What does the Knowledge Neuron Thesis Have to do with Knowledge?

Jingcheng Niu[14], Andrew Liu[2], Zining Zhu[134], Gerald Penn[14]
[1]University of Toronto, [2]University of Waterloo, [3]Stevens Institute of Technology, [4]Vector Institute

ICLR 2024
Spotlight Poster

CODE & DATA

KN Thesis

Geva et al. (2021): LMs operate like key-value memories. Key: textual patterns; Value: output vocabulary distribution.
Dai et al., 2022; Meng et al., 2022:

- Facts & knowledge are also stored like key-value pairs in Knowledge Neurons (KNs) in MLPs. Key: prompt templates; Value: knowledge.
- We can control & edit LMs by modifying MLP weights or activations.

So, what does the KN thesis have to do with knowledge?

The KN Thesis is an **oversimplification**. The KN thesis does not adequately explain the process of factual expression. MLP weights store **complex patterns** that are interpretable both syntactically and semantically; however, **these patterns do not constitute "knowledge."**

Model Editing Methods Based on the KN Thesis are not Robust

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**, ...

Not robust for **symmetric** relations.

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**.

Not robust for **synonym** replacements.

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** ...

Subj–Verb Agr.: only edits one subj–verb pair.

Formal *vs.* Functional Competence

Mahowald et al.'s (2024) distinction of linguistic competence:

- **Formal**: knowledge of linguistic rules and patterns.
  Data: BLiMP Paradigms.
  Frank likes ____ apples. that *vs.* those.
- **Functional**: understand and use language in the world.
  Data: PARAREL Relations.
  Canada's capital is ___. Ottawa *vs.* Vienna.

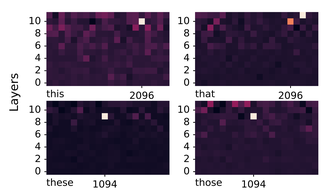
Informally referred to as **syntax** and **semantics**. *The Classical NLP Pipeline*.

- Can we localise syntactic phenomena using KN methods?
- How do the levels of localisation compare to each other?
- Are these results strong enough to support the KN thesis?

Conclusion: LMs may use the **same mechanisms** to process information related to these two types of competence.

Localising Syntactic Phenomena

Finding the determiner–noun agreement KNs.
BLiMP Paradigm: `determiner_noun_agreement_2`



Average KN attribution scores.

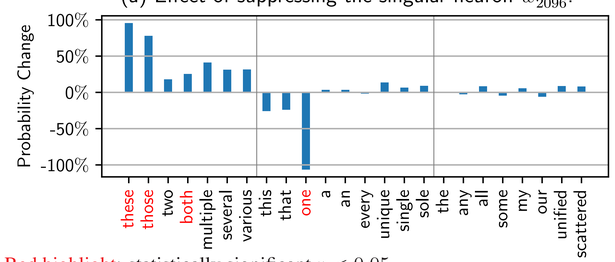
Neuron	this	that	these	those
$w^{(10)}_{2096}$	0.93	0.75	0	0
$w^{(9)}_{1094}$	0	0	1.00	1.00
$w^{(9)}_{2339}$	0.33	0	0.32	0
$w^{(11)}_{2686}$	0	0.81	0	0
...	...	...	...	...

Identified KNs for Det-N pairs.

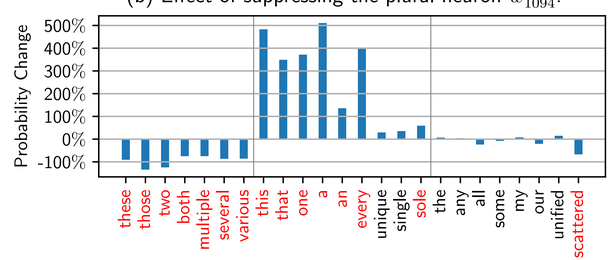
A common neuron ($w^{(10)}_{2096}$) activates for singular determiners (*this, that*) and another ($w^{(9)}_{1094}$) for plural determiners (*these, those*).
Are they the singular & plural neurons?

Effects of Suppressing the "Singular & Plural Neurons"

(a) Effect of suppressing the singular neuron $w^{(10)}_{2096}$.



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



Red highlight: statistically significant $p < 0.05$.

- The effect of KN suppression is pronounced.
- Suppress the singular KN: plural↑ singular↓ neutral−
- Suppress the plural KN: plural↓ singular↑ neutral−

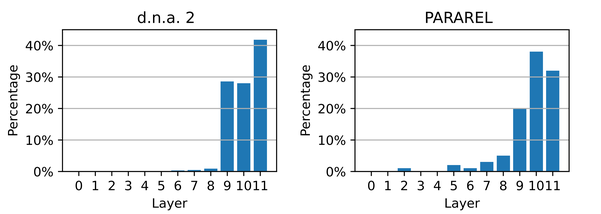
- Words like **scattered** also changed significantly, despite not expressing number agreement. E.g, *scattered rioting*.
- **They are not grammatical number KNs. The two KNs are affected by semantic number co-occurrence bias.**

Localisation Characteristics

BLiMP Paradigm	|KN|	τ	R^2_1
det_n_agr._1	3.94	0.71	0.56
det_n_agr._2	1.86	0.62	0.56
dna._irr._1	5.53	0.73	0.64
dna._irr._2	2.45	0.67	0.55
dna._w._adj_1	8.88	0.78	0.67
dna._w._adj_2	2.26	0.67	0.57

PARAREL Rels.	|KN|	τ	R^2_1
P101	0.167	0.515	0.399
P103	0.204	0.662	0.399
P106	1.292	0.607	0.365
P108	1.493	0.663	0.473
P1303	10.462	0.814	0.684
P140	2.008	0.689	0.263

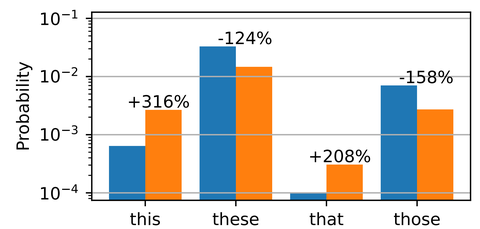
Same localisation statistics for syntactic & factual tasks.



Syntactic & factual KNs occupy the same range of layers.

KN Edit Effect

Editing the KNs is not enough to overturn the categorical predictions.



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%

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

Discussion & Conclusion

- Several syntactic agreement phenomena can be localised to MLP neurons.
- Syntactic phenomena has similar localisation characteristics to factual information.
- Formal and functional information may follow the same underlying mechanism.
- The effect of editing either cannot overturn the final prediction, or is limited to shallow cues such as token co-occurrence statistics.
- MLP neuron stores complex patterns that are interpretable linguistically, but **they are not knowledge**.
- We need a better framework: **circuits**?