Schema-learning and rebinding as mechanisms of in-context learning and emergence

NeurIPS 2023


Siva Swaminathan
sivark@


Antoine Dedieu
adedieu@


Rajkumar Vasudeva Raju
rajvraju@

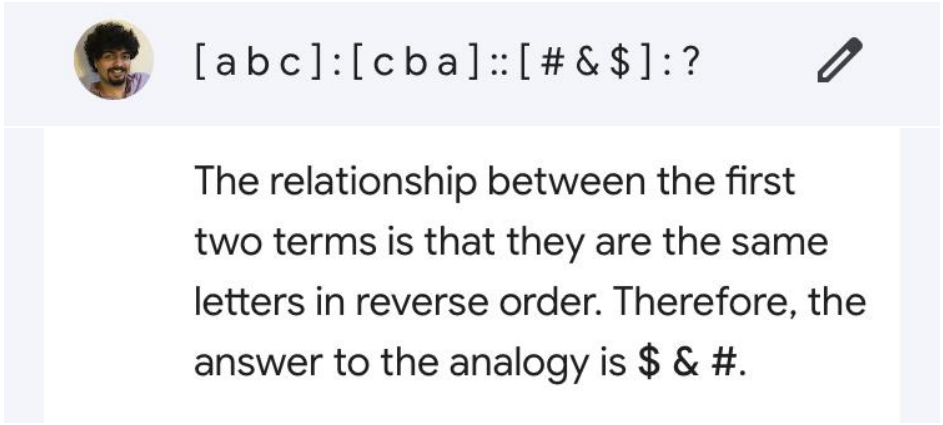

Murray Shanahan
mshanahan@


Miguel Lazaro-Gredilla
lazarogredilla@


Dileep George
dileepgeorge@

Google DeepMind

An example of in-context learning behavior



The model has likely never been trained on this *particular* sequence, but it manages to recall and use the *abstraction* of reversing a list.

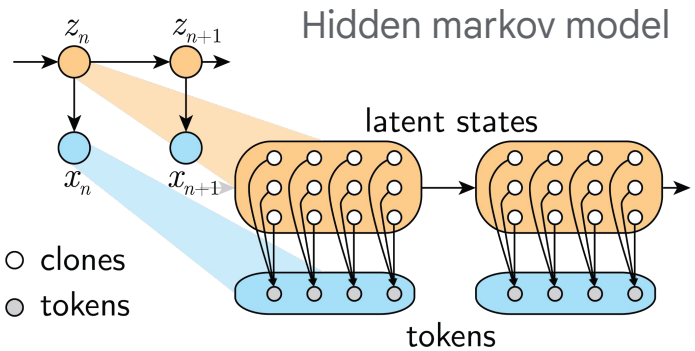
Sequence modeling with Clone-Structured Causal Graphs (CSCGs)

George et. al. Nature Communications 2021

"One bank robber eating milk and honey at river bank resort" ...

The model uses latent states to disambiguate different contexts for the same token, and then learns transitions between these latent states.

Multiple latent states a.k.a. "clones" bound to the same token

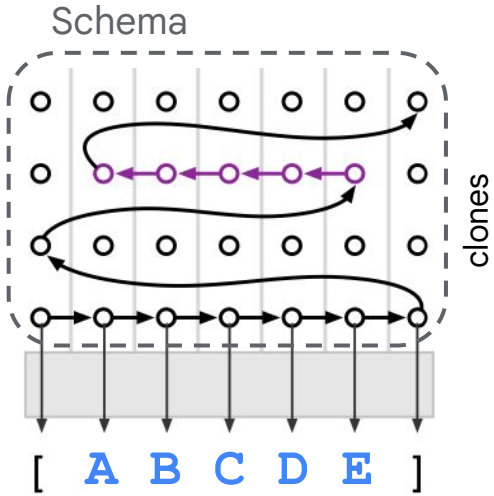


When trained on an example of list reversal

[A B C D E] [E D C B A]

How can the same model be applied to a novel prompt?

[P Q R S T] [T S R Q ?]



Preserve the pattern of flow between latent states as a "schema"

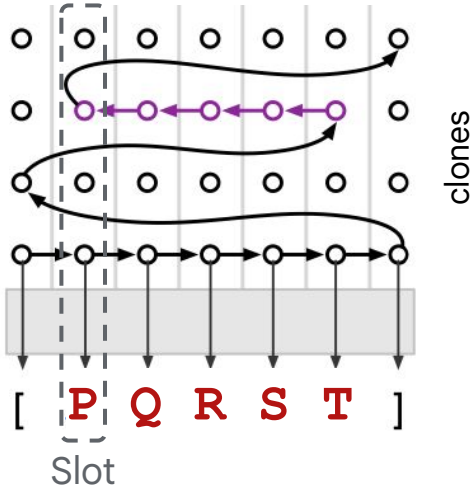
Rebinding enables a sequence model to represent abstractions

When trained on an example of list reversal

[A B C D E] [E D C B A]

How can the same model be applied to a novel prompt?

[P Q R S T] [T S R Q ?]



Preserve the pattern of flow between latent states as a "schema"

Rebind surprising prompt tokens to "slots" (clone groups) in the schema

$A \rightarrow P$, $B \rightarrow Q$, $C \rightarrow R$, $D \rightarrow S$, $E \rightarrow T$

What mechanism could select among competing abstractions?

Prompt `[ r t g o s k ] [ k s o g t r ] / [ x j d ] …`

Multiple training examples

`[ A B C D E F ] [ F E D C B A ] / …`	Reverse
`[ X Y R K M L ] [ Y R K M L X ] / …`	Bw. circ.
`[ J K L T Q P ] [ P J K L T Q ] / …`	Fw. circ.
`[ J K L T Q P ] J / …`	Query idx. 0

Peering under the hood of surprise-driven EM

Prompt [r t g o s k] [k s o g t r] / [x j d] …

Multiple training examples

[A B C D E F] [F E D C B A] / … Reverse

[X Y R K M L] [Y R K M L X] / … Bw. circ.

[J K L T Q P] [P J K L T Q] / … Fw. circ.

[J K L T Q P] J / … Query idx. 0

The process is bootstrapped using prediction surprise on the prompt

Peering under the hood of surprise-driven EM

Prompt [r t g o s k] [k s o g t r] / [x j d] …

Relevant training examples

[A B C D E F] [F E D C B A] / …	Reverse
[X Y R K M L] [Y R K M L X] / …	Bw. circ.
[J K L T Q P] [P J K L T Q] / …	Fw. circ.
~~[J K L T Q P] J / …~~	~~Query idx. 0~~

The process is bootstrapped using prediction surprise on the prompt

1. Unsurprising tokens act as “anchors” – serving as a template for partial selection of schemas

Peering under the hood of surprise-driven EM

Prompt `[ r t g o s k ] [ k s o g t r ] / [ x j d ] …`

Relevant training examples

`[ r t g o s k ] [ k s o g t r ] / …`	Reverse
`[ r t g o s k ] [ t g o s k r ] / …`	Bw. circ.
`[ r t g o s k ] [ k r t g o s ] / …`	Fw. circ.
~~`[ J K L T Q P ] J / …`~~	~~Query idx. 0~~

The process is bootstrapped using prediction surprise on the prompt

1. Unsurprising tokens act as “anchors” – serving as a template for partial selection of schemas
2. Surprising tokens bind to the “slots” in each remaining schema

Once the slots rebind to the surprising tokens in the prompt, consistency determines the correct abstraction.

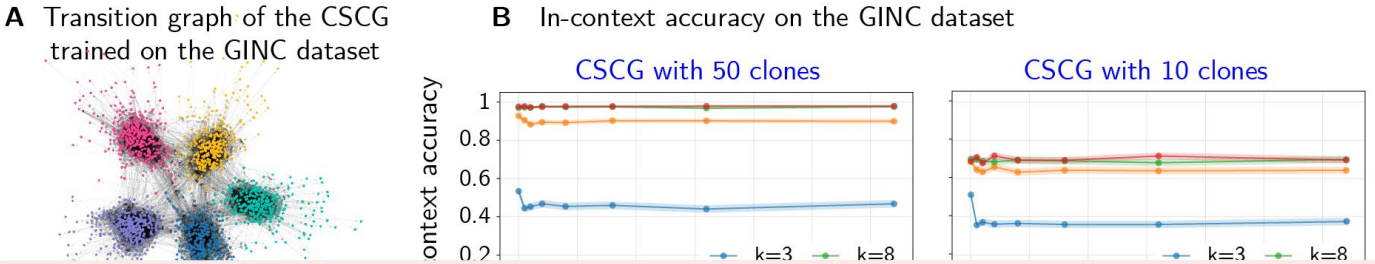
Mechanism for in-context learning

We propose that in-context learning is a combination of

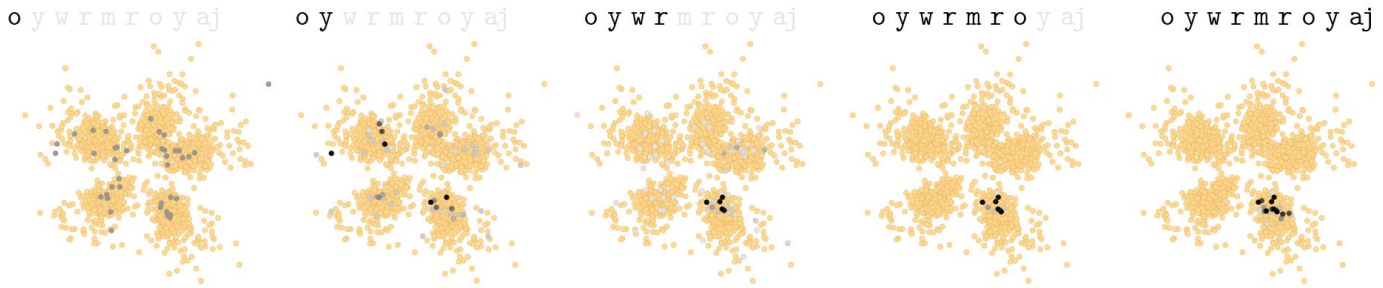
1. Learning schemas (template circuits) during training.
2. Retrieving relevant schemas in a context-sensitive manner.
3. Rebinding surprising prompt tokens to appropriate slots.

} In tandem, prompt-driven

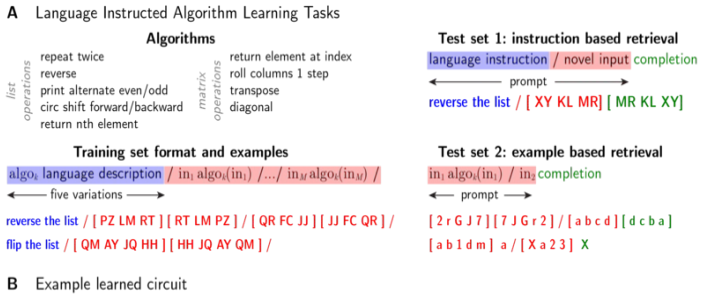
GINC dataset from *"An Explanation of In-context Learning as Implicit Bayesian Inference"* ICLR 2022

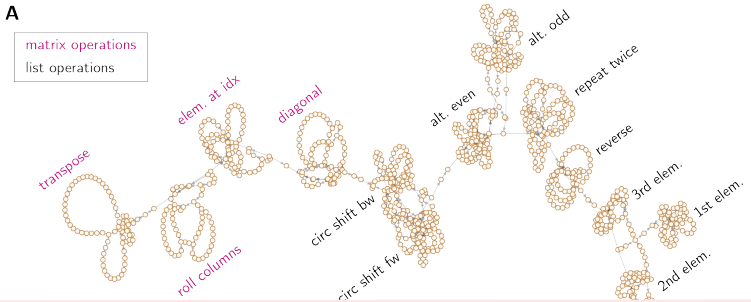


Subsumes prior work on Bayesian in-context learning as a special case which doesn't require rebinding and cannot handle novel tokens in the prompt

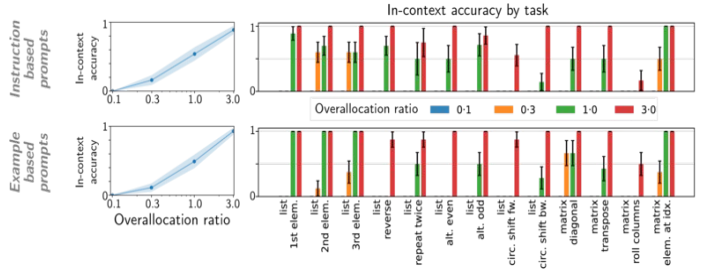


LIALT dataset for algorithm-learning tasks





A synthetic dataset to probe “algorithmic” generalization to novel prompts
where this mechanism results in high accuracy,
and the in-context learning process can be easily inspected



Schema-learning and rebinding as mechanisms of in-context learning and emergence

NeurIPS 2023


Siva Swaminathan
sivark@


Antoine Dedieu
adedieu@


Rajkumar Vasudeva Raju
rajvraju@


Murray Shanahan
mshanahan@


Miguel Lazaro-Gredilla
lazarogredilla@


Dileep George
dileepgeorge@

Google DeepMind