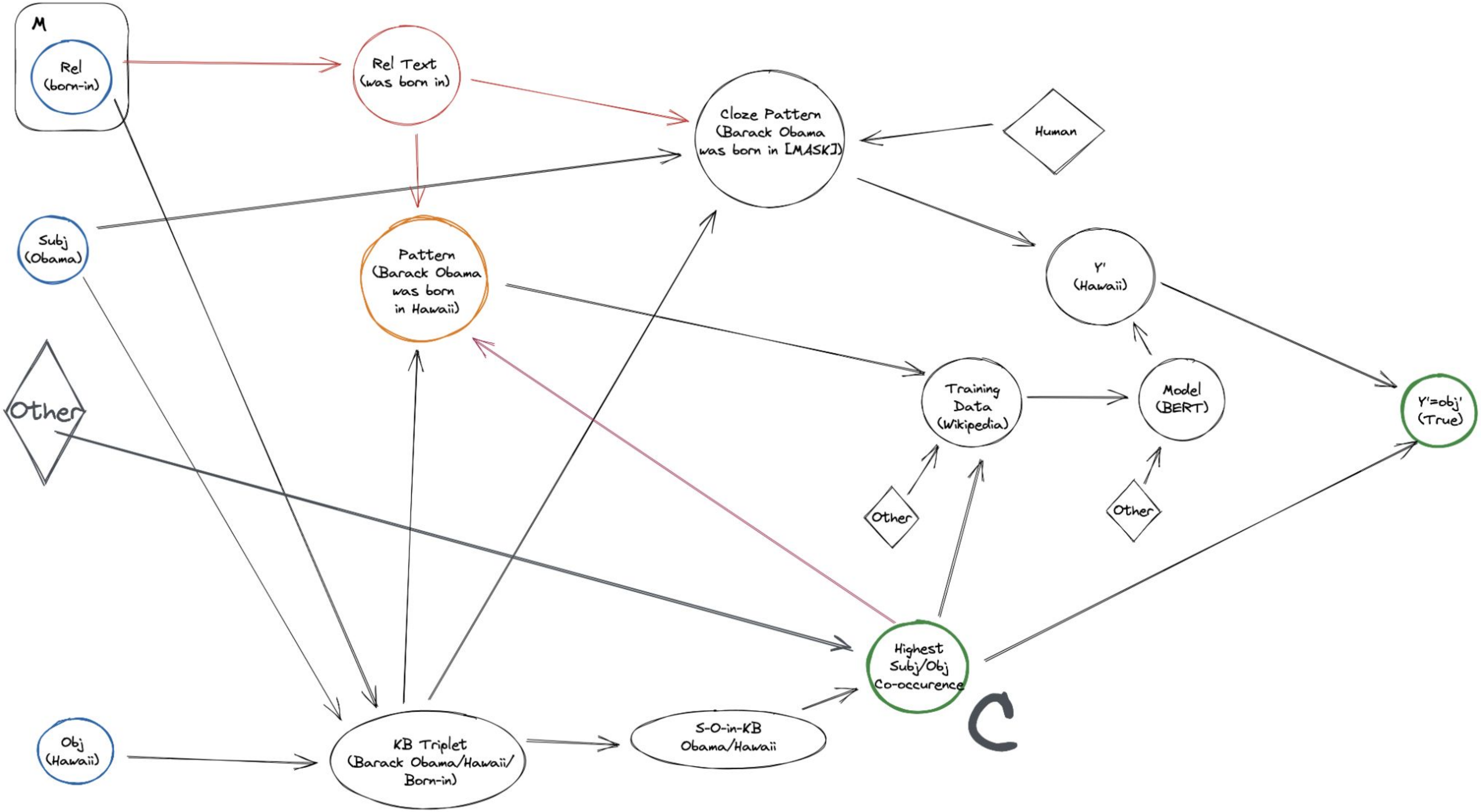

M
Rel
(born-in)
Subj
(Obama)
Obj
(Hawaii)


M
Rel
(born-in)
Rel Text
(was born in)
Subj
(Obama)
Pattern
(Barack Obama
was born
in Hawaii)
Training
Data
(Wikipedia)
Highest
Subj/Obj
Co-occurence
Obj
(Hawaii)
KB Triplet
(Barack Obama/Hawaii/
Born-in)
S-O-in-KB
Obama/Hawaii


M
Rel
(born-in)
Rel Text
(was born in)
Cloze Pattern
(Barack Obama
was born in [MASK])
Subj
(Obama)
Pattern
(Barack Obama
was born
in Hawaii)
Training
Data
(Wikipedia)
Highest
Subj/Obj
Co-occurence
Obj
(Hawaii)
KB Triplet
(Barack Obama/Hawaii/
Born-in)
S-O-in-KB
Obama/Hawaii


M
Rel
(born-in)
Rel Text
(was born in)
Cloze Pattern
(Barack Obama
was born in [MASK])
Subj
(Obama)
Pattern
(Barack Obama
was born
in Hawaii)
Y'
(Hawaii)
Training
Data
(Wikipedia)
Model
(BERT)
Highest
Subj/Obj
Co-occurence
C
Obj
(Hawaii)
KB Triplet
(Barack Obama/Hawaii/
Born-in)
S-O-in-KB
Obama/Hawaii


M
Rel
(born-in)
Rel Text
(was born in)
Cloze Pattern
(Barack Obama
was born in [MASK])
Human
Subj
(Obama)
Pattern
(Barack Obama
was born
in Hawaii)
Y'
(Hawaii)
Other
Training
Data
(Wikipedia)
Model
(BERT)
Highest
Subj/Obj
Co-occurence
C
Obj
(Hawaii)
KB Triplet
(Barack Obama/Hawaii/
Born-in)
S-O-in-KB
Obama/Hawaii


M
Rel
(born-in)
Rel Text
(was born in)
Cloze Pattern
(Barack Obama
was born in [MASK])
Human
Subj
(Obama)
Pattern
(Barack Obama
was born
in Hawaii)
Y'
(Hawaii)
Other
Training
Data
(Wikipedia)
Model
(BERT)
Y'=obj'
(True)
Other
Other
Highest
Subj/Obj
Co-occurence
C
Obj
(Hawaii)
KB Triplet
(Barack Obama/Hawaii/
Born-in)
S-O-in-KB
Obama/Hawaii

Explaining Knowledge - Causal Explanation

- Given that we believe this graph accurately describes the world...
- ... and we find the relevant back/front door criterion to control for confounding variables
- We can measure the effect of the heuristics on the models' predictions

NEAT! AND A STRONG RESULT

Explaining Knowledge - Causal Explanation

- If the effect is strong, what does it tell us about this model?
  - The model memorize, and uses correlations for making predictions
  - It has a limited understanding of linguistic relations
  - More?

Results

- Example:
  - ~~Barack Obama~~ was born in [MASK]. *(born-in)* ← *Pattern's preference*
  - Barack Obama ~~was born in~~ [MASK]. *(~~born-in~~)* ← *Subj-obj cooccurrences*
  - Barack Obama was born in [MASK]. *(~~born-in~~)* ← *Memorization*

Hypothesis	ATE
Pattern's Preference	4.1
Subj-Obj Co-occurrence	19.0
Memorization	10.4

Data as a Source of Explanations

Summary

- **Amnesic Probing**: a method that answers a causal question: “what is being used?”
- **Consistency** of PLMs knowledge is limited
- **Data as Explanation**: A graph describing causal relations
  - Allows to ask how concepts/heuristics **associated with training data** are used by models

Thanks!
Questions?

Yanai Elazar

@yanaiela

yanaiela.github.io