Evaluating neural language models

Roger Levy

9.19: Computational Psycholinguistics

8 November 2021

Agenda for today

- Impressionistic assessment of the text NLMs generate
- Perplexity-based evaluation
- Targeted grammatical evaluation: subject–verb agreement
    - Left-to-right prediction paradigm
    - Psycholinguistics of subject–verb agreement
    - Evaluation on "colorless green" (*nonce*) sentences
    - Controlled stimuli
    - Ablation tests to reveal circuit-level processing in models

What do language models generate?

- Half of these are generated from *n*-gram models, half from an LSTM:

What do language models generate?

- Half of these are generated from *n*-gram models, half from an LSTM:

Last December through the way to preserve the Hudson corporation N. B. E. C. Taylor would seem to complete the major central planners one point five percent of U. S. E. has already old M. X. corporation of living on information such as more frequently fishing to keep her

Another person familiar with the diaries pointed her tonight

The state is insisting extensive sanctions against wooden emigration until I'd publish Mr. Collor 's own birth conservation control with him

Months the my and issue of year foreign new exchange 's September were recession exchange new endorsed a acquire to six executives

They also point to ninety nine point six billion dollars from two hundred four oh six three percent of the rates of interest stores as Mexico and Brazil on market conditions

It has always played something so an man who asked thanks to David Smith 's case would open its Hindu jail , Brooke Wright , the president and manager of the American Express Foundation in U.S. Providence, Va.

What do language models generate?

- Half of these are generated from *n*-gram models, half from an LSTM:

2-gram	Last December through the way to preserve the Hudson corporation N. B. E. C. Taylor would seem to complete the major central planners one point five percent of U. S. E. has already old M. X. corporation of living on information such as more frequently fishing to keep her
LSTM	Another person familiar with the diaries pointed her tonight
LSTM	The state is insisting extensive sanctions against wooden emigration until I'd publish Mr. Collor 's own birth conservation control with him
1-gram	Months the my and issue of year foreign new exchange 's September were recession exchange new endorsed a acquire to six executives
3-gram	They also point to ninety nine point six billion dollars from two hundred four oh six three percent of the rates of interest stores as Mexico and Brazil on market conditions
LSTM	It has always played something so an man who asked thanks to David Smith 's case would open its Hindu jail , Brooke Wright , the president and manager of the American Express Foundation in U.S. Providence, Va.

Review: perplexity-based LM evaluation

Perplexity: inverse (geometric) mean word probability

$$\textbf{PPL}(w_{1\ldots N}) = \sqrt[-N]{\prod_{i=1}^{N} P(w_i \mid w_{1\ldots i-1})}$$

Review: perplexity-based LM evaluation

Perplexity: inverse (geometric) mean word probability

$$\textbf{PPL}(w_{1\ldots N}) = \sqrt[-N]{\prod_{i=1}^{N} P(w_i \mid w_{1\ldots i-1})}$$

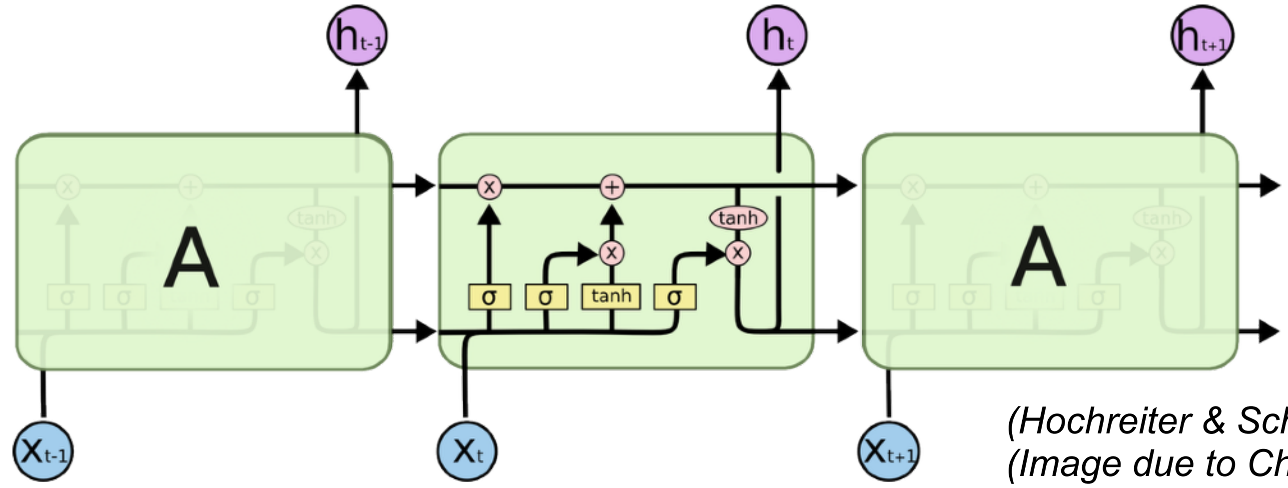
equivalently,

$$\textbf{PPL}(w_{1\ldots N}) = \exp\left[\frac{1}{N}\sum_{i=1}^{N} \log \frac{1}{P(w_i \mid w_{1\ldots i-1})}\right]$$

$$= 2^{-\frac{1}{N}\sum_{i=1}^{N} \log_2 P(w_i|w_{1\ldots i-1})}$$

(Recommended: prove the equivalence!)

Deep learning has revolutionized language modeling



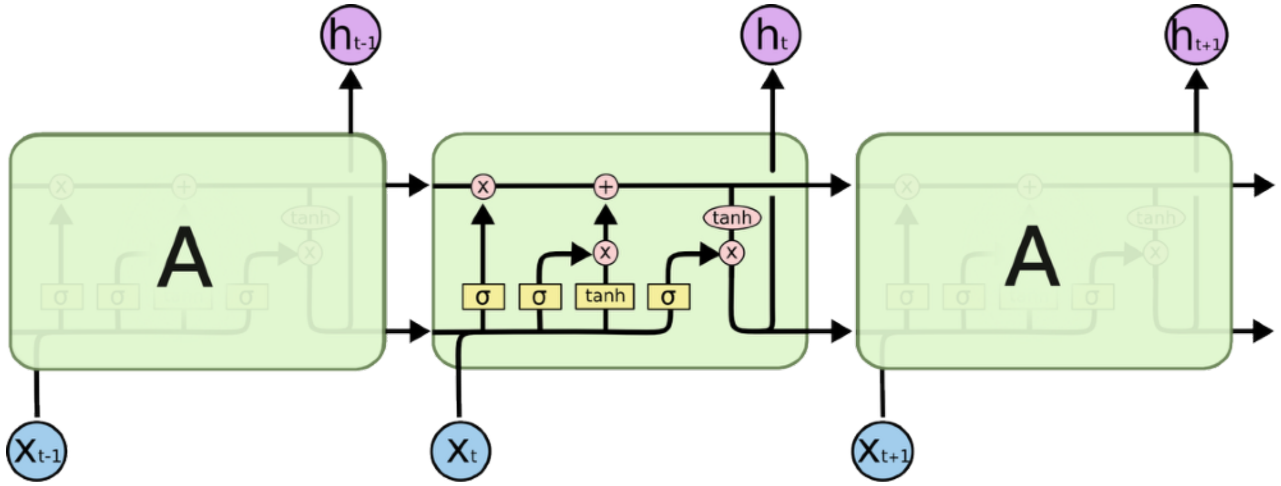
(Hochreiter & Schmidhuber, 1997)
(Image due to Christopher Olah)

Evaluation on "Billion word benchmark" train/dev/test:

MODEL	TEST PERPLEXITY
SIGMOID-RNN-2048 (JI ET AL., 2015A)	68.3
INTERPOLATED KN 5-GRAM, 1.1B N-GRAMS (CHELBA ET AL., 2013)	67.6
SPARSE NON-NEGATIVE MATRIX LM (SHAZEER ET AL., 2015)	52.9
RNN-1024 + MAXENT 9-GRAM FEATURES (CHELBA ET AL., 2013)	51.3
LSTM-512-512	54.1
LSTM-1024-512	48.2
LSTM-2048-512	43.7
LSTM-8192-2048 (NO DROPOUT)	37.9
LSTM-8192-2048 (50% DROPOUT)	32.2
2-LAYER LSTM-8192-1024 (BIG LSTM)	30.6
BIG LSTM+CNN INPUTS *(Jozefowicz et al., 2016)*	**30.0**

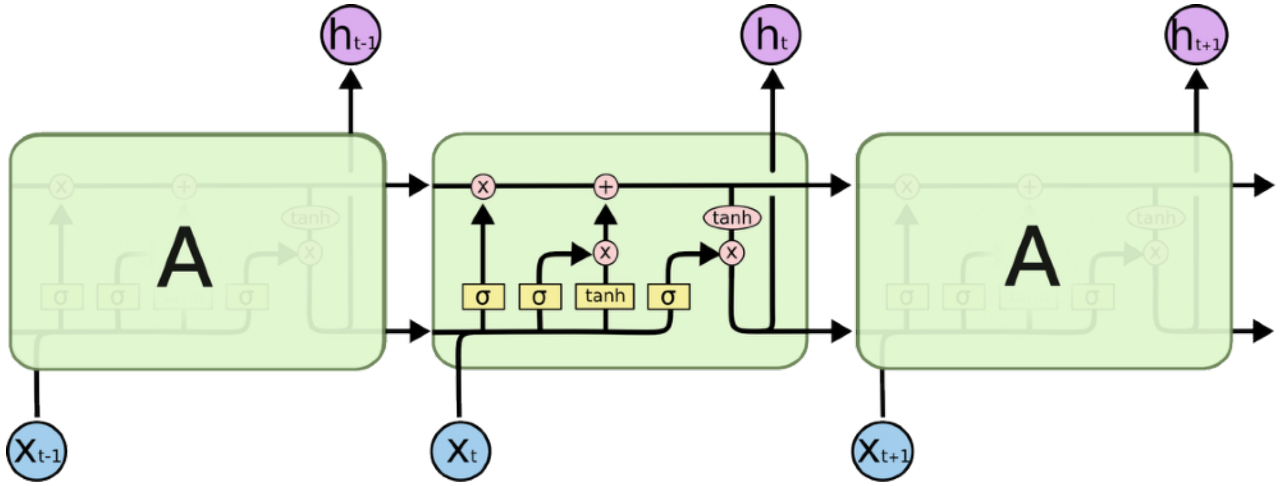
(More recent models do even better than this! Hopefully we will have a chance to look at them in a subsequent week)

What have these models learned as “English”?



(continuations generated by model of Jozefowicz et al., 2016)

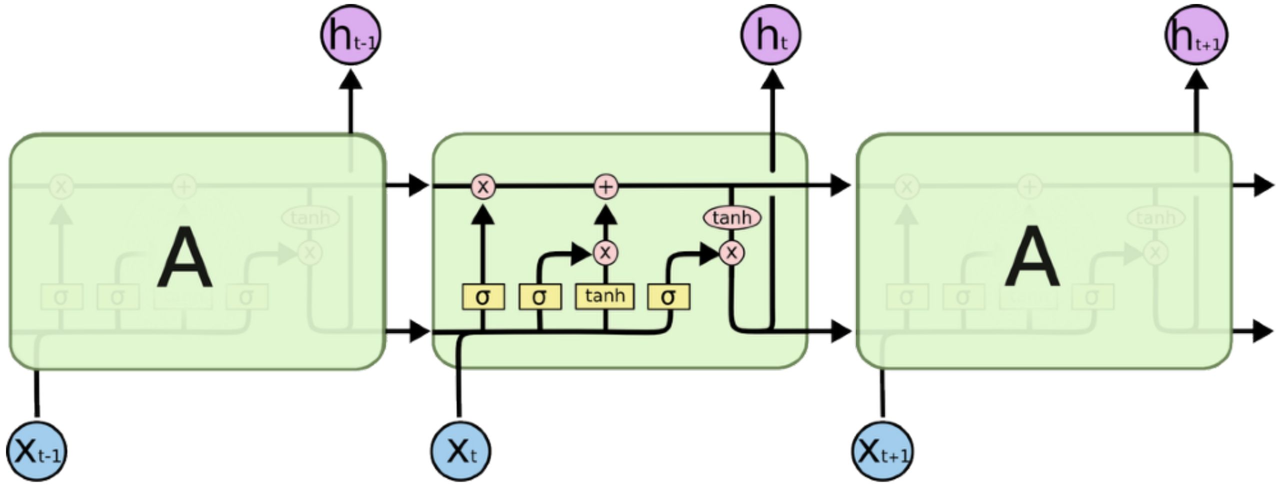
What have these models learned as “English”?



The girl who the newspaper…

(continuations generated by model of Jozefowicz et al., 2016)

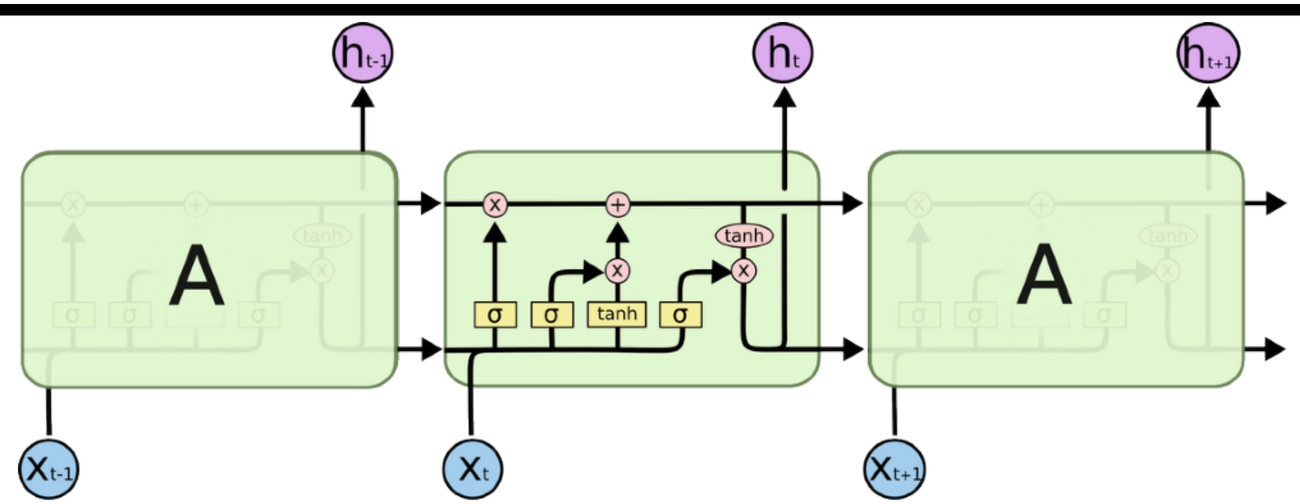
What have these models learned as “English”?



The girl who the newspaper now calls his girlfriend has really been hateful .

(continuations generated by model of Jozefowicz et al., 2016)

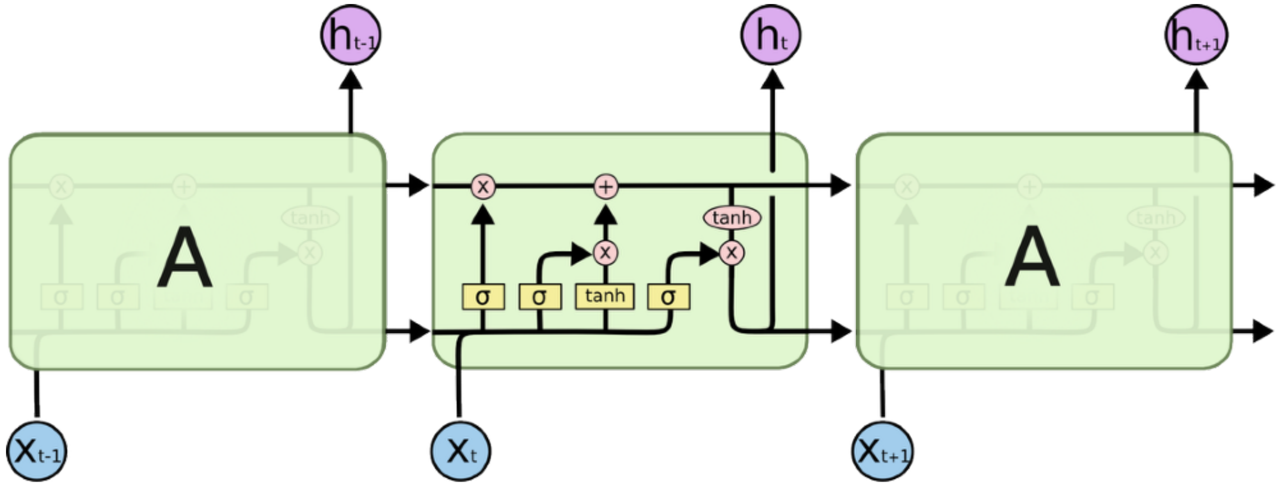
What have these models learned as “English”?



The girl who the newspaper now calls his girlfriend has really been hateful . ✓

(continuations generated by model of Jozefowicz et al., 2016)

What have these models learned as “English”?

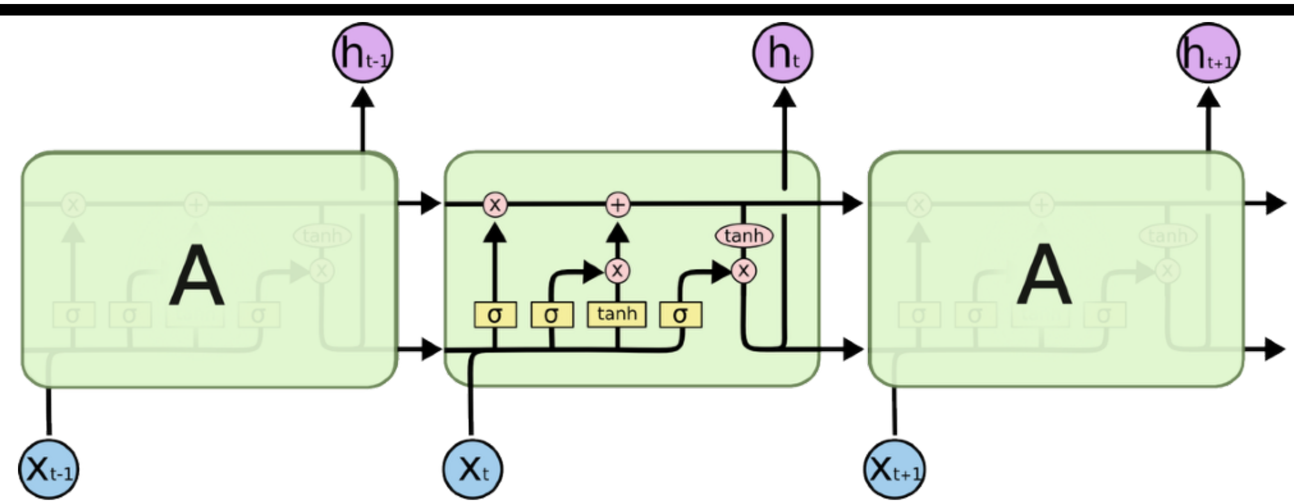


The girl who the newspaper now calls his girlfriend has really been hateful . ✓

The monologue that the actor who the movie industry…

(continuations generated by model of Jozefowicz et al., 2016)

What have these models learned as “English”?

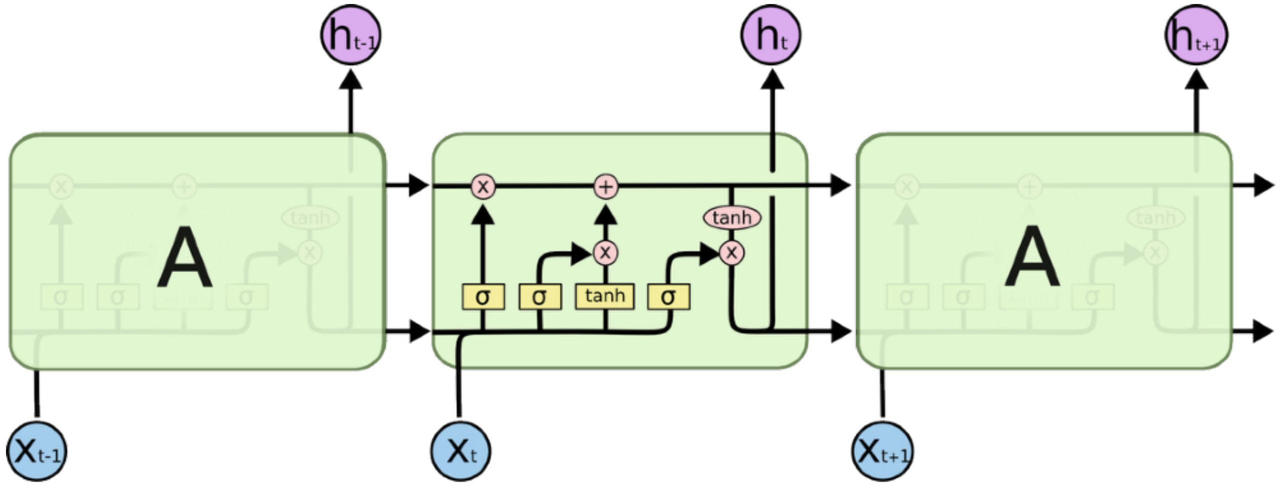


The girl who the newspaper now calls his girlfriend has really been hateful . ✓

The monologue that the actor who the movie industry likes made silent was being uploaded .

(continuations generated by model of Jozefowicz et al., 2016)

What have these models learned as "English"?

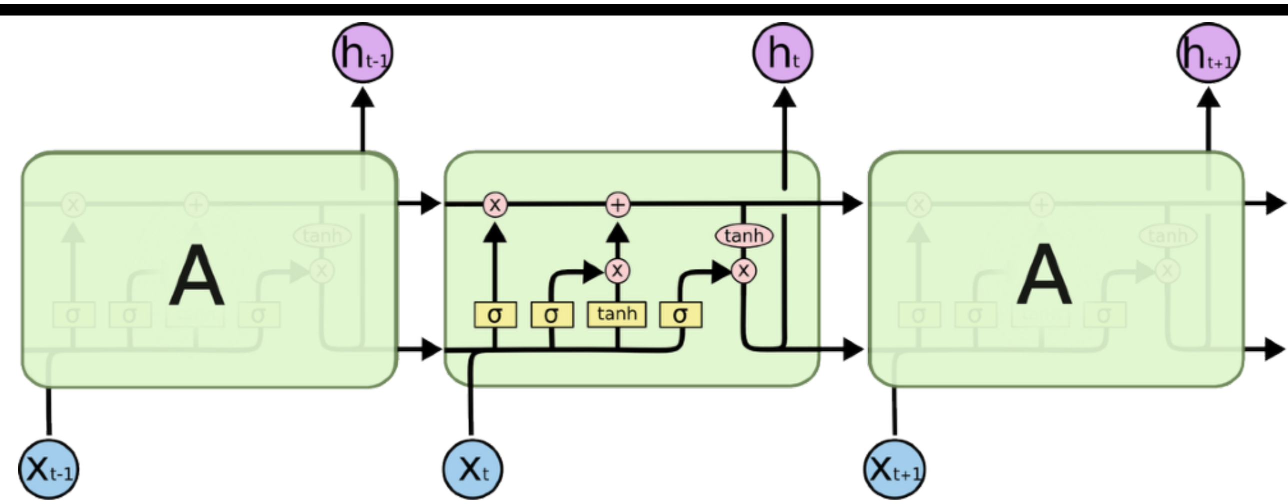


The girl who the newspaper now calls his girlfriend has really been hateful . ✓

The monologue that the actor who the movie industry likes made silent was being uploaded . ✓ (!)

(continuations generated by model of Jozefowicz et al., 2016)

What have these models learned as “English”?



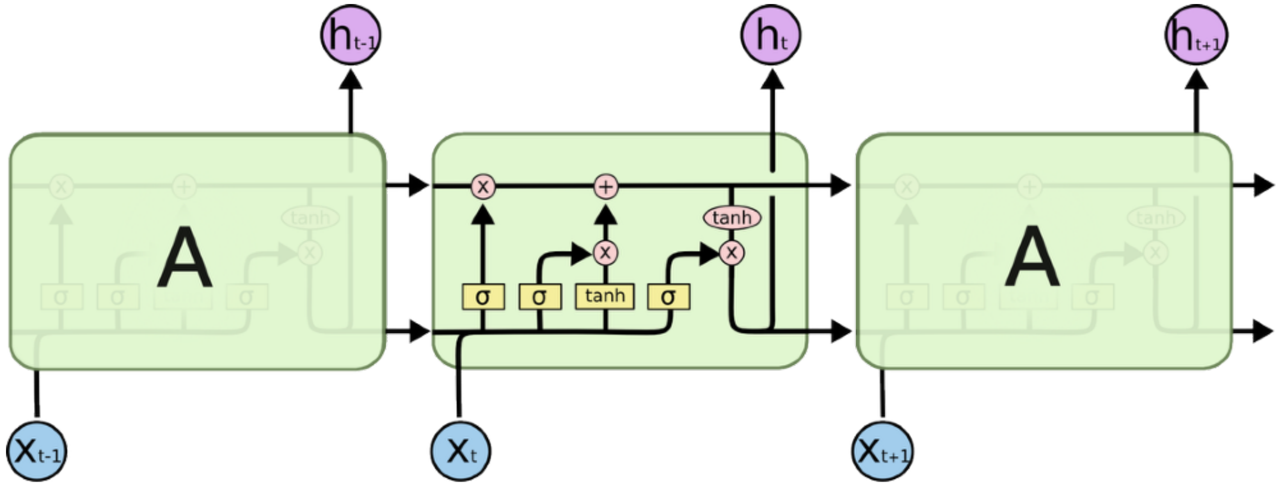
The girl who the newspaper now calls his girlfriend has really been hateful . ✓

The monologue that the actor who the movie industry likes made silent was being uploaded . ✓ (!)

The man who the car…

(continuations generated by model of Jozefowicz et al., 2016)

What have these models learned as “English”?



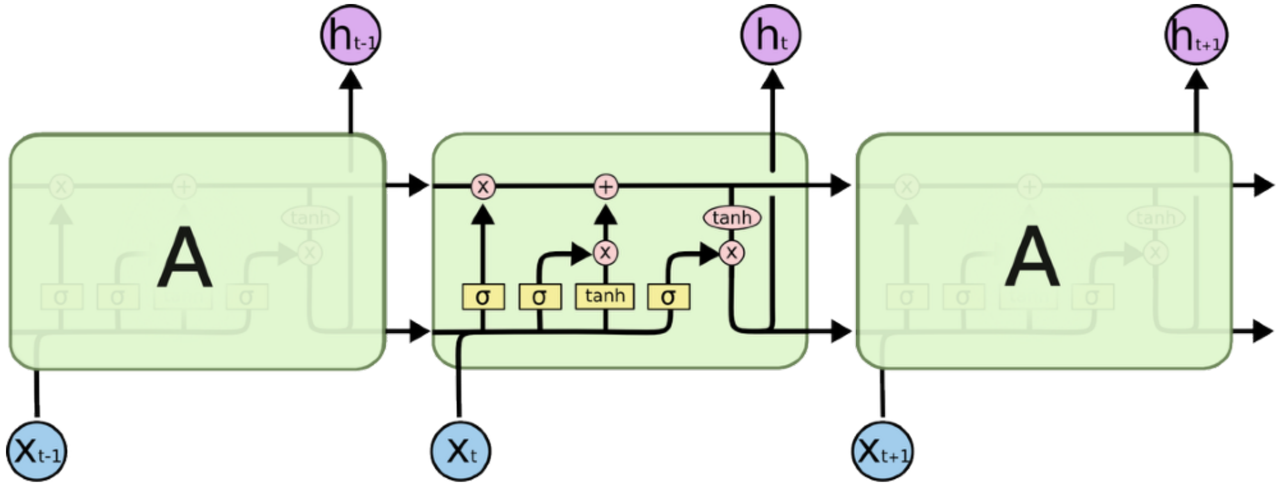
The girl who the newspaper now calls his girlfriend has really been hateful . ✓

The monologue that the actor who the movie industry likes made silent was being uploaded . ✓ (!)

The man who the car has gazed longingly at for years .

(continuations generated by model of Jozefowicz et al., 2016)

What have these models learned as “English”?



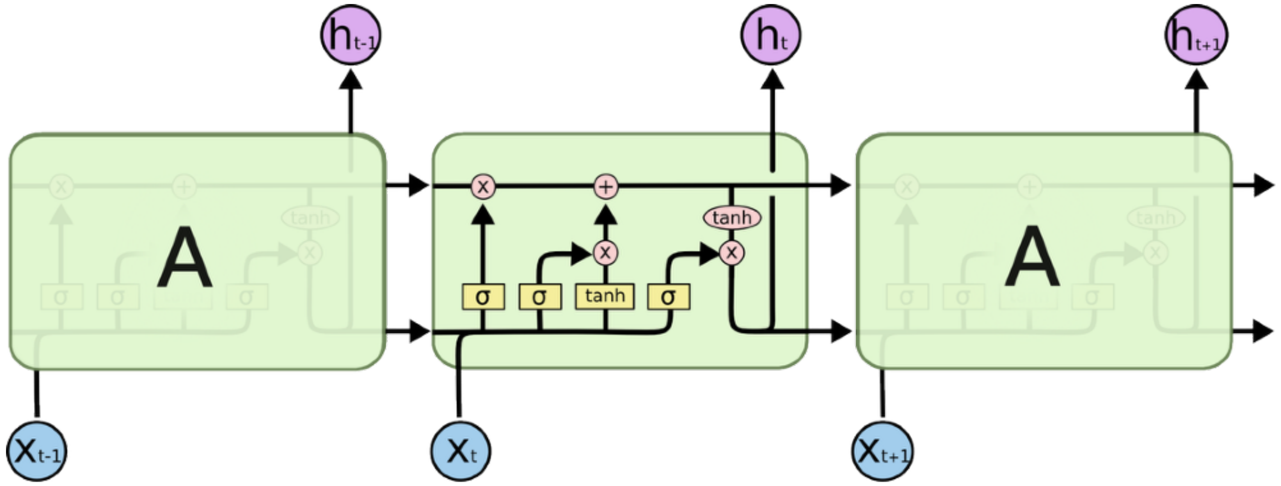
The girl who the newspaper now calls his girlfriend has really been hateful . ✓

The monologue that the actor who the movie industry likes made silent was being uploaded . ✓ (!)

The man who the car has gazed longingly at for years . ✗

(continuations generated by model of Jozefowicz et al., 2016)

What have these models learned as “English”?



The girl who the newspaper now calls his girlfriend has really been hateful . ✓

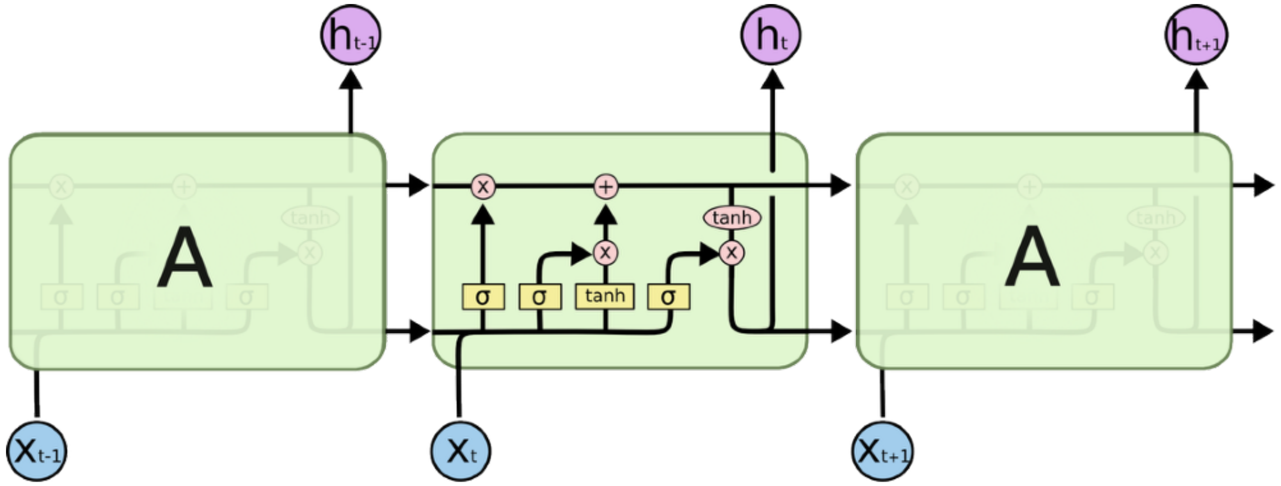
The monologue that the actor who the movie industry likes made silent was being uploaded . ✓ (!)

The man who the car has gazed longingly at for years . ✗

The athlete who the restaurant…

(continuations generated by model of Jozefowicz et al., 2016)

What have these models learned as “English”?



The girl who the newspaper now calls his girlfriend has really been hateful .

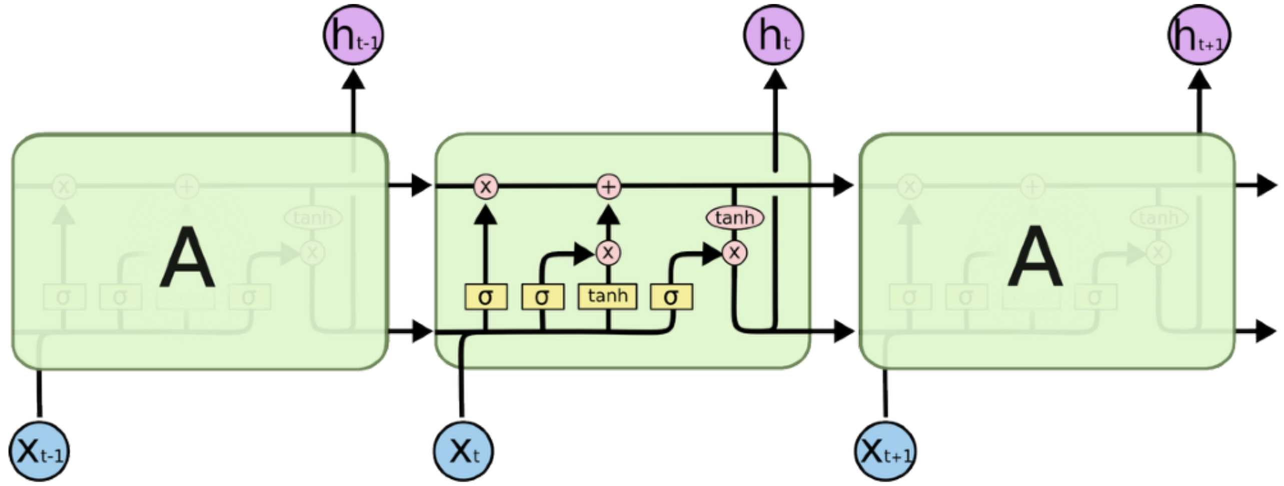
The monologue that the actor who the movie industry likes made silent was being uploaded . (!)

The man who the car has gazed longingly at for years .

The athlete who the restaurant would justify decided to add the main West Coast restaurants to his menu and who hadn 't upgraded from his previous suite , into a more <UNK> steakhouse in New York .

(continuations generated by model of Jozefowicz et al., 2016)

What have these models learned as “English”?



The girl who the newspaper now calls his girlfriend has really been hateful . ✓

The monologue that the actor who the movie industry likes made silent was being uploaded . ✓ (!)

The man who the car has gazed longingly at for years . ✗

The athlete who the restaurant would justify decided to add the main West Coast restaurants to his menu and who hadn 't upgraded from his previous suite , into a more <UNK> steakhouse in New York . ???

(continuations generated by model of Jozefowicz et al., 2016)

What do language models learn?

What do language models learn?

- In the case of *n-gram* models…

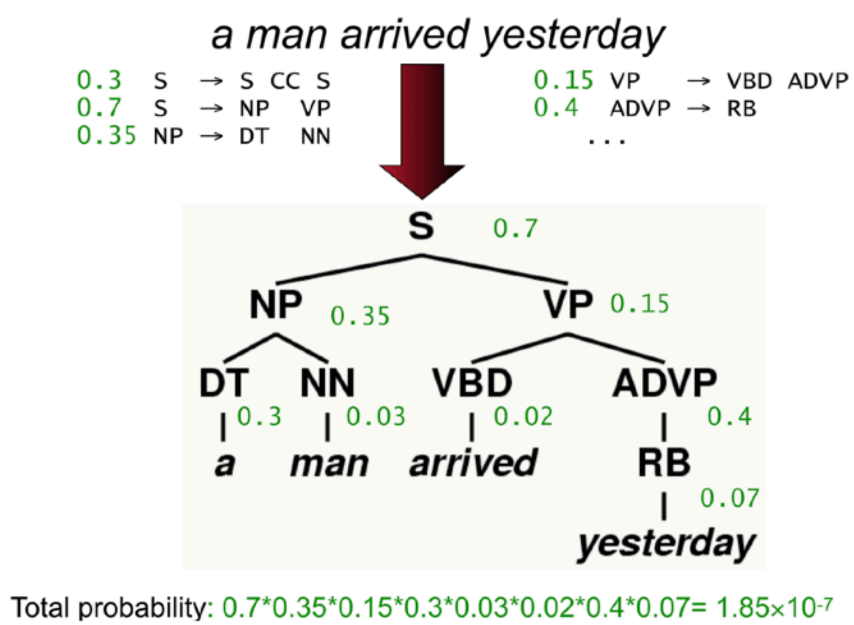
$$P(\texttt{\$ dogs chase cats \$}) \approx P(\texttt{\$}|\texttt{cats})P(\texttt{cats}|\texttt{chase})P(\texttt{chase}|\texttt{dogs})P(\texttt{dogs}|\texttt{\$})$$

What do language models learn?

- In the case of *n-gram* models…

$$P(\texttt{\$ dogs chase cats \$}) \approx P(\texttt{\$}|\texttt{cats})P(\texttt{cats}|\texttt{chase})P(\texttt{chase}|\texttt{dogs})P(\texttt{dogs}|\texttt{\$})$$

- …and probabilistic context-free grammars…

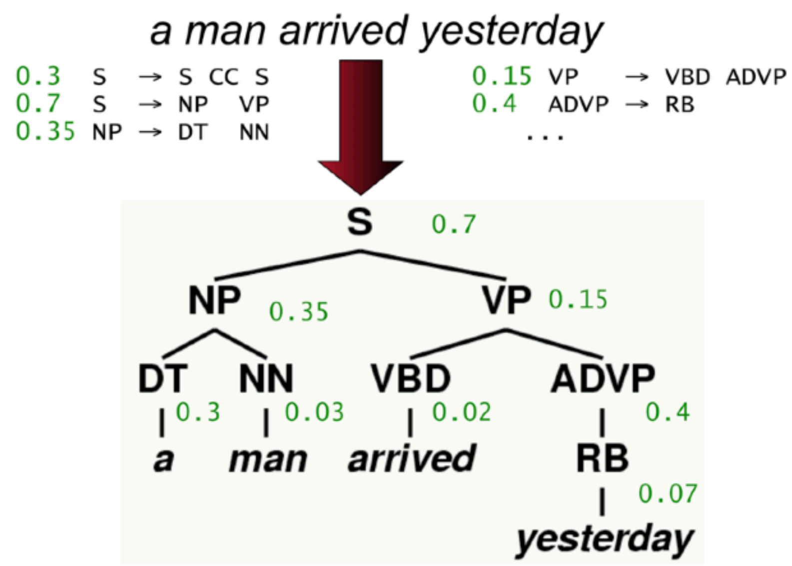


What do language models learn?

- In the case of *n-gram* models…

$$P(\$\ \texttt{dogs chase cats}\ \$) \approx P(\$|\texttt{cats})P(\texttt{cats}|\texttt{chase})P(\texttt{chase}|\texttt{dogs})P(\texttt{dogs}|\$)$$

- …and probabilistic context-free grammars…



Total probability: 0.7*0.35*0.15*0.3*0.03*0.02*0.4*0.07= 1.85×10^{-7}

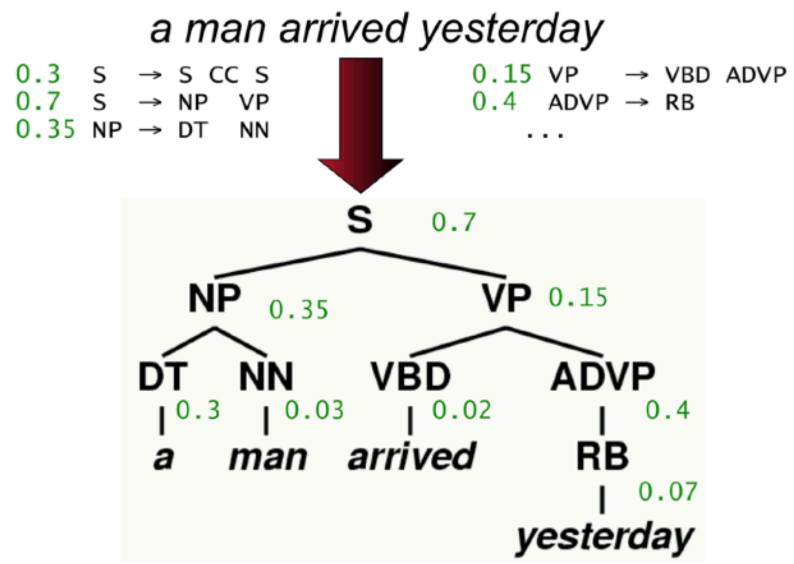
- …we can reason (fairly) easily about model predictions.

What do language models learn?

- In the case of *n-gram* models…

$$P(\texttt{\$ dogs chase cats \$}) \approx P(\texttt{\$}|\texttt{cats})P(\texttt{cats}|\texttt{chase})P(\texttt{chase}|\texttt{dogs})P(\texttt{dogs}|\texttt{\$})$$

- …and probabilistic context-free grammars…



Total probability: 0.7*0.35*0.15*0.3*0.03*0.02*0.4*0.07= 1.85×10-7

- …we can reason (fairly) easily about model predictions.

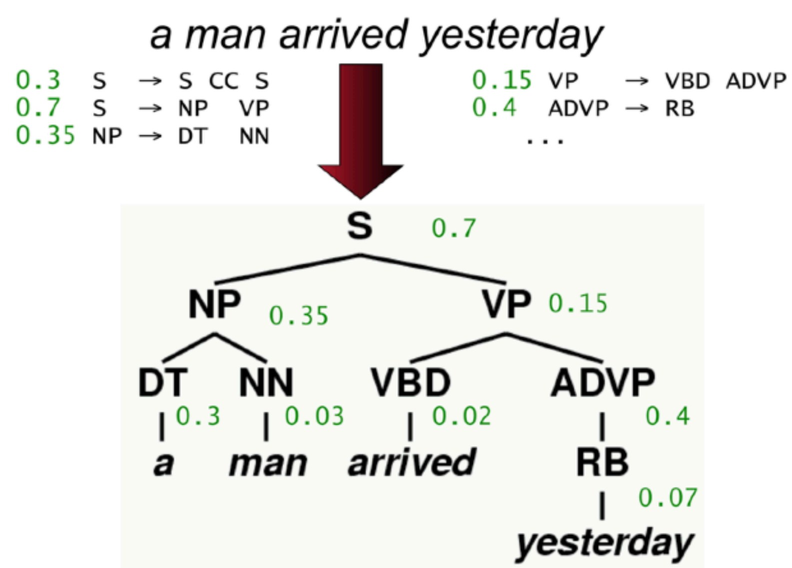
A friend of my parents…

What do language models learn?

- In the case of *n-gram* models…

$$P(\texttt{\$ dogs chase cats \$}) \approx P(\texttt{\$}|\texttt{cats})P(\texttt{cats}|\texttt{chase})P(\texttt{chase}|\texttt{dogs})P(\texttt{dogs}|\texttt{\$})$$

- …and probabilistic context-free grammars…



Total probability: 0.7*0.35*0.15*0.3*0.03*0.02*0.4*0.07= 1.85×10^{-7}

- …we can reason (fairly) easily about model predictions.

  A friend of my parents…

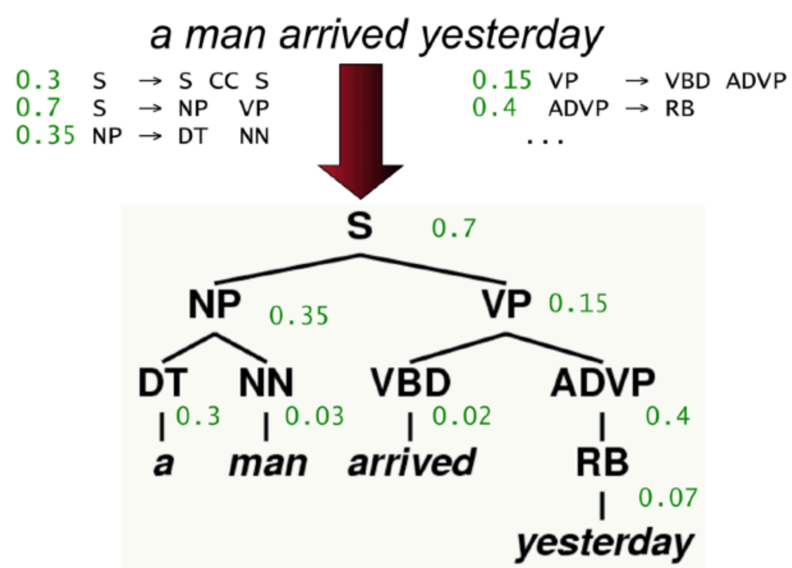
- These are ***highly explainable*** probabilistic models.

What do language models learn?

- In the case of *n-gram* models…

$$P(\$\ \texttt{dogs chase cats}\ \$) \approx P(\$|\texttt{cats})P(\texttt{cats}|\texttt{chase})P(\texttt{chase}|\texttt{dogs})P(\texttt{dogs}|\$)$$

- …and probabilistic context-free grammars…



Total probability: 0.7*0.35*0.15*0.3*0.03*0.02*0.4*0.07= 1.85×10^{-7}

- …we can reason (fairly) easily about model predictions.

A friend of my parents… → $P_{trigram}(w|\textit{A friend of my parents…})$

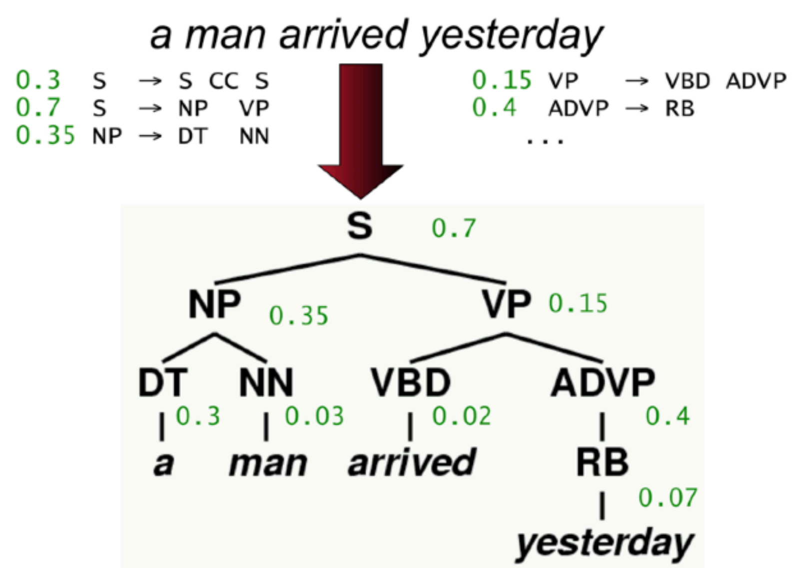
- These are ***highly explainable*** probabilistic models.

What do language models learn?

- In the case of *n-gram* models…

$$P(\$\ \texttt{dogs chase cats}\ \$) \approx P(\$|\texttt{cats})P(\texttt{cats}|\texttt{chase})P(\texttt{chase}|\texttt{dogs})P(\texttt{dogs}|\$)$$

- …and probabilistic context-free grammars…



Total probability: 0.7*0.35*0.15*0.3*0.03*0.02*0.4*0.07= 1.85×10^{-7}

- …we can reason (fairly) easily about model predictions.

A friend of my parents… → $P_{trigram}(w|\textit{A friend of my parents…})$

A friend of my parents… → $P_{PCFG}(w|\textit{A friend of my parents…})$

- These are ***highly explainable*** probabilistic models.

Explainable models, explainable AI

Explainable models, explainable AI

- An ideal “explainable” state-of-the-art AI system:

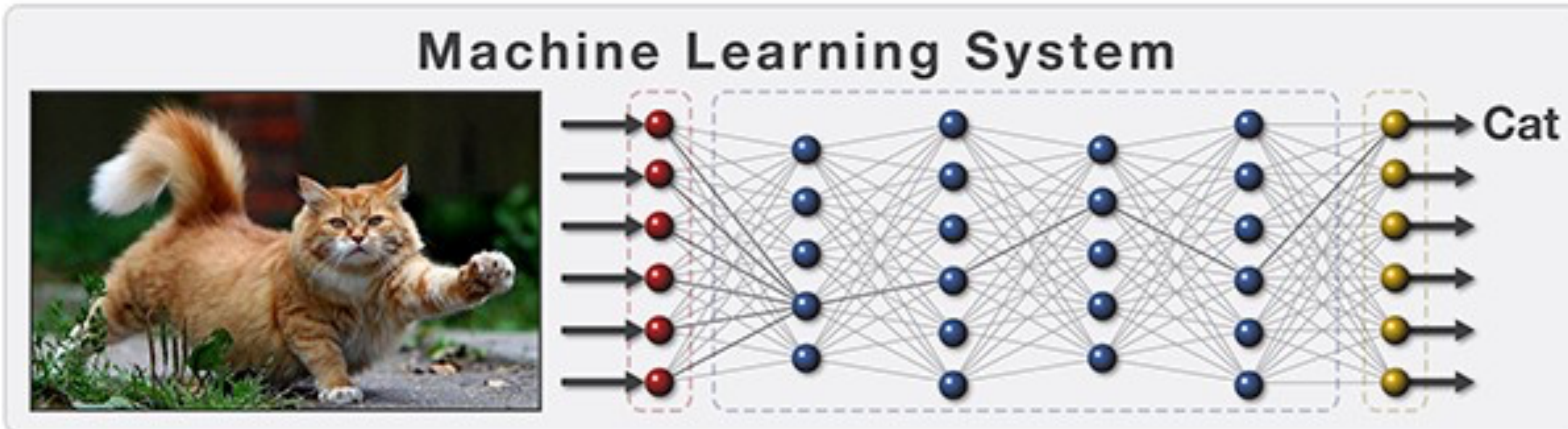


Image due to Cenk Sezgin, https://cenksezgin.com/explainable-artificial-intelligence-xai-third-wave-of-ai-5879d6f40a26

Explainable models, explainable AI

- An ideal “explainable” state-of-the-art AI system:

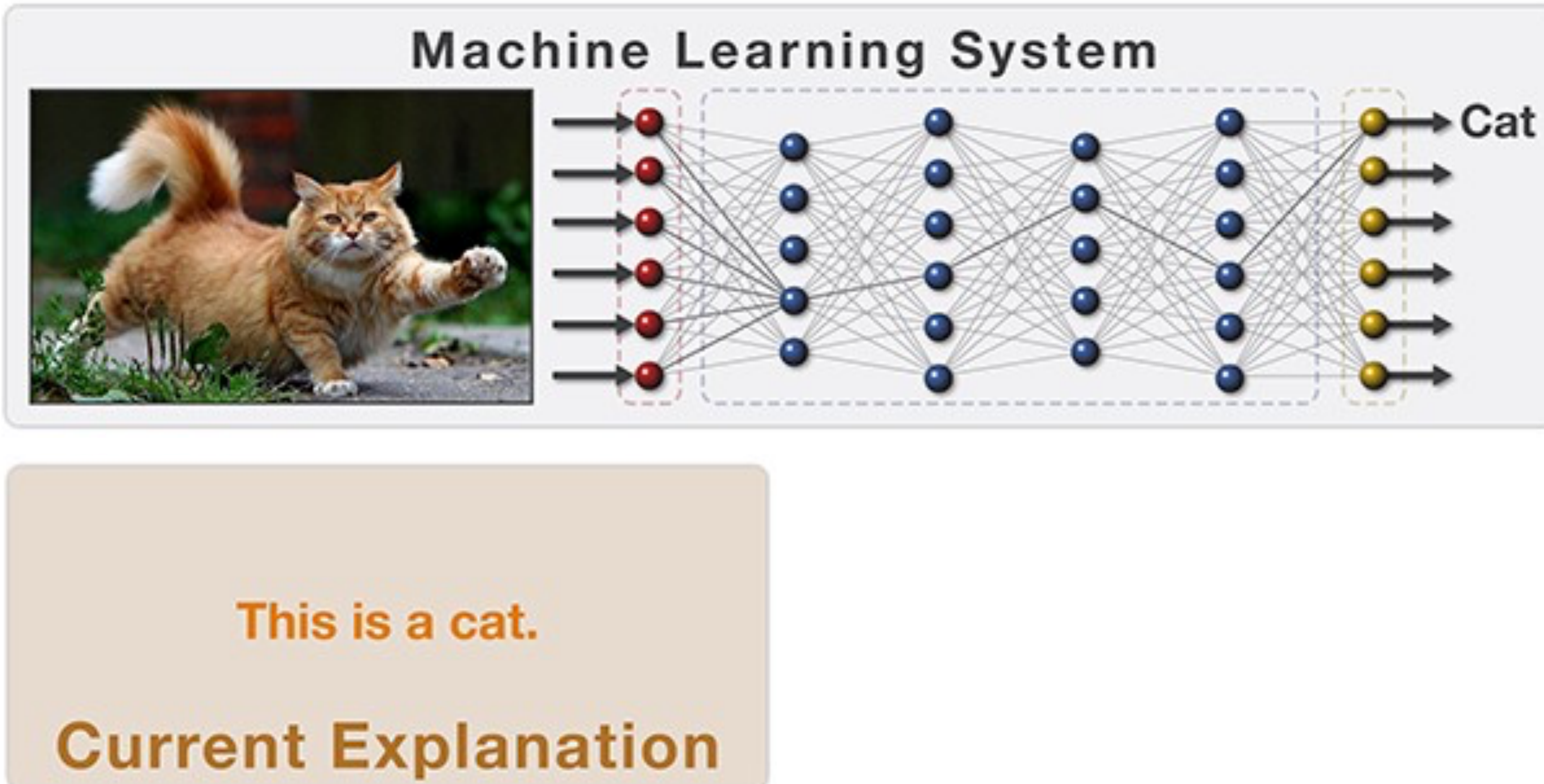


Image due to Cenk Sezgin, https://cenksezgin.com/explainable-artificial-intelligence-xai-third-wave-of-ai-5879d6f40a26

Explainable models, explainable AI

- An ideal “explainable” state-of-the-art AI system:

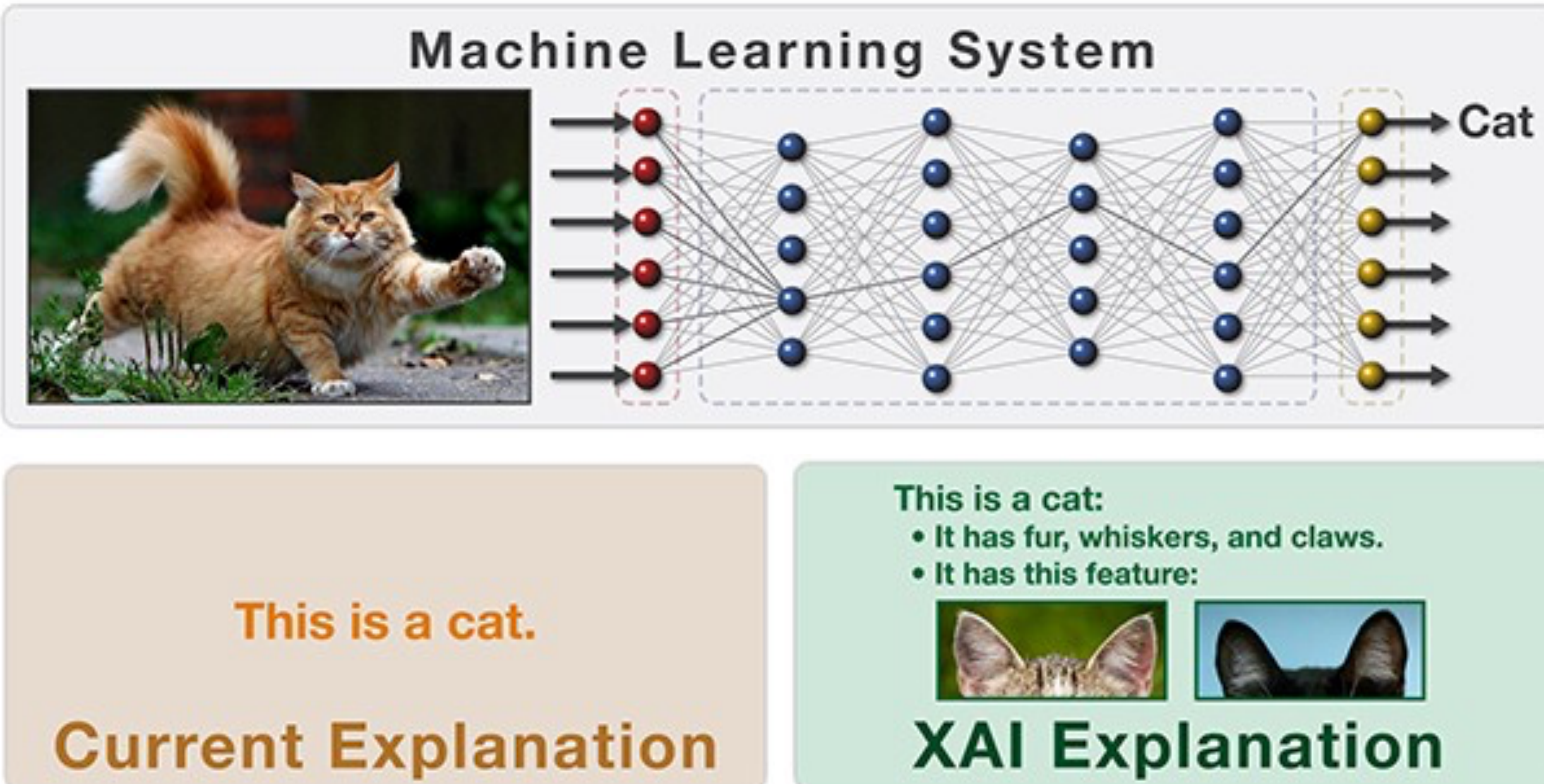
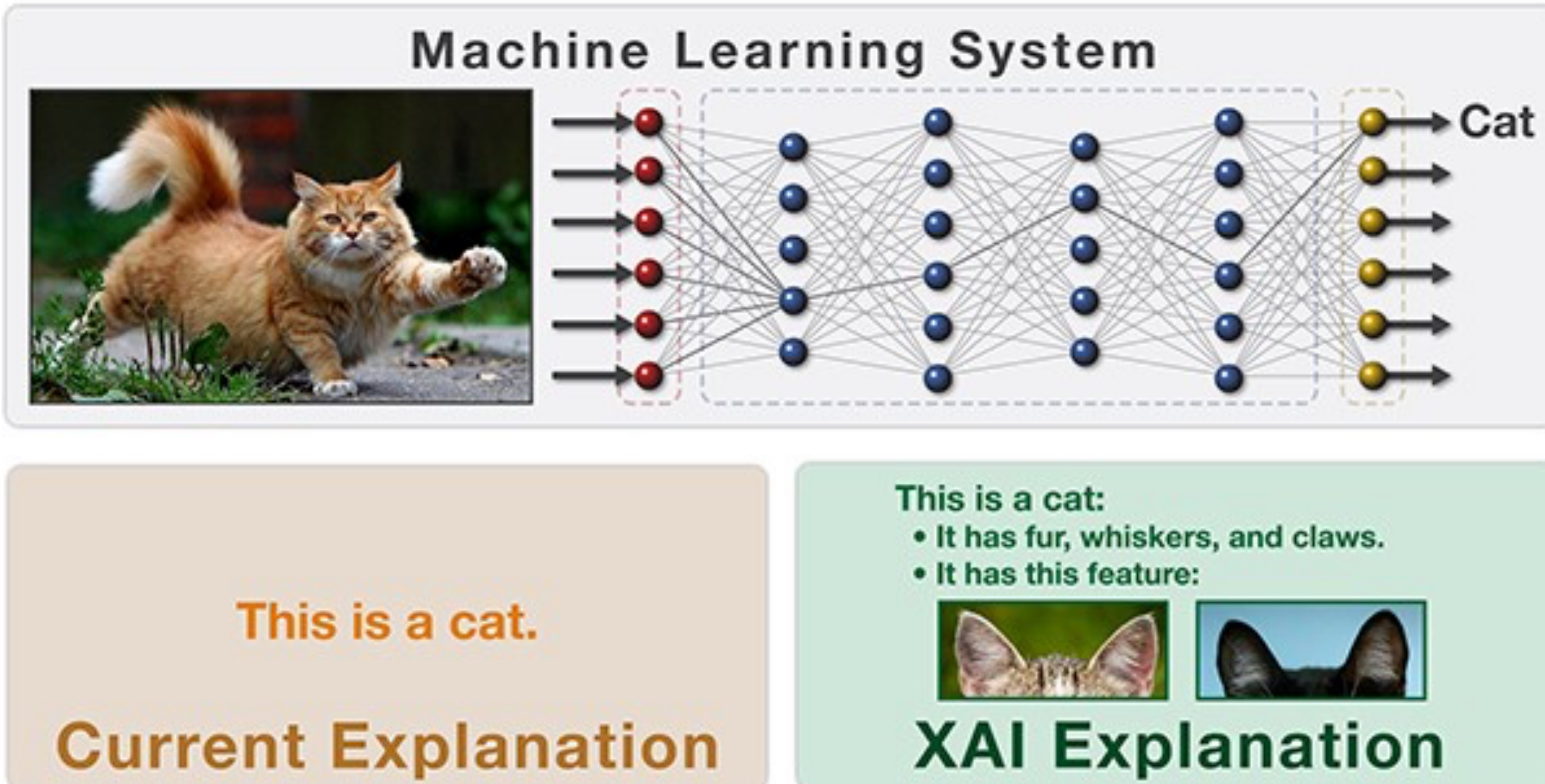


Image due to Cenk Sezgin, https://cenksezgin.com/explainable-artificial-intelligence-xai-third-wave-of-ai-5879d6f40a26

Explainable models, explainable AI

- An ideal "explainable" state-of-the-art AI system:



- State-of-the-art AI systems are increasingly ***opaque***, presenting challenges for explainability

Image due to Cenk Sezgin, https://cenksezgin.com/explainable-artificial-intelligence-xai-third-wave-of-ai-5879d6f40a26

Importance of explainability

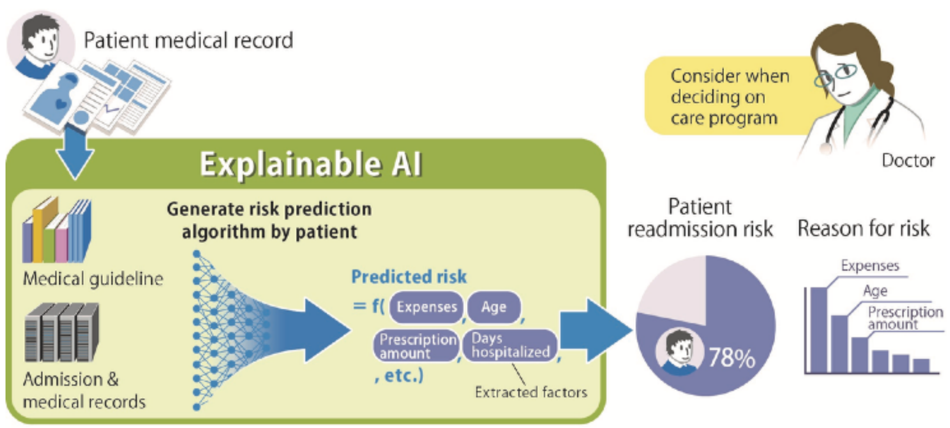
Importance of explainability

- Unexplainable systems are not ***accountable***, presenting problems for:

Importance of explainability

- Unexplainable systems are not ***accountable***, presenting problems for:

Medical decision-making

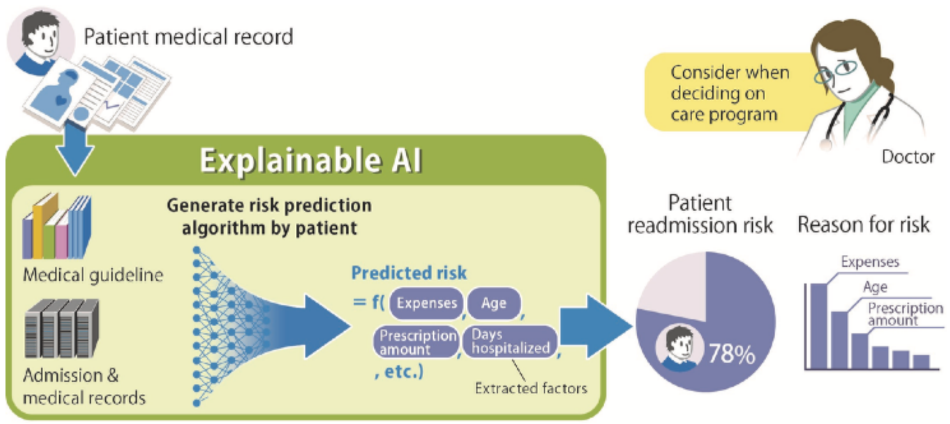


http://www.hitachi.us/sites/default/files/AI_health.jpg

Importance of explainability

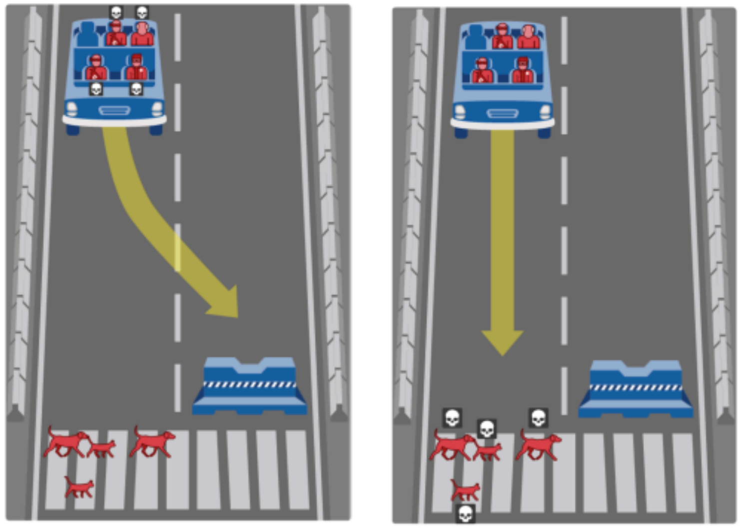
- Unexplainable systems are not ***accountable***, presenting problems for:

Medical decision-making



http://www.hitachi.us/sites/default/files/AI_health.jpg

Self-driving cars

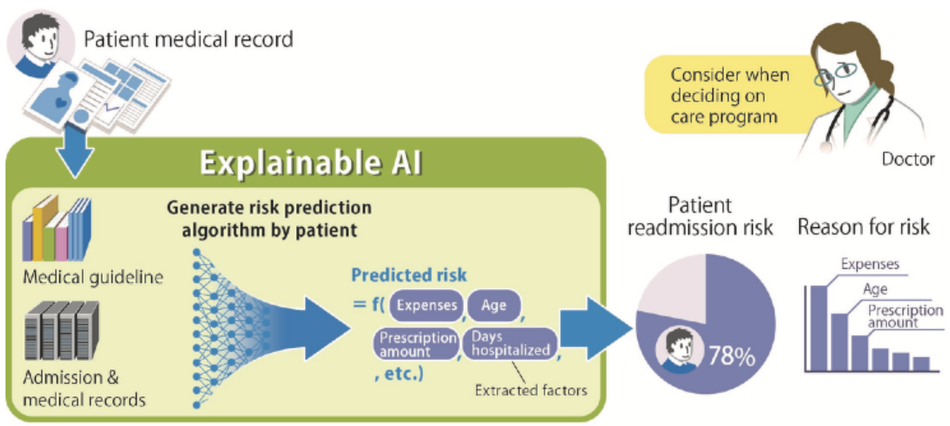


http://aptgadget.com/wp-content/uploads/2016/10/what-should-the-self-driving-car-do.png

Importance of explainability

- Unexplainable systems are not ***accountable***, presenting problems for:

Medical decision-making

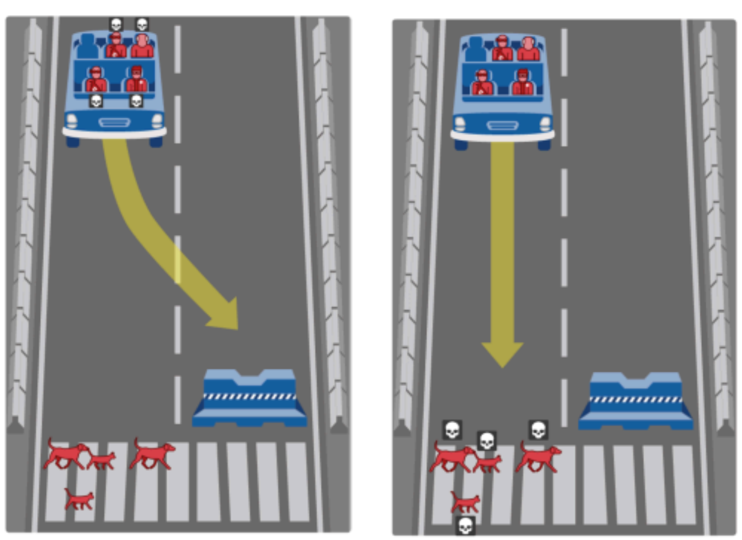


http://www.hitachi.us/sites/default/files/AI_health.jpg

Automated personnel decisions

https://www.wcn.uk/files/2016-05/making-the-most-of-automation-in-recruitment-wcn.png

Self-driving cars

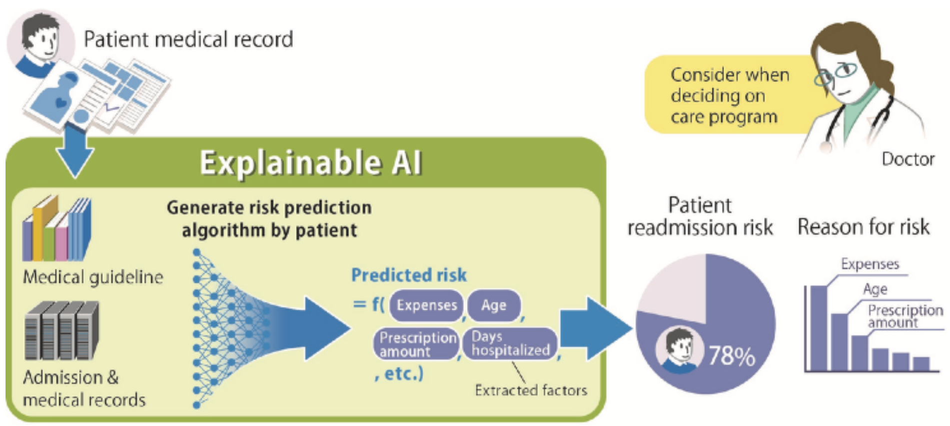


http://aptgadget.com/wp-content/uploads/2016/10/what-should-the-self-driving-car-do.png

Importance of explainability

- Unexplainable systems are not ***accountable***, presenting problems for:

Medical decision-making

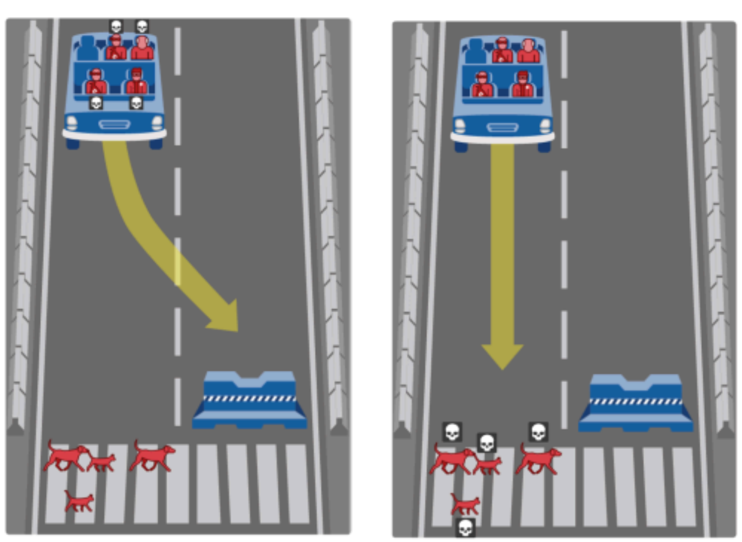


http://www.hitachi.us/sites/default/files/AI_health.jpg

Automated personnel decisions

https://www.wcn.uk/files/2016-05/making-the-most-of-automation-in-recruitment-wcn.png

Self-driving cars



http://aptgadget.com/wp-content/uploads/2016/10/what-should-the-self-driving-car-do.png

…and more!

Explaining is fundamentally human

Explaining is fundamentally human

"Knocking over a tower is the best part of building it!"

Explaining is fundamentally human

“Knocking over a tower is the best part of building it!”

“I was jealous of my brother”

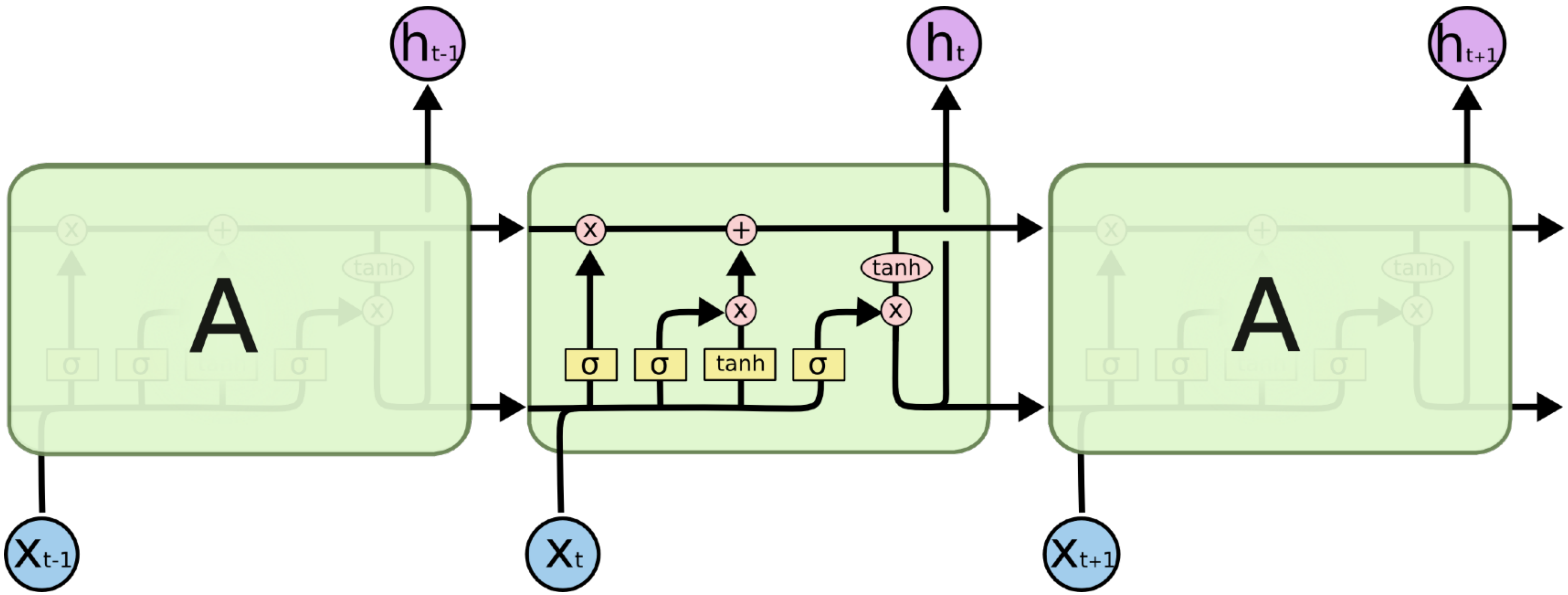
Explaining is fundamentally human

“Knocking over a tower is the best part of building it!”

“The tower was in the way of the TV set”

“I was jealous of my brother”

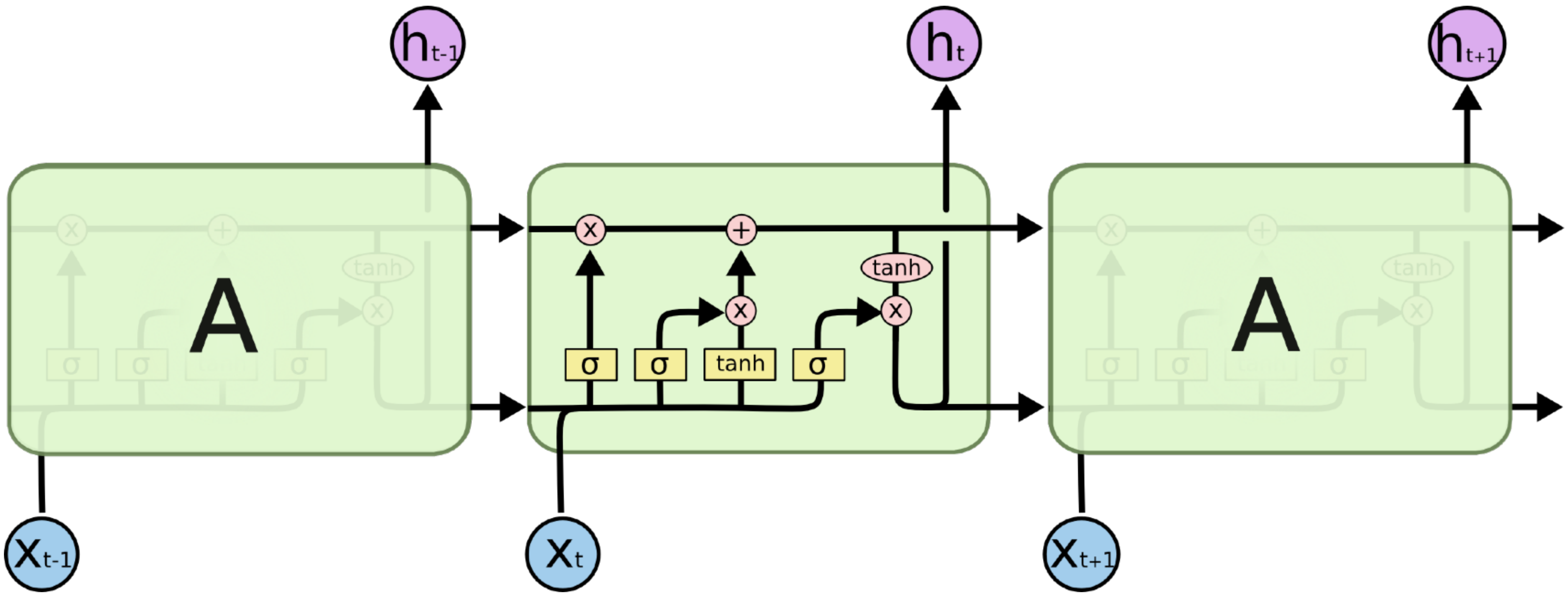
Understanding & explaining NLM behavior



visualization due to Christopher Olah, http://colah.github.io/posts/2015-08-Understanding-LSTMs/

Understanding & explaining NLM behavior

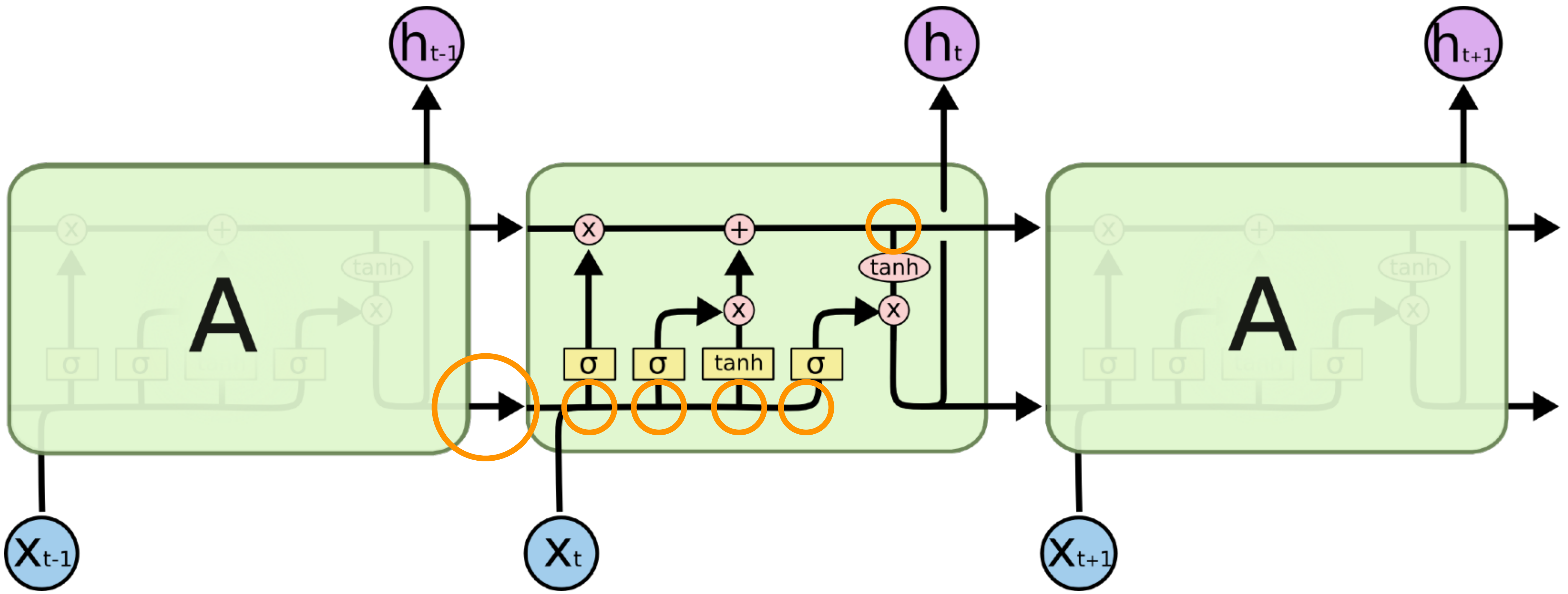
- Remember, this is the state of the art for language models:



visualization due to Christopher Olah, http://colah.github.io/posts/2015-08-Understanding-LSTMs/

Understanding & explaining NLM behavior

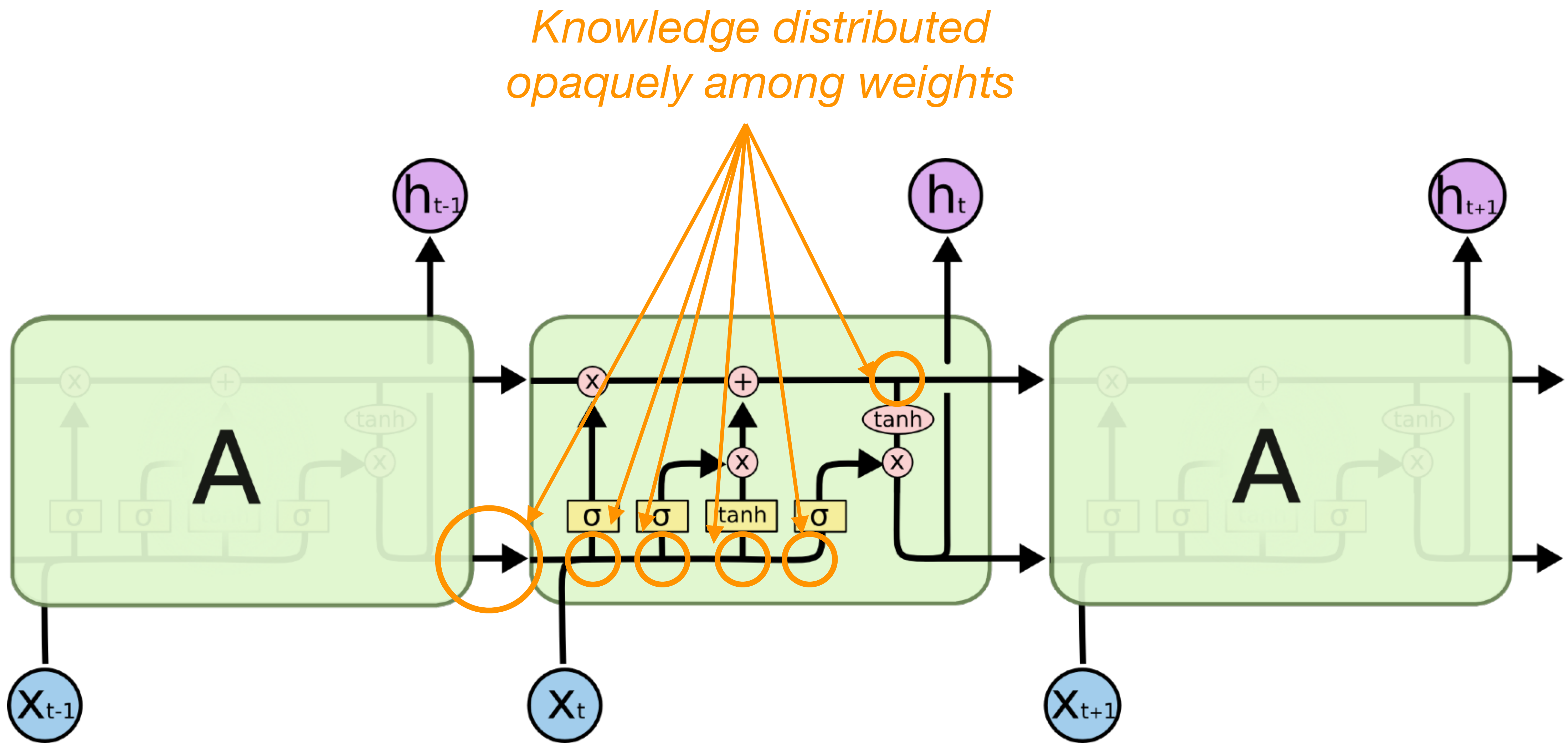
- Remember, this is the state of the art for language models:



visualization due to Christopher Olah, http://colah.github.io/posts/2015-08-Understanding-LSTMs/

Understanding & explaining NLM behavior

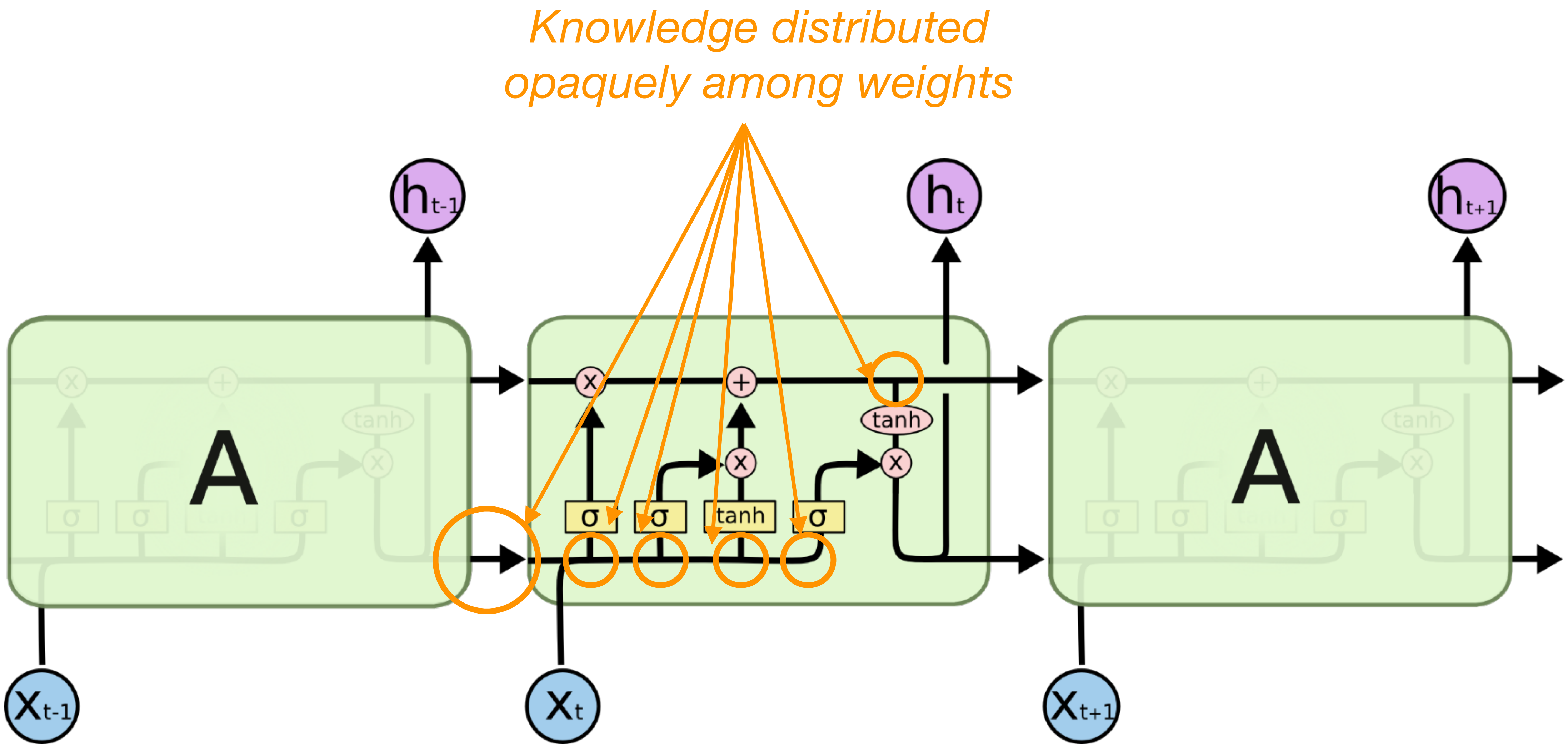
- Remember, this is the state of the art for language models:



visualization due to Christopher Olah, http://colah.github.io/posts/2015-08-Understanding-LSTMs/

Understanding & explaining NLM behavior

- Remember, this is the state of the art for language models:

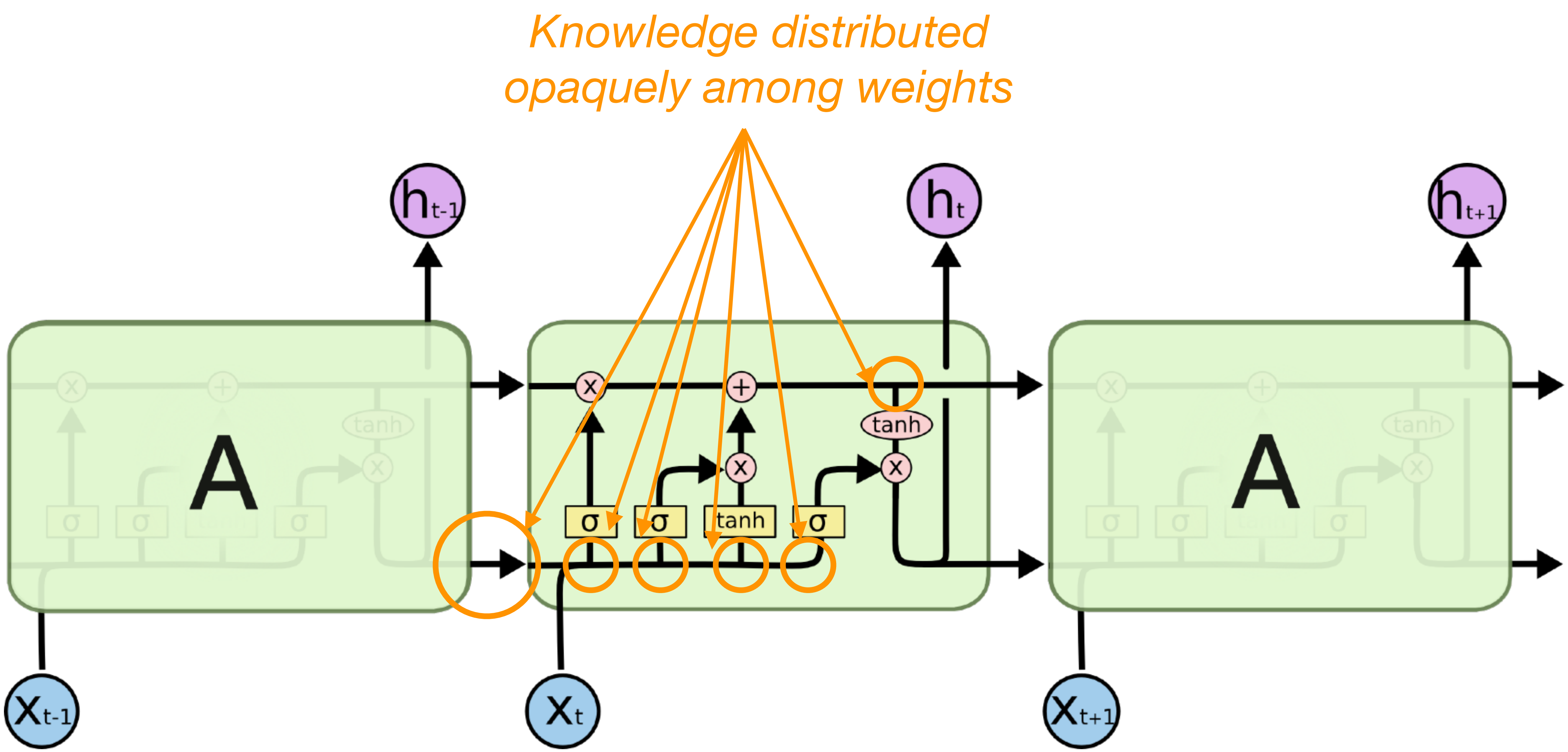


- How can we make such a system explainable?

visualization due to Christopher Olah, http://colah.github.io/posts/2015-08-Understanding-LSTMs/

Understanding & explaining NLM behavior

- Remember, this is the state of the art for language models:

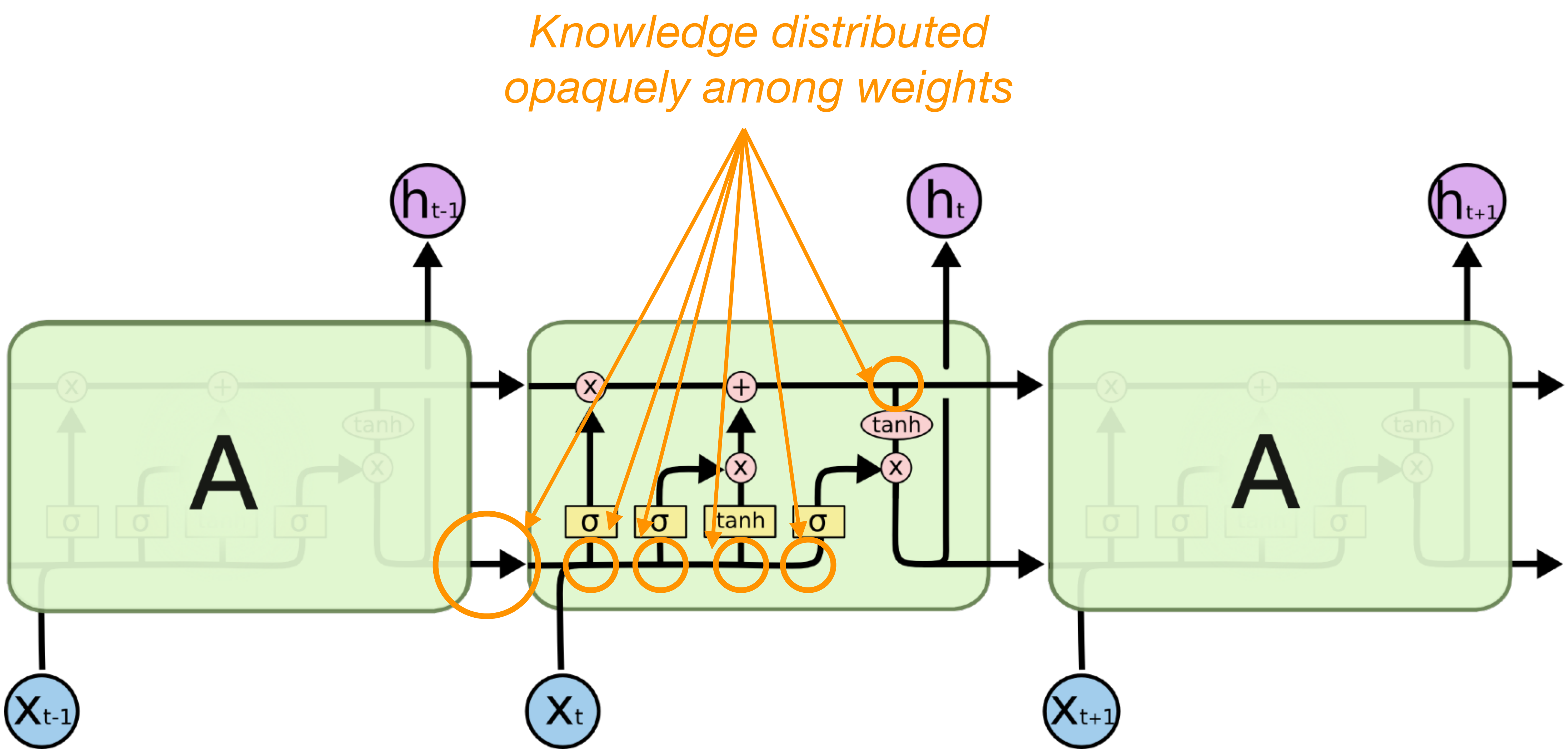


- How can we make such a system explainable?
  - Controlled tests of a model's capabilities & behaviors

visualization due to Christopher Olah, http://colah.github.io/posts/2015-08-Understanding-LSTMs/

Understanding & explaining NLM behavior

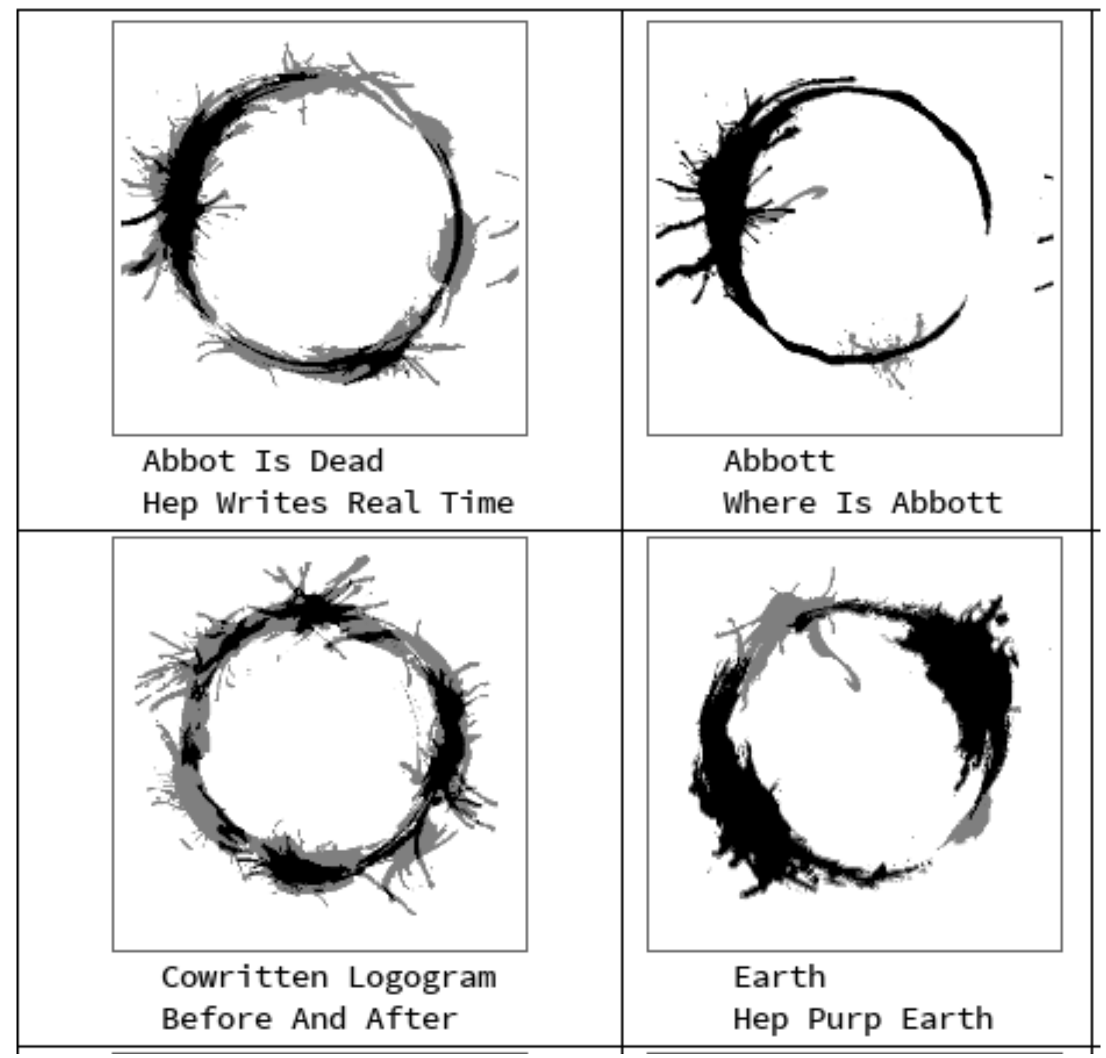
- Remember, this is the state of the art for language models:



- How can we make such a system explainable?
  - Controlled tests of a model's capabilities & behaviors
  - Probing the circuitry underlying these behaviors

Technical question:

What generalizations are these models learning?



Theoretical question:

How well would positive* input data alone deliver the right linguistic generalizations to a generic flexible learner without strong hierarchical bias?

No negative evidence!

Theoretical question:

How well would positive* input data alone deliver the right linguistic generalizations to a generic flexible learner without strong hierarchical bias?

***No negative evidence!**

Explaining a model's linguistic behavior

Explaining a model's linguistic behavior

- Let's now spend some time brainstorming how to examine a model's behavior in more detail

Explaining a model's linguistic behavior

- Let's now spend some time brainstorming how to examine a model's behavior in more detail
- Papers that have examined this more systematically:

Explaining a model's linguistic behavior

- Let's now spend some time brainstorming how to examine a model's behavior in more detail
- Papers that have examined this more systematically:

Assessing the Ability of LSTMs to Learn Syntax-Sensitive Dependencies

Tal Linzen[1,2] **Emmanuel Dupoux**[1]
LSCP[1] & IJN[2], CNRS,
EHESS and ENS, PSL Research University
{tal.linzen,
emmanuel.dupoux}@ens.fr

Yoav Goldberg
Computer Science Department
Bar Ilan University
yoav.goldberg@gmail.com

Linzen et al., 2016, TACL

Neural Language Models as Psycholinguistic Subjects: Representations of Syntactic State

Richard Futrell[1], **Ethan Wilcox**[2], **Takashi Morita**[3,4], **Peng Qian**[5], **Miguel Ballesteros**[6], and **Roger Levy**[5]

[1]Department of Language Science, UC Irvine, rfutrell@uci.edu
[2]Department of Linguistics, Harvard University, wilcoxeg@g.harvard.edu
[3]Primate Research Institute, Kyoto University, tmorita@alum.mit.edu
[4]Department of Linguistics and Philosophy, MIT
[5]Department of Brain and Cognitive Sciences, MIT, rplevy@mit.edu
[6]IBM Research, MIT-IBM Watson AI Lab, miguel.ballesteros@ibm.com

Futrell et al., 2019, NAACL

Colorless green recurrent networks dream hierarchically

Kristina Gulordava*
Department of Linguistics
University of Geneva
kristina.gulordava@unige.ch

Piotr Bojanowski
Facebook AI Research
Paris
bojanowski@fb.com

Edouard Grave
Facebook AI Research
New York
egrave@fb.com

Tal Linzen
Department of Cognitive Science
Johns Hopkins University
tal.linzen@jhu.edu

Marco Baroni
Facebook AI Research
Paris
mbaroni@fb.com

Gulordava et al., 2018, NAACL

What do RNN Language Models Learn about Filler–Gap Dependencies?

Ethan Wilcox[1], **Roger Levy**[2], **Takashi Morita**[3,4], and **Richard Futrell**[5]

[1]Department of Linguistics, Harvard University, wilcoxeg@g.harvard.edu
[2]Department of Brain and Cognitive Sciences, MIT, rplevy@mit.edu
[3]Primate Research Institute, Kyoto University, tmorita@alum.mit.edu
[4]Department of Linguistics and Philosophy, MIT
[5]Department of Language Science, UC Irvine, rfutrell@uci.edu

Wilcox et al., 2018, BlackBox NLP

Subject–verb agreement

The author laughs.	*The authors laughs.
*The author laugh.	The authors laugh.

Subject–verb agreement

`The author laughs.`	`*The authors laughs.`
`*The author laugh.`	`The authors laugh.`

Postmodification by prepositional phrase:

` The `**`author`**` of the novels `**`laughs`**`.`

`*The `**`author`**` of the novels `**`laugh`**`.`

Subject–verb agreement

`The author laughs.`	`*The authors laughs.`
`*The author laugh.`	`The authors laugh.`

Postmodification by prepositional phrase:

` The` **`author`** `of the novels` **`laughs.`**

`*The` **`author`** `of the novels` **`laugh.`** *"Attractor"*

Subject–verb agreement

```
 The author laughs.        *The authors laughs.
*The author laugh.          The authors laugh.
```

Postmodification by prepositional phrase:

 The **author** of the novels **laughs.**

*The **author** of the novels **laugh.**

"*Attractor*"

Embedding in complement clause:

 Readers know that the **author laughs.**

*Readers know that the **author laugh.**

Subject–verb agreement

The author laughs.	*The authors laughs.
*The author laugh.	The authors laugh.

Postmodification by prepositional phrase:

The **author** of the novels **laughs.**

*The **author** of the novels **laugh.**

"*Attractor*"

Embedding in complement clause:

Readers know that the **author laughs.**

*Readers know that the **author laugh.**

Readers know that the **author** who wrote the novels, regardless of what people may think, **laughs** in unusual circumstances.

*Readers know that the **author** who wrote the novels, regardless of what people may think, **laugh** in unusual circumstances.

Subject–verb agreement

The author laughs.
*The author laugh.

*The authors laughs.
The authors laugh.

Postmodification by prepositional phrase:

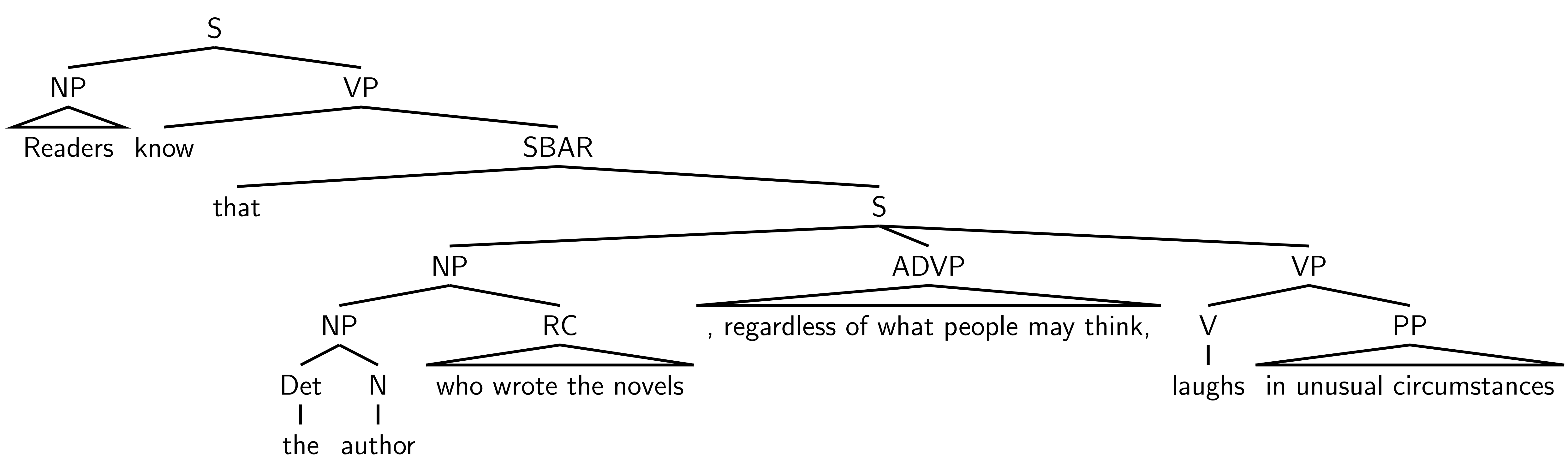
The **author** of the novels **laughs.**
*The **author** of the novels **laugh.**

"*Attractor*"

Embedding in complement clause:

Readers know that the **author laughs.**
*Readers know that the **author laugh.**

Readers know that the **author** who wrote the novels, regardless of what people may think, **laughs** in unusual circumstances.
*Readers know that the **author** who wrote the novels, regardless of what people may think, **laugh** in unusual circumstances.



Subject–verb agreement

```
 The author laughs.            *The authors laughs.
*The author laugh.              The authors laugh.
```

Postmodification by prepositional phrase:

 The **author** of the novels **laughs**.

*The **author** of the novels **laugh**.

"Attractor"

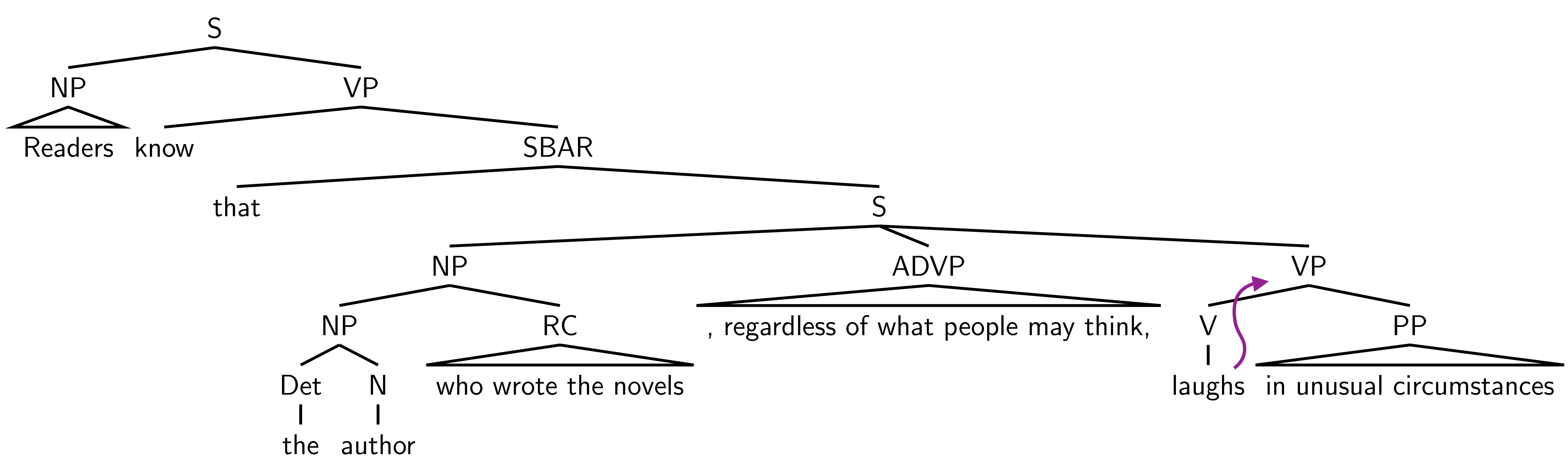
Embedding in complement clause:

 Readers know that the **author laughs**.

*Readers know that the **author laugh**.

 Readers know that the **author** who wrote the novels, regardless of what people may think, **laughs** in unusual circumstances.

*Readers know that the **author** who wrote the novels, regardless of what people may think, **laugh** in unusual circumstances.



Subject–verb agreement

The author laughs.
*The author laugh.

*The authors laughs.
The authors laugh.

Postmodification by prepositional phrase:

The **author** of the novels **laughs**.
*The **author** of the novels **laugh**.

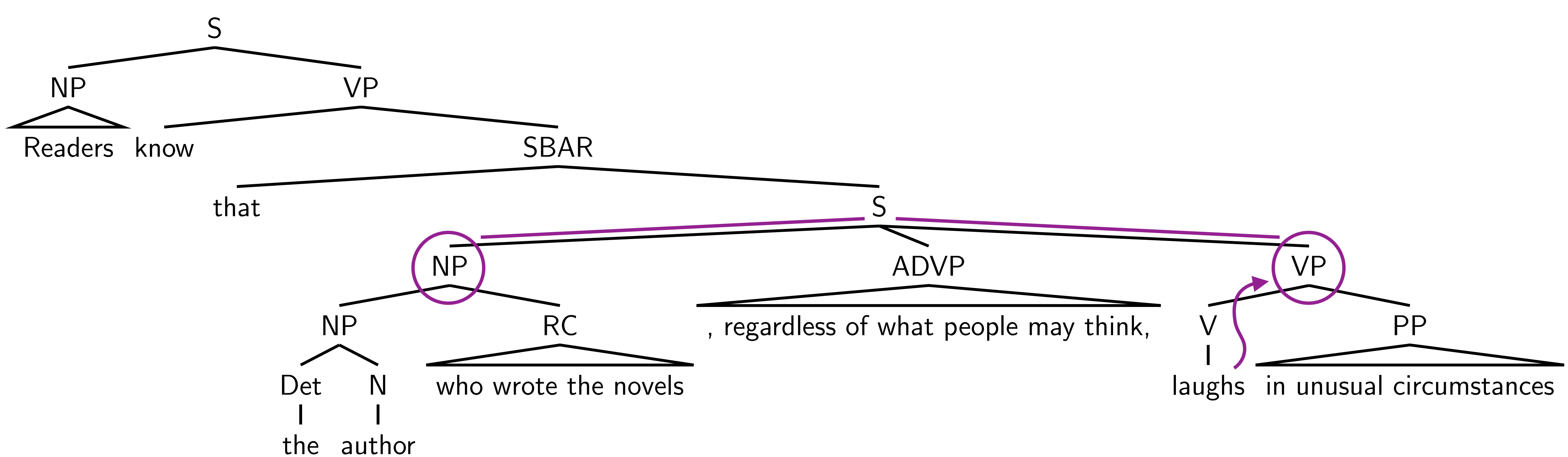
"*Attractor*"

Embedding in complement clause:

Readers know that the **author laughs**.
*Readers know that the **author laugh**.

Readers know that the **author** who wrote the novels, regardless of what people may think, **laughs** in unusual circumstances.

*Readers know that the **author** who wrote the novels, regardless of what people may think, **laugh** in unusual circumstances.



Subject–verb agreement

```
 The author laughs.            *The authors laughs.
*The author laugh.              The authors laugh.
```

Postmodification by prepositional phrase:

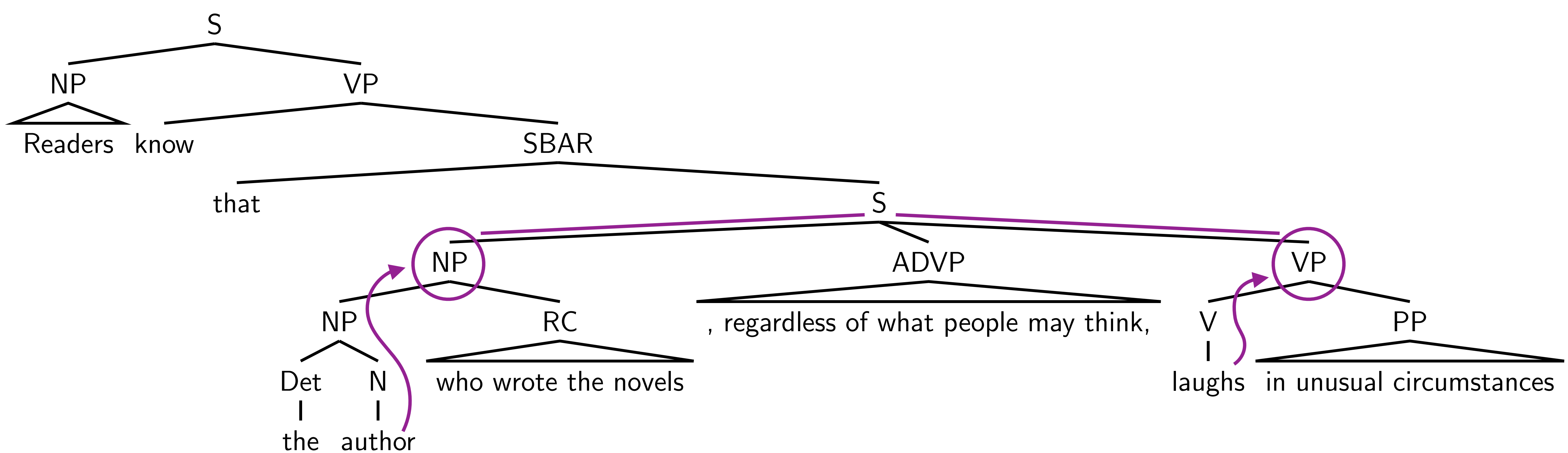
 The **author** of the novels **laughs**.
*The **author** of the novels **laugh**.

"Attractor"

Embedding in complement clause:

 Readers know that the **author laughs**.
*Readers know that the **author laugh**.

 Readers know that the **author** who wrote the novels, regardless of what people may think, **laughs** in unusual circumstances.
*Readers know that the **author** who wrote the novels, regardless of what people may think, **laugh** in unusual circumstances.



Subject–verb agreement

```
 The author laughs.         *The authors laughs.
*The author laugh.           The authors laugh.
```

Postmodification by prepositional phrase:

 The **author** of the novels **laughs**.

*The **author** of the novels **laugh**.

"Attractor"

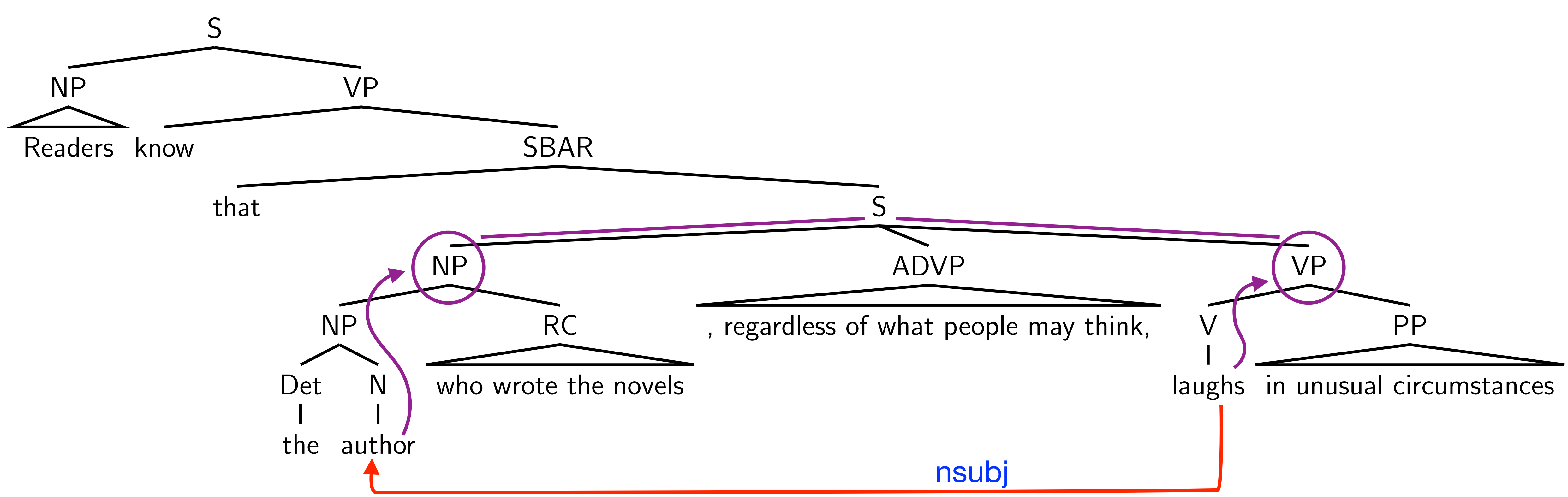
Embedding in complement clause:

 Readers know that the **author laughs**.

*Readers know that the **author laugh**.

 Readers know that the **author** who wrote the novels, regardless of what people may think, **laughs** in unusual circumstances.

*Readers know that the **author** who wrote the novels, regardless of what people may think, **laugh** in unusual circumstances.



Testing NLMs on subject–verb agreement

```
 The key to the cabinets is on the table.
*The key to the cabinets are on the table.
```

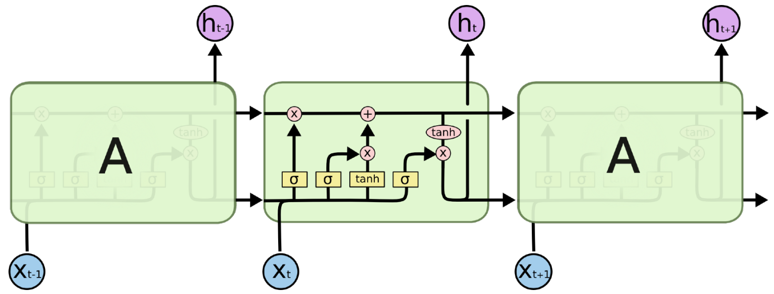
(*Linzen et al., 2016*)

Testing NLMs on subject–verb agreement

```
 The key to the cabinets is on the table.
*The key to the cabinets are on the table.
```

(*Linzen et al., 2016*)

Testing NLMs on subject–verb agreement

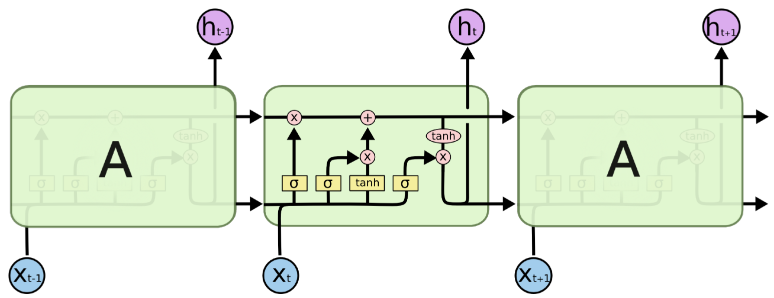



```
 The key to the cabinets | is on the table.
*The key to the cabinets | are on the table.
```

(Linzen et al., 2016)

Testing NLMs on subject–verb agreement

$$P(\texttt{is}|\text{Context}) \stackrel{?}{>} P(\texttt{are}|\text{Context})$$

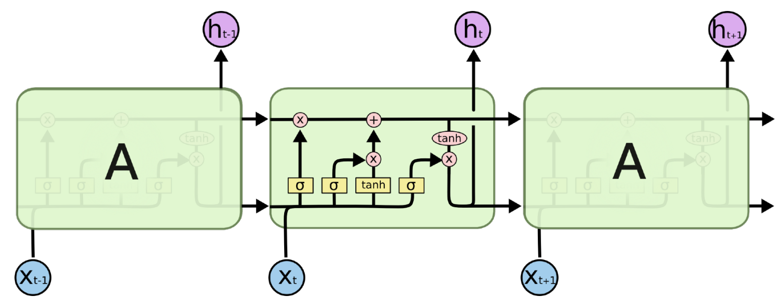



```
 The key to the cabinets | is on the table.
*The key to the cabinets | are on the table.
```

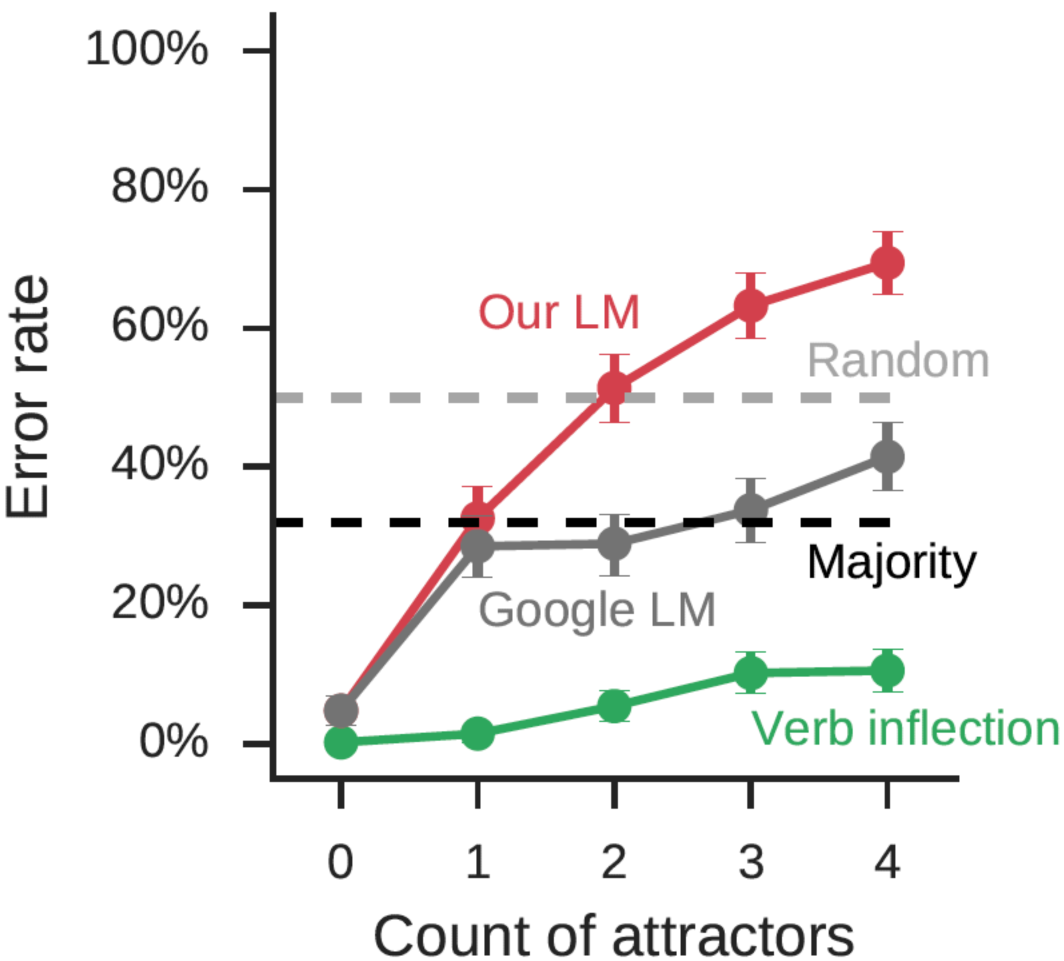
(Linzen et al., 2016)

Testing NLMs on subject–verb agreement

$$P(\texttt{is}|\text{Context}) \overset{?}{>} P(\texttt{are}|\text{Context})$$



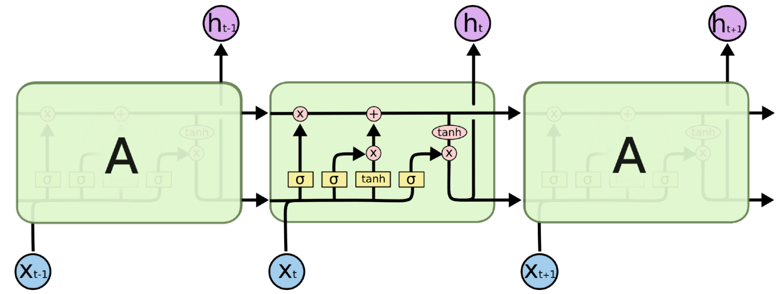

```
 The key to the cabinets is on the table.
*The key to the cabinets are on the table.
```



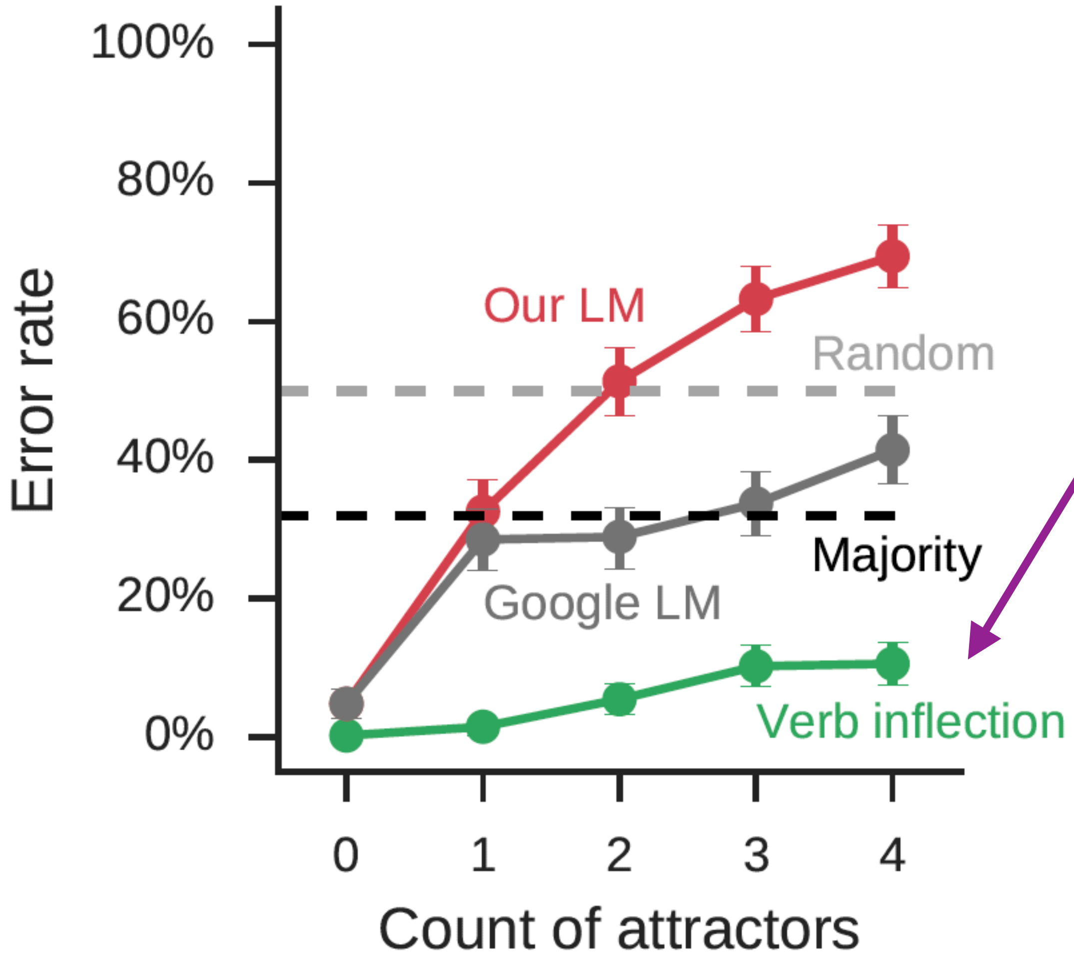
(Linzen et al., 2016)

Testing NLMs on subject–verb agreement

$$P(\texttt{is}|\text{Context}) \overset{?}{>} P(\texttt{are}|\text{Context})$$




```
 The key to the cabinets is on the table.
*The key to the cabinets are on the table.
```

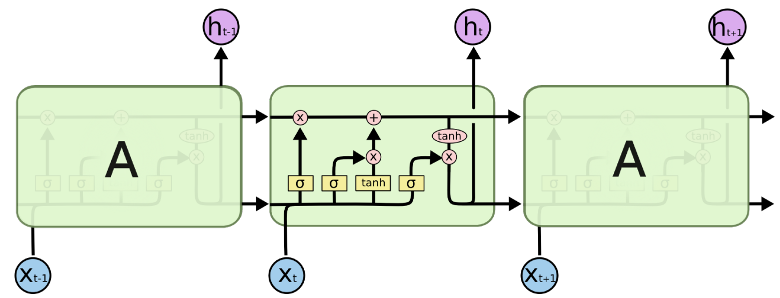


RNN models trained solely for the verb inflection task

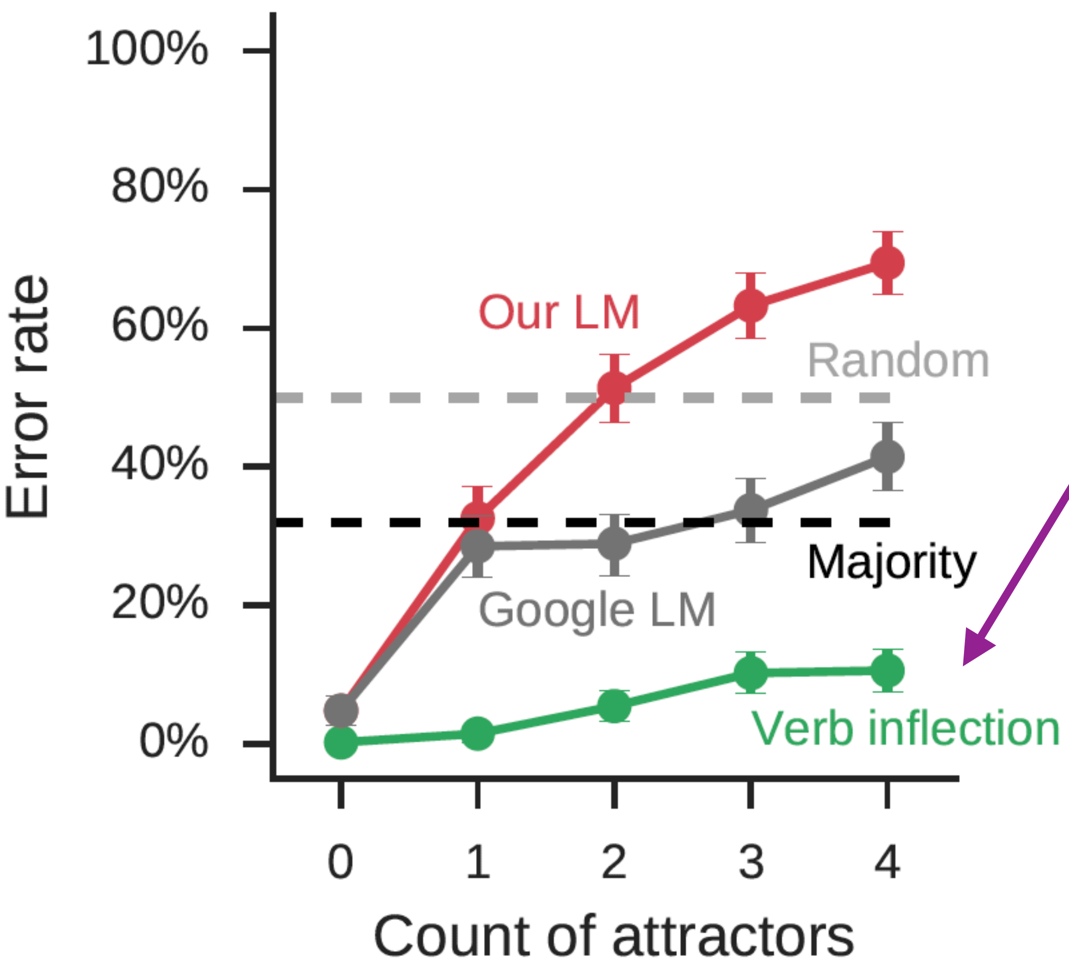
(Linzen et al., 2016)

Testing NLMs on subject–verb agreement

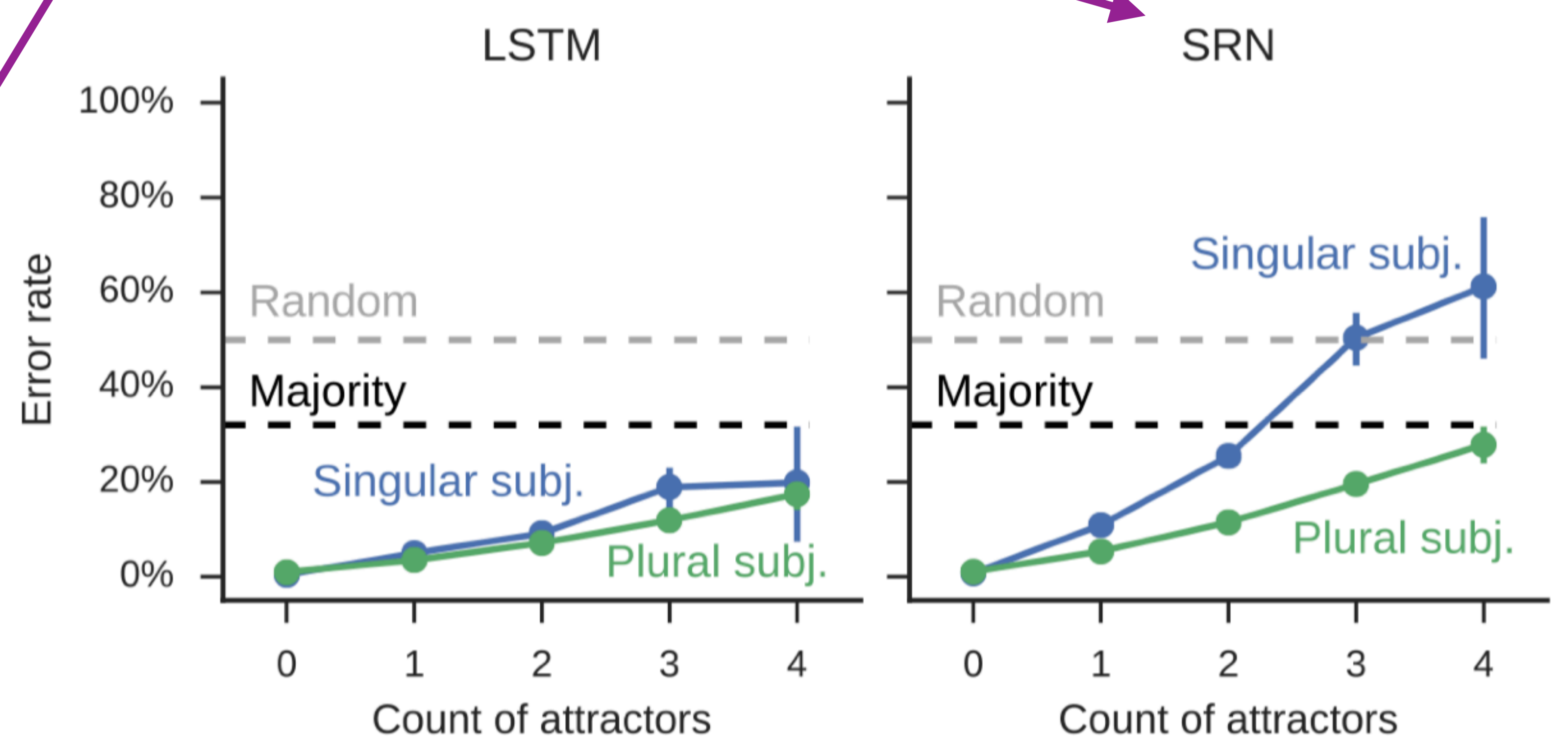
$$P(\texttt{is}|\text{Context}) \overset{?}{>} P(\texttt{are}|\text{Context})$$




```
 The key to the cabinets | is on the table.
*The key to the cabinets | are on the table.
```



RNN models trained solely for the verb inflection task



(Linzen et al., 2016)

Psycholinguistics of agreement production

- Subject–verb agreement is one of the most common "errorful" human language production behaviors in English

After Deadline

New York Times Blog

Ugly Disagreements

BY PHILIP B. CORBETT MARCH 8, 2016 8:00 AM 49

…

__The push__ by Mr. Xi's to assert state control over the markets and the economy __go__ against the philosophy of China's early reformers under Deng Xiaoping, the paramount leader who sought to give more space to the market.

As frequently happens, the phrase between the subject and verb threw us off. Most often, as here, a singular subject is followed by a plural noun, and we are misled into using a plural verb. Make it "the push … goes." (Also, there's no need for the possessive "Xi's.")

…

Her __criticism__ of "medieval" punishments in Saudi Arabia and of Israeli violence against Palestinians __have led__ to diplomatic breaches — and __have prompted__ Ms. Wallstrom to be compared to Mr. Palme.

And yet again! Despite the intervening phrases, the subject is "criticism," so make it "has led" and "has prompted."

Psycholinguistics of subject–verb agreement

- In human language production, agreement performance typically studied with a *preamble completion task:*

```
The key to the cabinets...
```

(Bock & Miller, 1991; Eberhard et al., 2005; Franck et al., 2002; Wagers et al., 2009; many others)

Psycholinguistics of subject–verb agreement

- In human language production, agreement performance typically studied with a *preamble completion task:*

(Bock & Miller, 1991; Eberhard et al., 2005; Franck et al., 2002; Wagers et al., 2009; many others)

Psycholinguistics of subject–verb agreement

- In human language production, agreement performance typically studied with a *preamble completion task:*

The key to the cabinets was on the table.

(Bock & Miller, 1991; Eberhard et al., 2005; Franck et al., 2002; Wagers et al., 2009; many others)

Psycholinguistics of subject–verb agreement

- In human language production, agreement performance typically studied with a *preamble completion task:*

✓ The key to the cabinets was on the table.

(Bock & Miller, 1991; Eberhard et al., 2005; Franck et al., 2002; Wagers et al., 2009; many others)

Psycholinguistics of subject–verb agreement

- In human language production, agreement performance typically studied with a *preamble completion task:*

✓ The key to the cabinets was on the table.

The key to the cabinets doesn't work anymore.

(*Bock & Miller, 1991; Eberhard et al., 2005; Franck et al., 2002; Wagers et al., 2009; many others*)

Psycholinguistics of subject–verb agreement

- In human language production, agreement performance typically studied with a *preamble completion task:*

✓ The key to the cabinets was on the table.

✓ The key to the cabinets doesn't work anymore.

(Bock & Miller, 1991; Eberhard et al., 2005; Franck et al., 2002; Wagers et al., 2009; many others)

Psycholinguistics of subject–verb agreement

- In human language production, agreement performance typically studied with a *preamble completion task:*

✓ The key to the cabinets was on the table.

✓ The key to the cabinets doesn't work anymore.

The key to the cabinets are rusty.

(Bock & Miller, 1991; Eberhard et al., 2005; Franck et al., 2002; Wagers et al., 2009; many others)

Psycholinguistics of subject–verb agreement

- In human language production, agreement performance typically studied with a *preamble completion task:*

✓ The key to the cabinets was on the table.

✓ The key to the cabinets doesn't work anymore.

✗ The key to the cabinets are rusty.

(Bock & Miller, 1991; Eberhard et al., 2005; Franck et al., 2002; Wagers et al., 2009; many others)

Psycholinguistics of subject–verb agreement

- In human language production, agreement performance typically studied with a *preamble completion task:*

✓ The key to the cabinets was on the table.

✓ The key to the cabinets doesn't work anymore.

✗ The key to the cabinets are rusty.

The key to the cabinets would be really helpful to have right now!

(Bock & Miller, 1991; Eberhard et al., 2005; Franck et al., 2002; Wagers et al., 2009; many others)

Psycholinguistics of subject–verb agreement

- In human language production, agreement performance typically studied with a *preamble completion task:*

✓ The key to the cabinets was on the table.

✓ The key to the cabinets doesn't work anymore.

✗ The key to the cabinets are rusty.

n/a The key to the cabinets would be really helpful to have right now!

(Bock & Miller, 1991; Eberhard et al., 2005; Franck et al., 2002; Wagers et al., 2009; many others)

Psycholinguistics of subject–verb agreement

- In human language production, agreement performance typically studied with a *preamble completion task:*

✓ The key to the cabinets was on the table.

✓ The key to the cabinets doesn't work anymore.

✗ The key to the cabinets are rusty.

n/a The key to the cabinets would be really helpful to have right now!

The keys to the cabinets are in my pocket.

(Bock & Miller, 1991; Eberhard et al., 2005; Franck et al., 2002; Wagers et al., 2009; many others)

Psycholinguistics of subject–verb agreement

- In human language production, agreement performance typically studied with a *preamble completion task:*

✓ The key to the cabinets was on the table.

✓ The key to the cabinets doesn't work anymore.

✗ The key to the cabinets are rusty.

n/a The key to the cabinets would be really helpful to have right now!

n/a The keys to the cabinets are in my pocket.

(Bock & Miller, 1991; Eberhard et al., 2005; Franck et al., 2002; Wagers et al., 2009; many others)

Psycholinguistics of subject–verb agreement

- In human language production, agreement performance typically studied with a *preamble completion task:*

✓ The key to the cabinets was on the table.

✓ The key to the cabinets doesn't work anymore.

✗ The key to the cabinets are rusty.

n/a The key to the cabinets would be really helpful to have right now!

n/a The keys to the cabinets are in my pocket.

$\frac{1}{3} = 33\,\%$ error rate

(Bock & Miller, 1991; Eberhard et al., 2005; Franck et al., 2002; Wagers et al., 2009; many others)

Psycholinguistics of subject–verb agreement

- In human language production, agreement performance typically studied with a *preamble completion task:*

✓ The key to the cabinets was on the table.

✓ The key to the cabinets doesn't work anymore.

✗ The key to the cabinets are rusty.

n/a The key to the cabinets would be really helpful to have right now!

n/a The keys to the cabinets are in my pocket.

$\frac{1}{3} = 33\,\%$ error rate

sg	**sg**	`The key to the cabinet...`
sg	**pl**	`The key to the cabinets...`
pl	**sg**	`The keys to the cabinet...`
pl	**pl**	`The keys to the cabinets...`

(Bock & Miller, 1991; Eberhard et al., 2005; Franck et al., 2002; Wagers et al., 2009; many others)

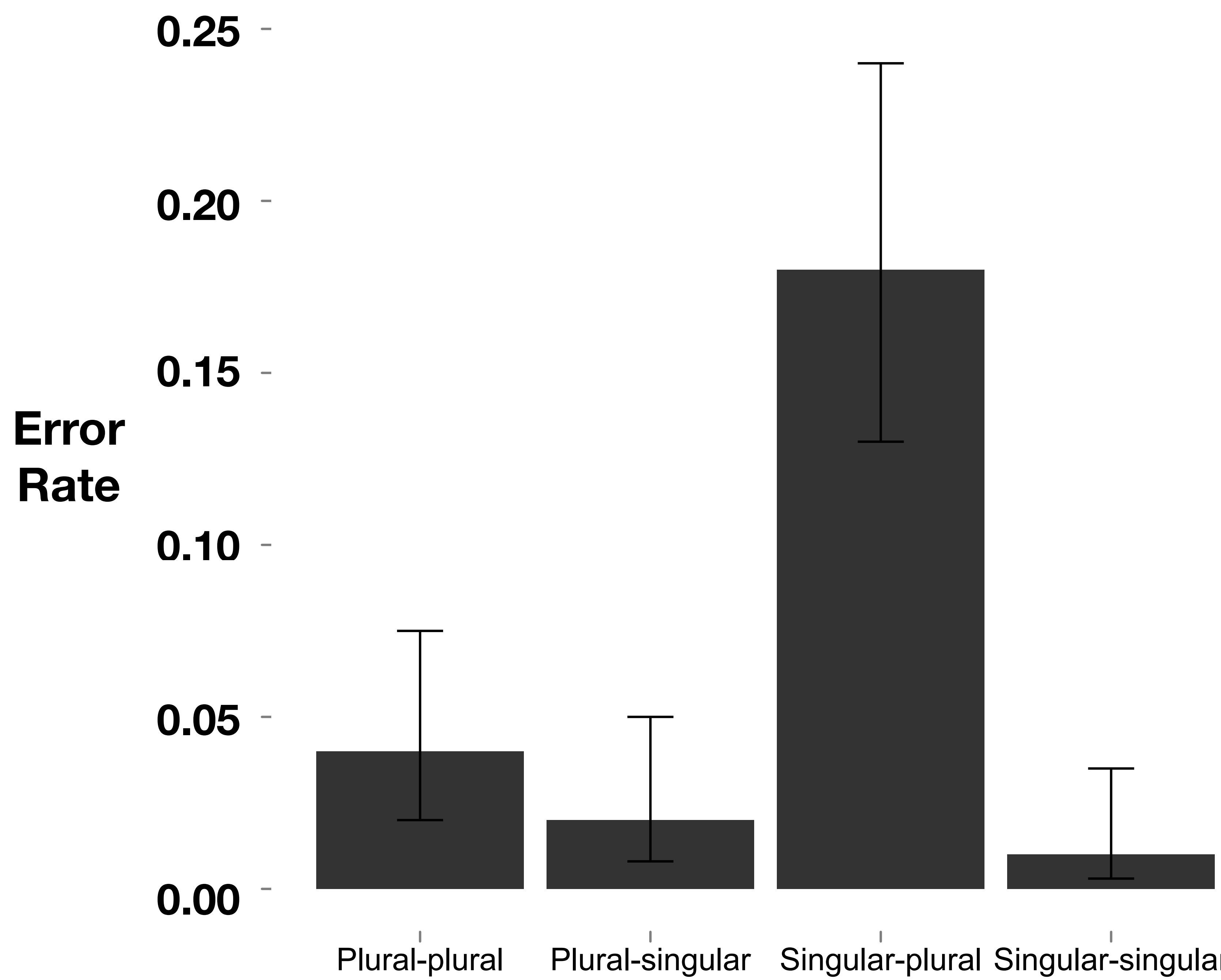
Production error rates

sg sg `The key to the cabinet...`

sg pl `The key to the cabinets...`

pl sg `The keys to the cabinet...`

pl pl `The keys to the cabinets...`



(Bock & Miller, 1991; data from Bergen & Gibson, unpublished)

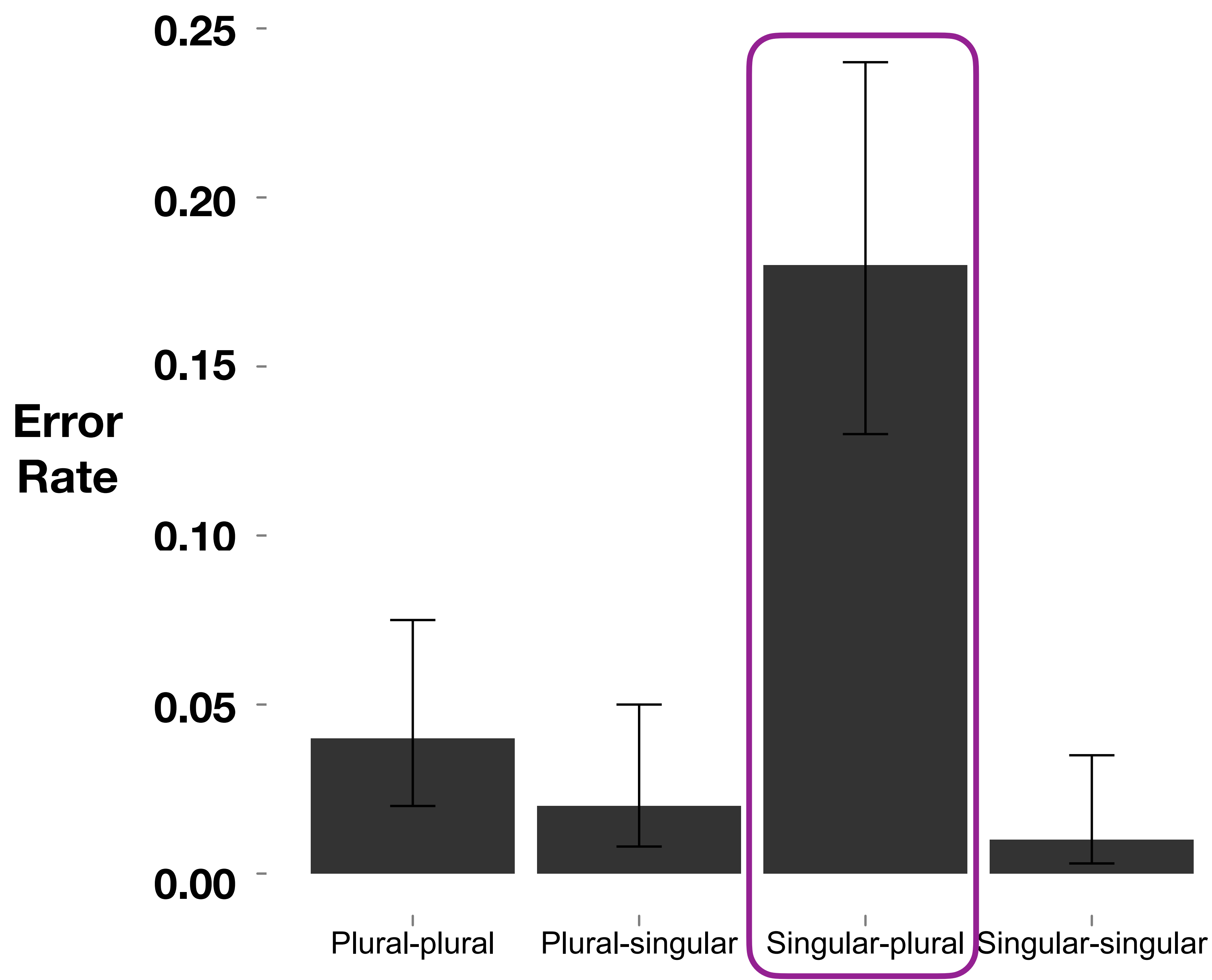
Production error rates

sg sg `The key to the cabinet...`

sg pl `The key to the cabinets...`

pl sg `The keys to the cabinet...`

pl pl `The keys to the cabinets...`



(Bock & Miller, 1991; data from Bergen & Gibson, unpublished)

Comprehension of agreement errors

The musician (SG) who the reviewer praise so highly will probably win ...

(Wagers et al., 2009)

Comprehension of agreement errors

The musician (SG) / musicians (PL) who the reviewer praise so highly will probably win ...

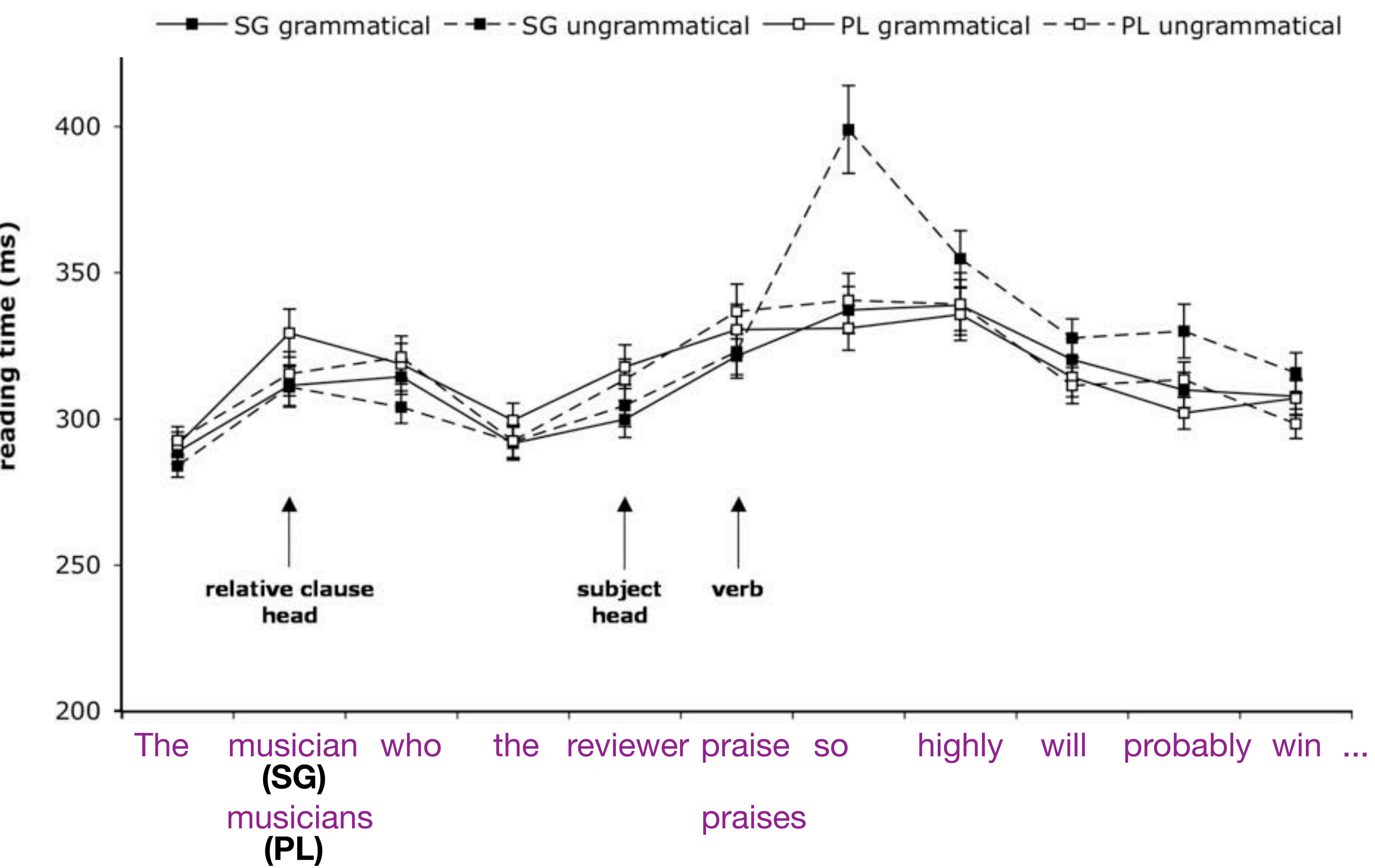
(*Wagers et al., 2009*)

Comprehension of agreement errors

The	musician **(SG)**	who	the	reviewer	praise	so	highly	will	probably	win	...
	musicians **(PL)**				praises						

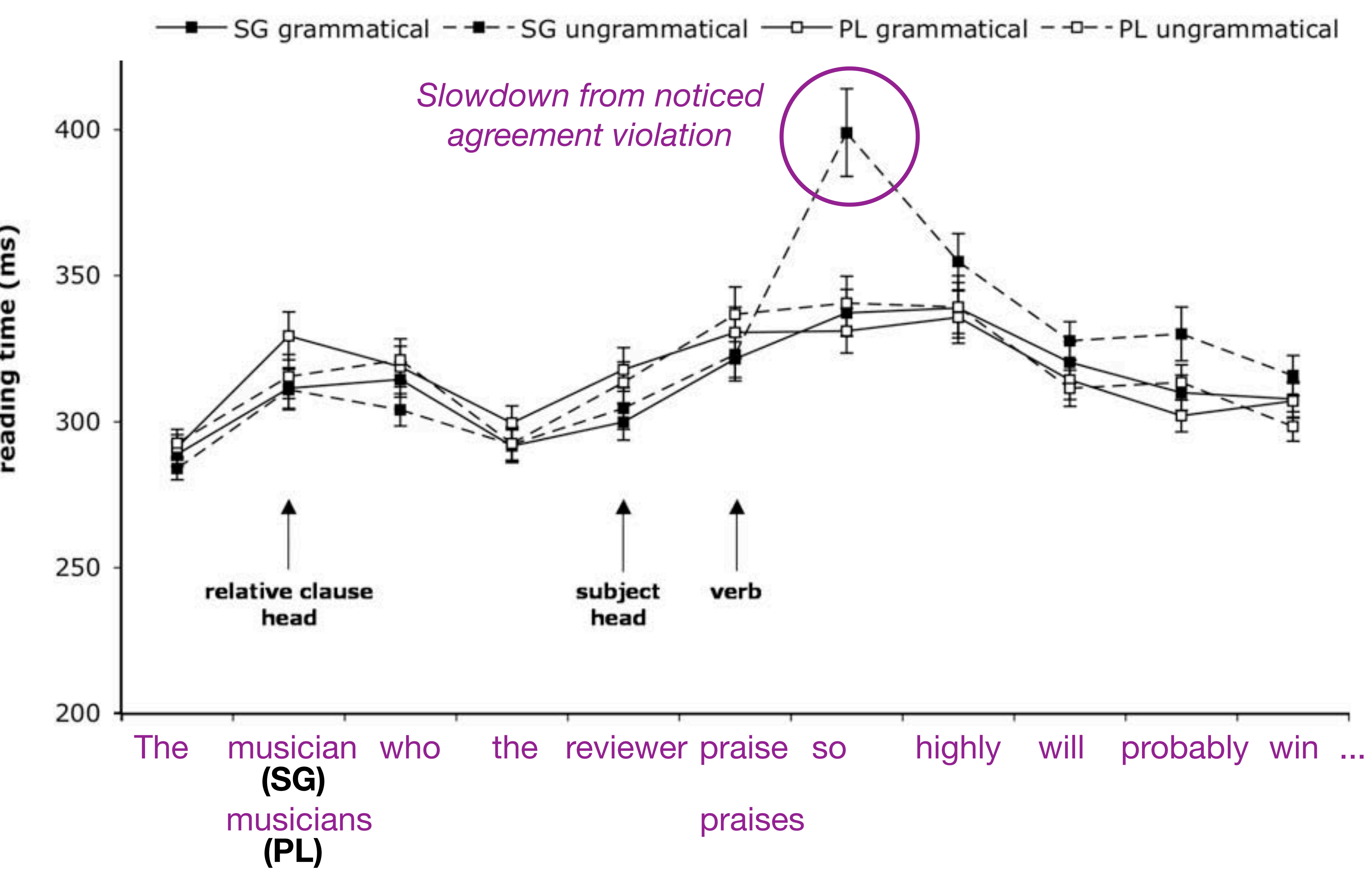
(Wagers et al., 2009)

Comprehension of agreement errors



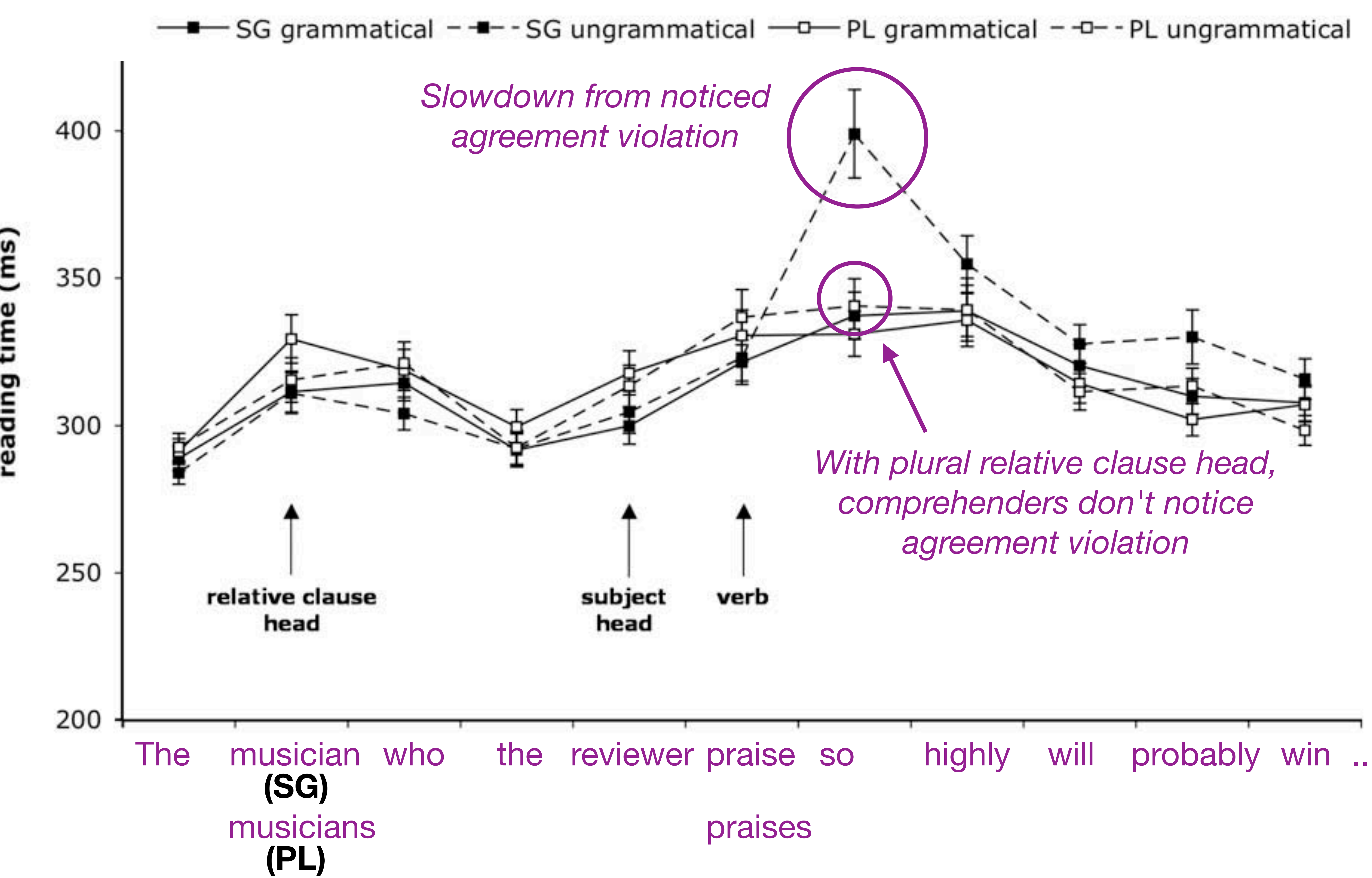
(Wagers et al., 2009)

Comprehension of agreement errors



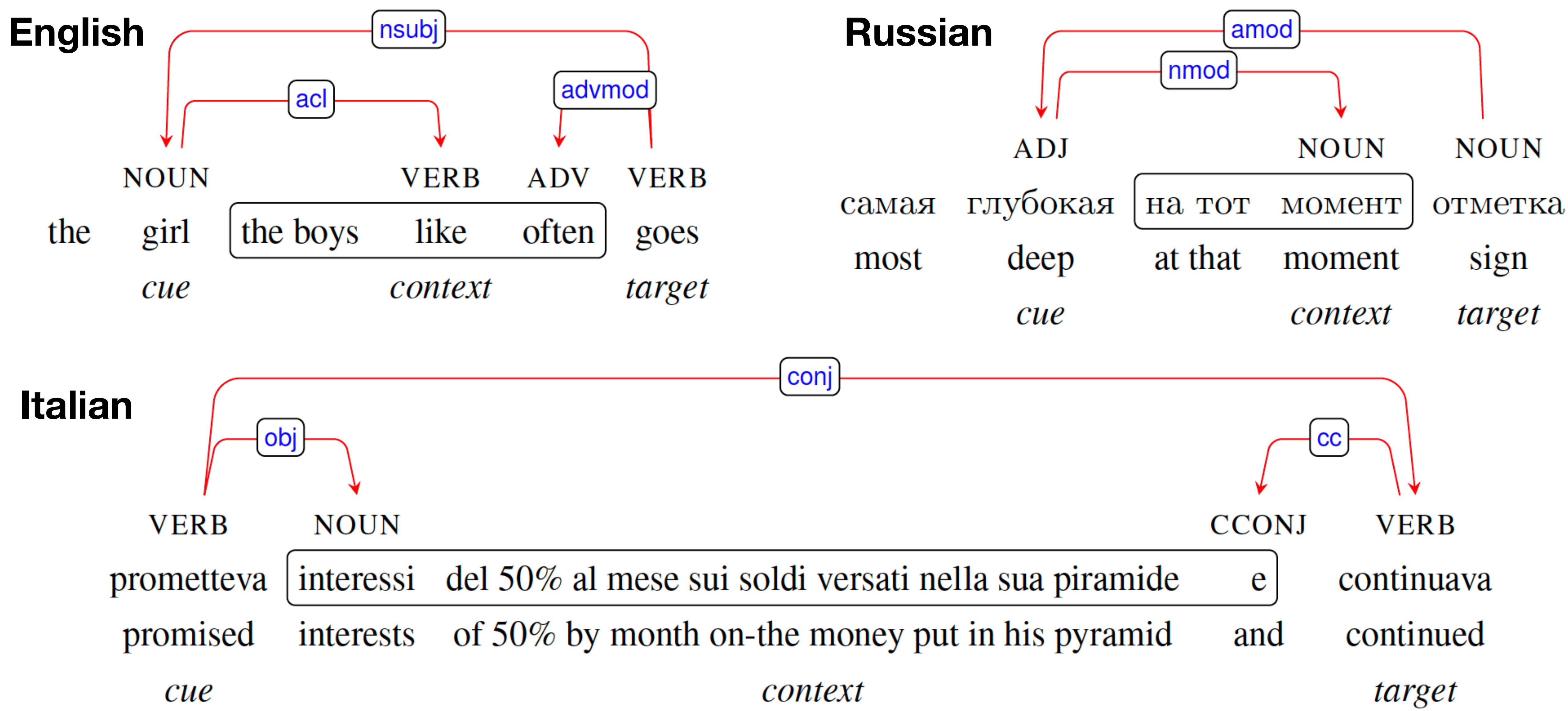
(Wagers et al., 2009)

Comprehension of agreement errors



(Wagers et al., 2009)

Corpus-based long-distance agreement benchmark



Hebrew (not shown)

(Gulordava et al., 2018)

Real and "artificial" examples

Original to "colorless green" (*nonce*) sentence conversion by part-of-speech-preserving content-word substitution:

It **presents** the **case** for **marriage** **equality** and **states**...

(Gulordava et al., 2018)

Real and "artificial" examples

Original to "colorless green" (*nonce*) sentence conversion by part-of-speech-preserving content-word substitution:

`It` **`presents`** `the` **`case`** `for` **`marriage equality`** `and` **`states`**`...`

↓ ↓ ↓ ↓ ↓

`It` **`stays`** `the` **`shuttle`** `for` **`honesty insurance`** `and` **`finds`**`...`

(Gulordava et al., 2018)

Real and "artificial" examples

Original to "colorless green" (*nonce*) sentence conversion by part-of-speech-preserving content-word substitution:

It **presents** the **case** for **marriage** **equality** and **states**...

It **stays** the **shuttle** for **honesty** **insurance** and **finds**...

Original

```
I guess the questions on which I should like your steer      is|are
The main thing people do there                               is|are
The first thing you notice when you arrive on location is
that the waiting line literally goes out the door and        spill|spills
Please let me know if you have any questions or              need|needs
```

"Colorless green" (*nonce*)

```
the ecological wines we believe in our night                 assume|assumes
You assume the picketing and                                 seek|seeks
The sauce to recruit in pastor                               advise|advises
Can you expire that the ground we are happening              produce|produces
```

(Gulordava et al., 2018)

Real and "artificial" examples

Original to "colorless green" (*nonce*) sentence conversion by part-of-speech-preserving content-word substitution:

`It` **`presents`** `the` **`case`** `for` **`marriage equality`** `and` **`states`**`...`

↓ ↓ ↓ ↓ ↓

`It` **`stays`** `the` **`shuttle`** `for` **`honesty insurance`** `and` **`finds`**`...`

Original

`I guess the questions on which I should like your steer`	`is|are`
`The main thing people do there`	`is|are`
`The first thing you notice when you arrive on location is that the waiting line literally goes out the door and`	`spill|spills`
`Please let me know if you have any questions or`	`need|needs`

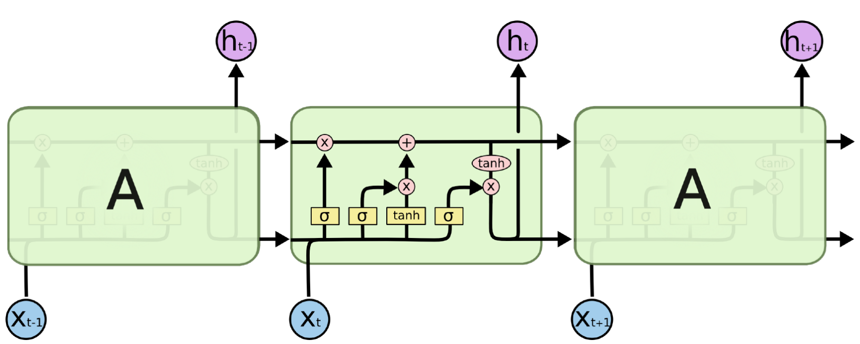
"Colorless green" (*nonce*)

`the ecological wines we believe in our night`	`assume|assumes`
`You assume the picketing and`	`seek|seeks`
`The sauce to recruit in pastor`	`advise|advises`
`Can you expire that the ground we are happening`	`produce|produces`

(Gulordava et al., 2018)

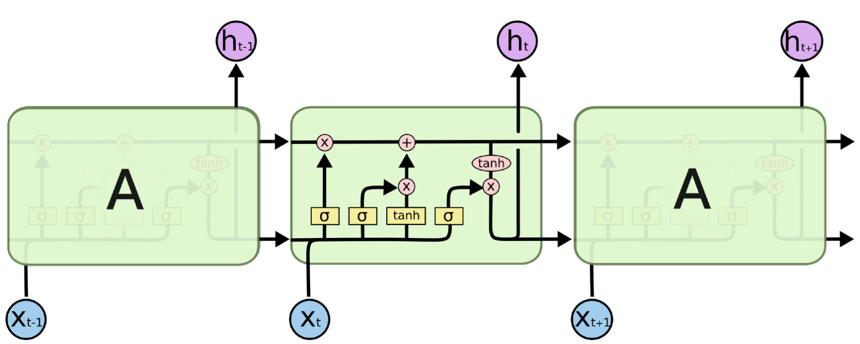
Human and LSTM performance

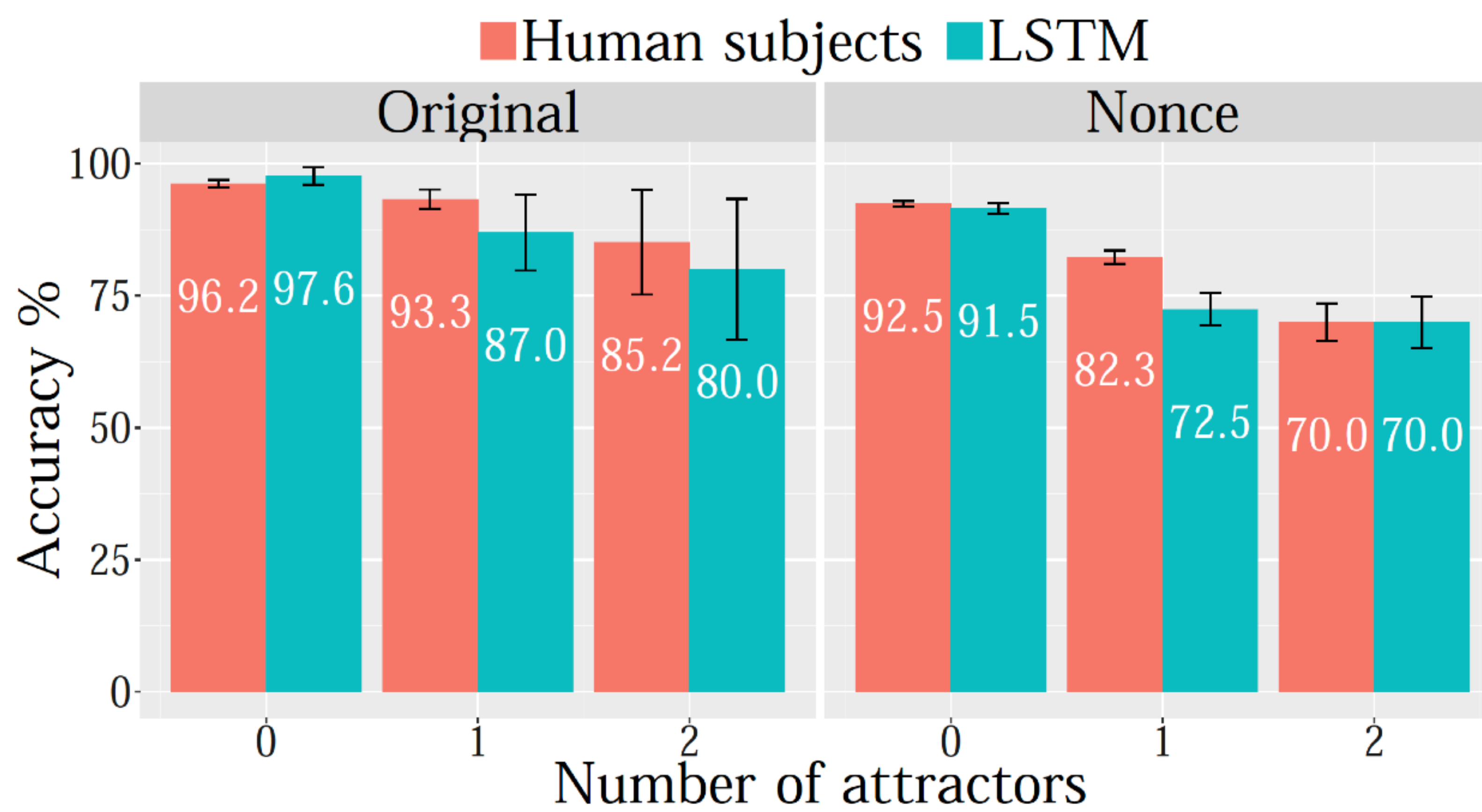
90m words per language



Human and LSTM performance

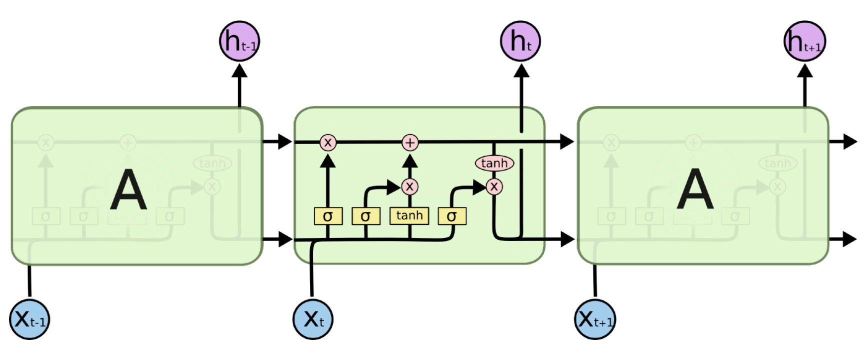
90m words per language

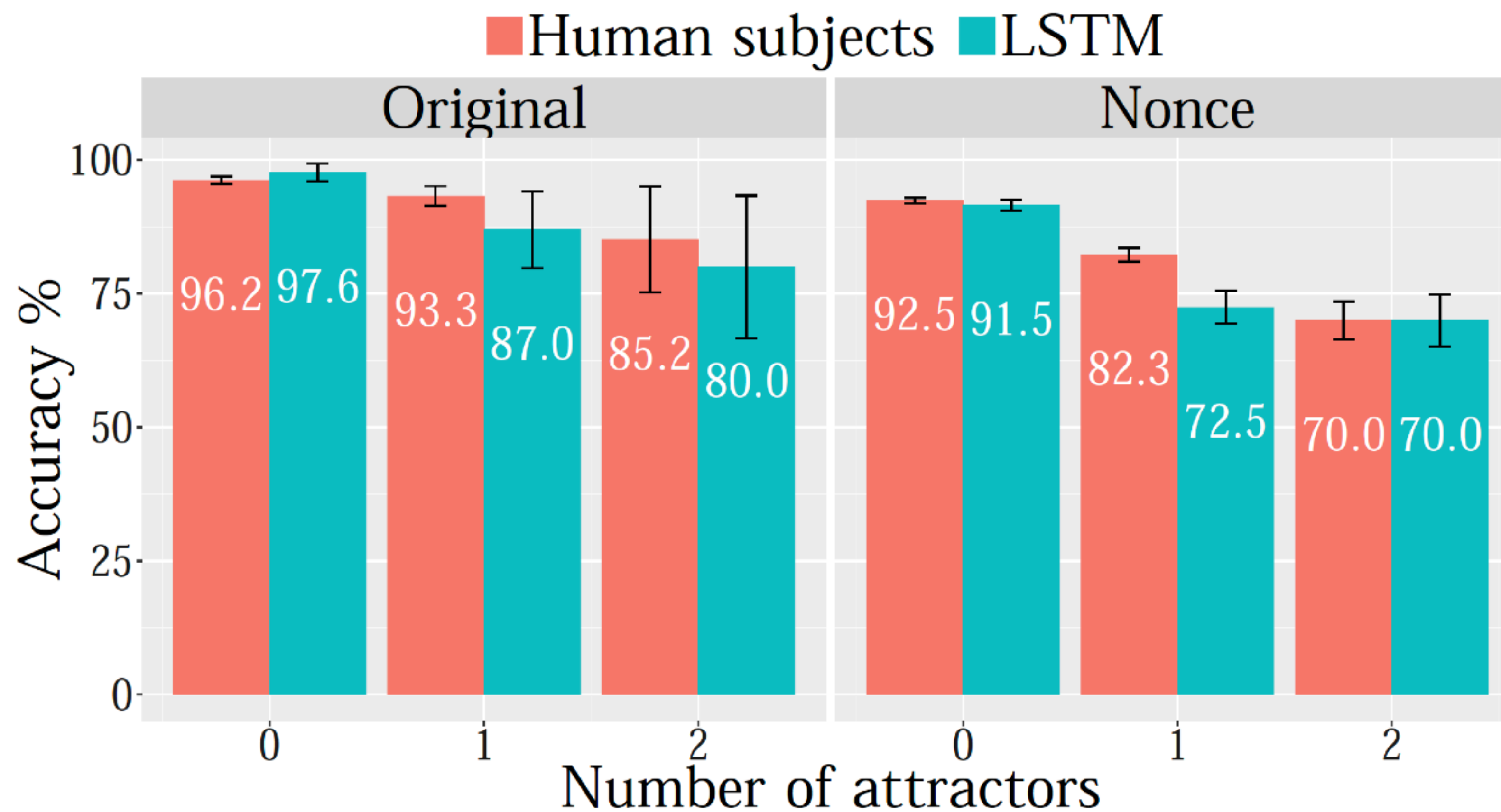




Human and LSTM performance

90m words per language





Construction	#original	Original Subjects	Original LSTM	Nonce Subjects	Nonce LSTM
DET [AdjP] NOUN	14	98.7	$98.6_{\pm 3.2}$	98.1	$91.7_{\pm 0.4}$
NOUN [RelC / PartP] clitic VERB	6	93.1	$100_{\pm 0.0}$	95.4	$97.8_{\pm 0.8}$
NOUN [RelC / PartP] VERB	27	97.0	$93.3_{\pm 4.1}$	92.3	$92.5_{\pm 2.1}$
ADJ [conjoined ADJs] ADJ	13	98.5	$100_{\pm 0.0}$	98.0	$98.1_{\pm 1.1}$
NOUN [AdjP] relpron VERB	10	95.9	$98.0_{\pm 4.5}$	89.5	$84.0_{\pm 3.3}$
NOUN [PP] ADVERB ADJ	13	91.5	$98.5_{\pm 3.4}$	79.4	$76.9_{\pm 1.4}$
NOUN [PP] VERB (participial)	18	87.1	$77.8_{\pm 3.9}$	73.4	$71.1_{\pm 3.3}$
VERB [NP] CONJ VERB	18	94.0	$83.3_{\pm 10.4}$	86.8	$78.5_{\pm 1.7}$
(Micro) average		94.5	$92.1_{\pm 1.6}$	88.4	$85.5_{\pm 0.7}$

Table 3: Subject and LSTM accuracy on the Italian test set, by construction and averaged.

Targeted syntactic evaluation

(Marvin & Linzen, 2018)

Targeted syntactic evaluation

- Frequency *in naturalistic usage* is important for understanding our linguistic environment & for applications

(Marvin & Linzen, 2018)

Targeted syntactic evaluation

- Frequency *in naturalistic usage* is important for understanding our linguistic environment & for applications
- But targeted evaluation, where *we choose the stimuli*, may reveal more to us about a system

(Marvin & Linzen, 2018)

Targeted syntactic evaluation

- Frequency *in naturalistic usage* is important for understanding our linguistic environment & for applications
- But targeted evaluation, where *we choose the stimuli*, may reveal more to us about a system

Across PP	`The farmer near the parents` **`smiles`**
Across Subject RC	`The officers that love the skater` **`smile`**
Across Coordination	`The senator smiles and` **`laughs`**
Inside Object RC	`The farmer that the parents` **`love`** `swims`
⋮	⋮

(Marvin & Linzen, 2018)

Targeted syntactic evaluation

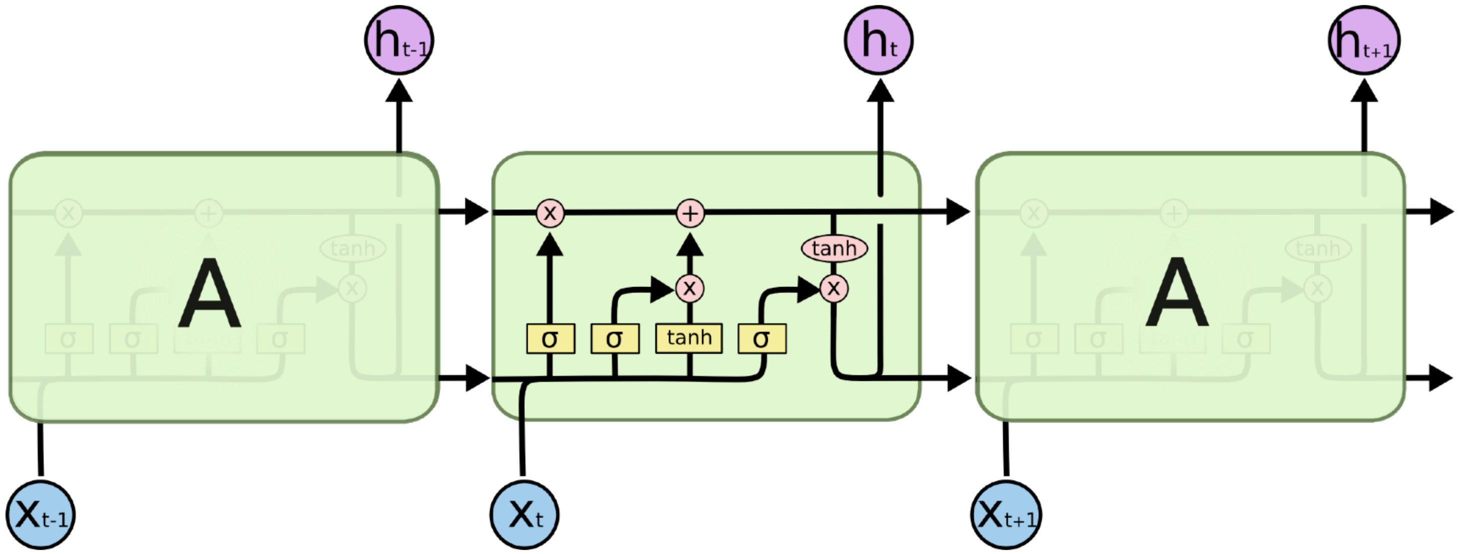
- Frequency *in naturalistic usage* is important for understanding our linguistic environment & for applications
- But targeted evaluation, where *we choose the stimuli*, may reveal more to us about a system

Across PP	`The farmer near the parents` **`smiles`**
Across Subject RC	`The officers that love the skater` **`smile`**
Across Coordination	`The senator smiles and` **`laughs`**
Inside Object RC	`The farmer that the parents` **`love`** `swims`
⋮	⋮

	RNN	*n*-gram	Humans
SUBJECT-VERB AGREEMENT:			
Simple	0.94	0.79	0.96
In a sentential complement	0.99	0.79	0.93
Short VP coordination	0.90	0.51	0.94
Long VP coordination	0.61	0.50	0.82
Across a prepositional phrase	0.57	0.50	0.85
Across a subject relative clause	0.56	0.50	0.88
Across an object relative clause	0.50	0.50	0.85
Across an object relative (no *that*)	0.52	0.50	0.82
In an object relative clause	0.84	0.50	0.78
In an object relative (no *that*)	0.71	0.50	0.79

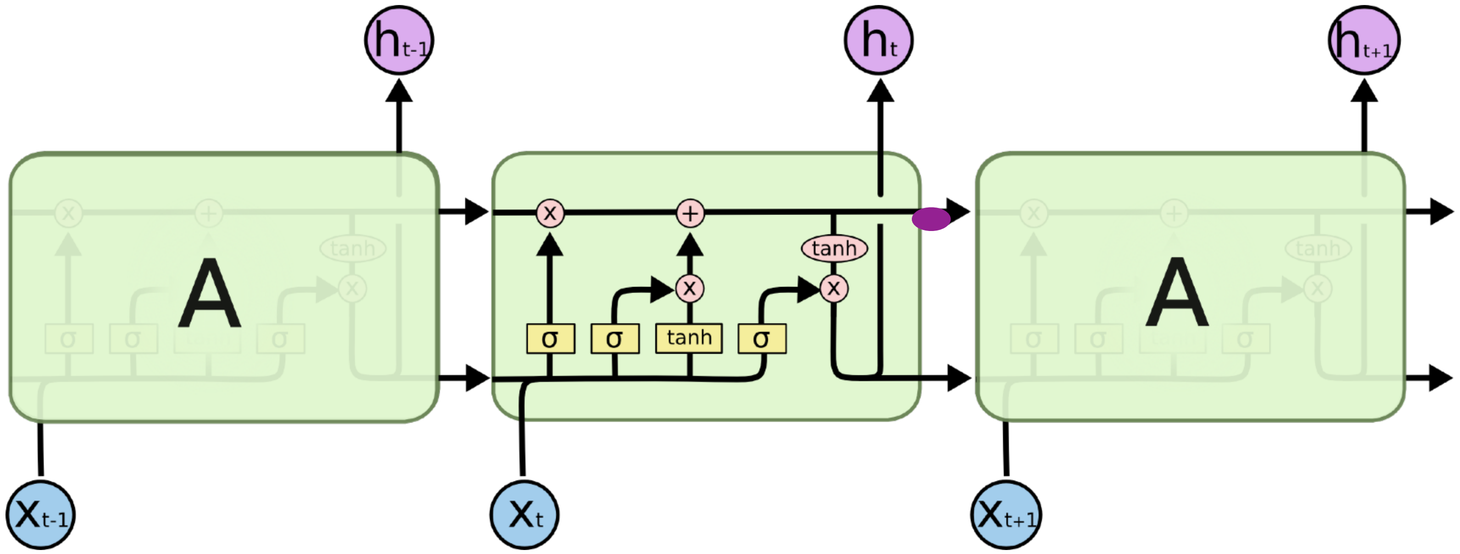
(Marvin & Linzen, 2018)

LSTM learned circuitry for S–V agreement



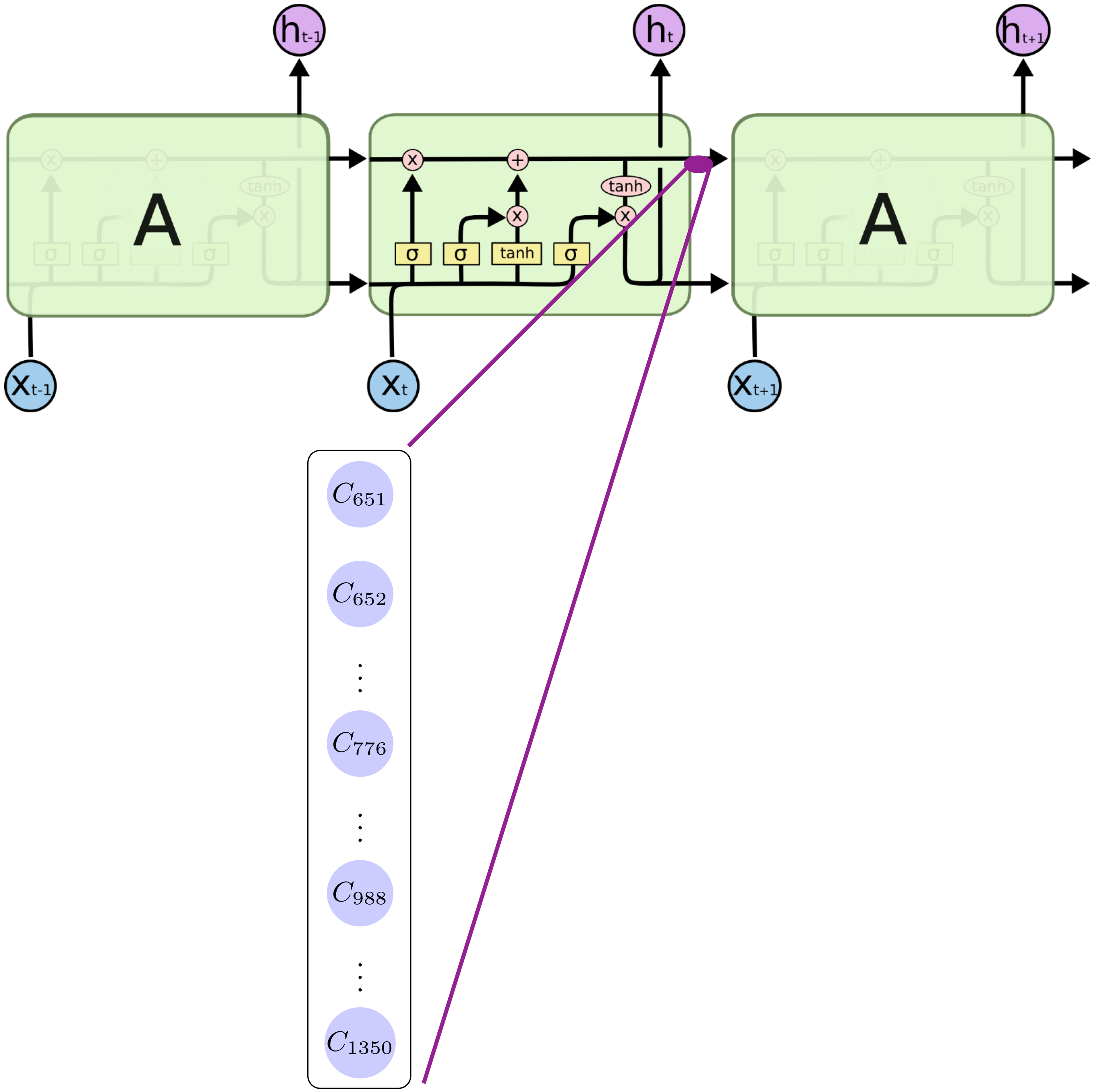
(Lakretz et al., 2019)

LSTM learned circuitry for S–V agreement



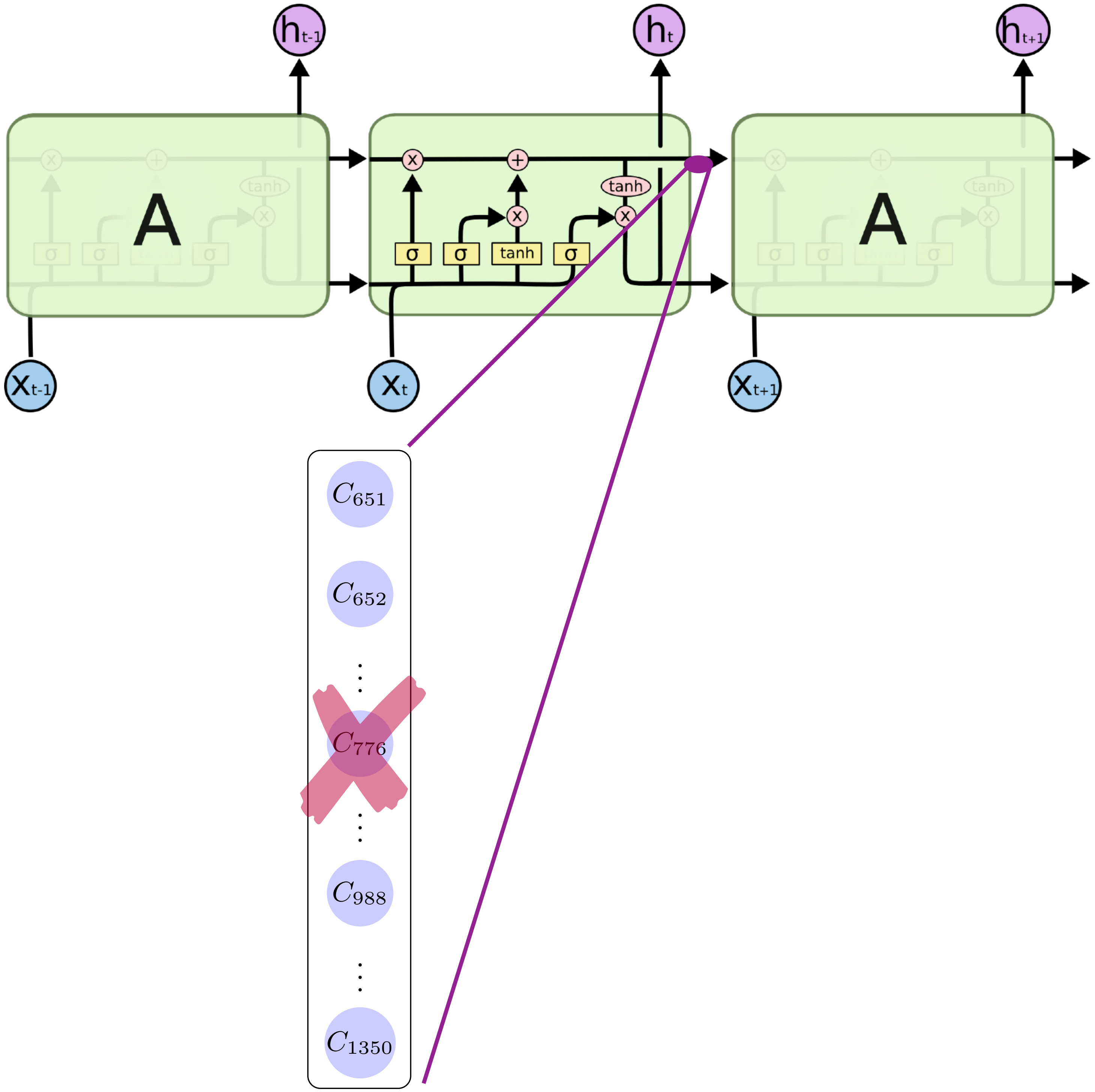
(Lakretz et al., 2019)

LSTM learned circuitry for S–V agreement



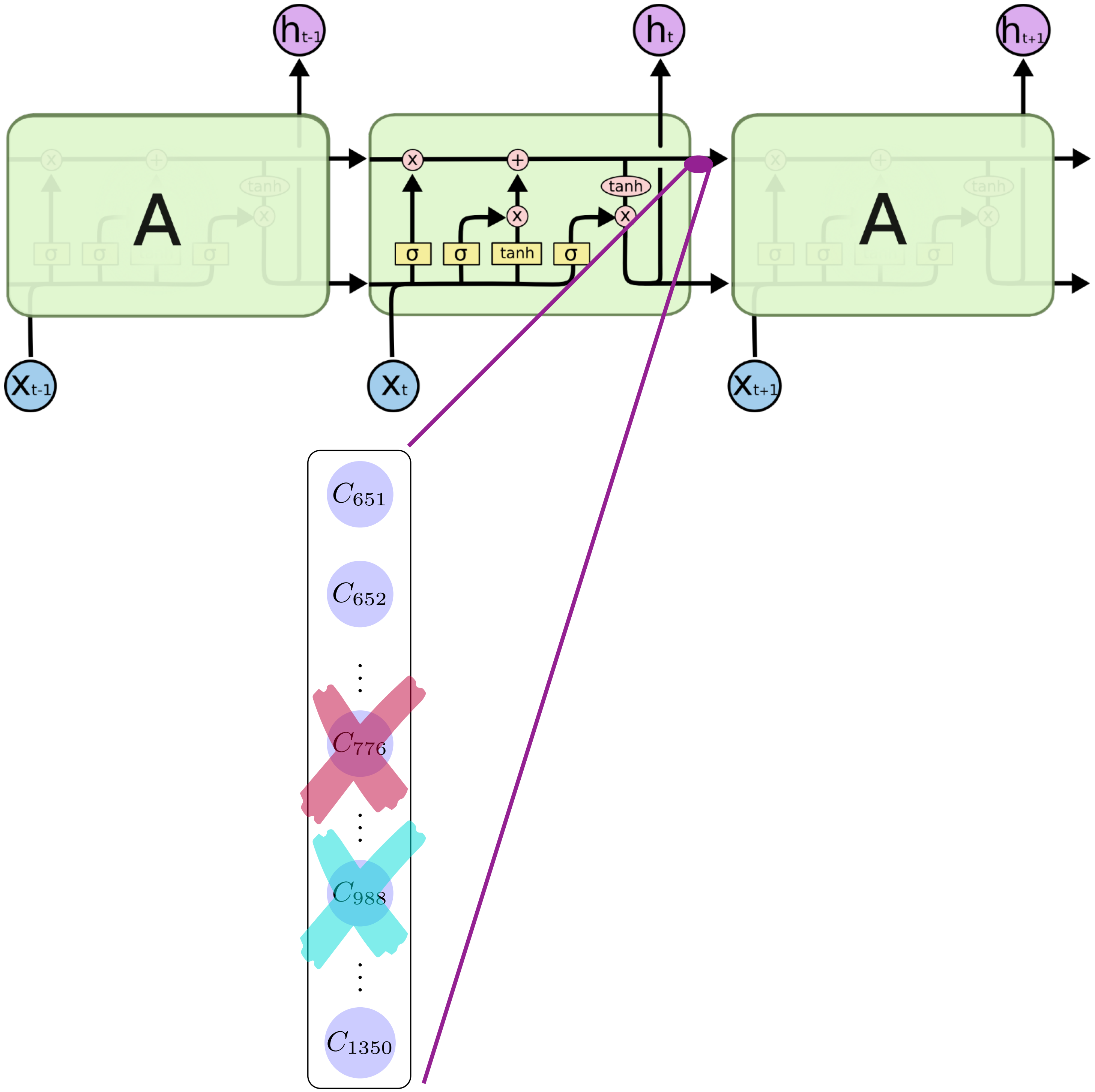
(Lakretz et al., 2019)

LSTM learned circuitry for S–V agreement



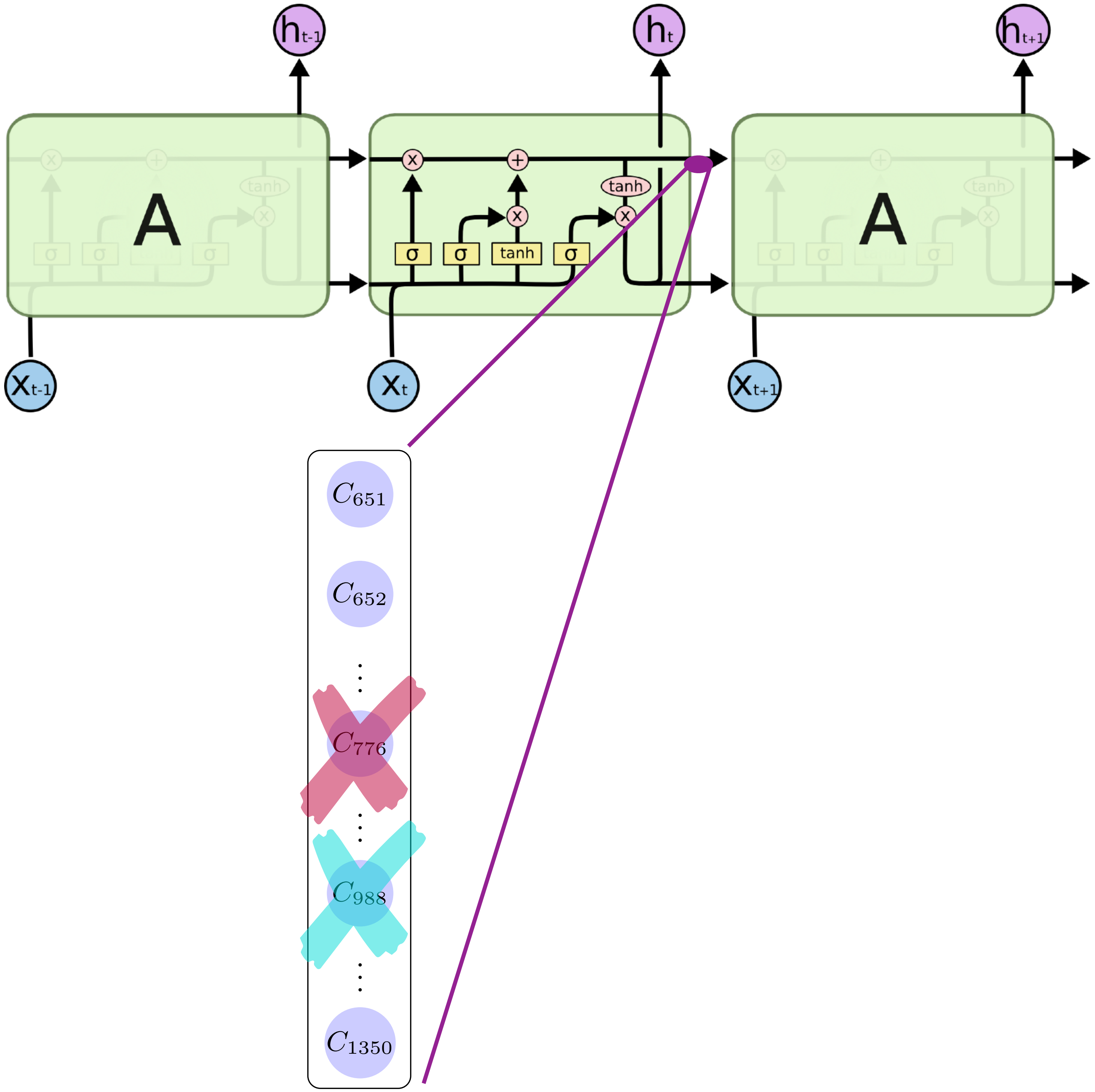
(Lakretz et al., 2019)

LSTM learned circuitry for S–V agreement



(Lakretz et al., 2019)

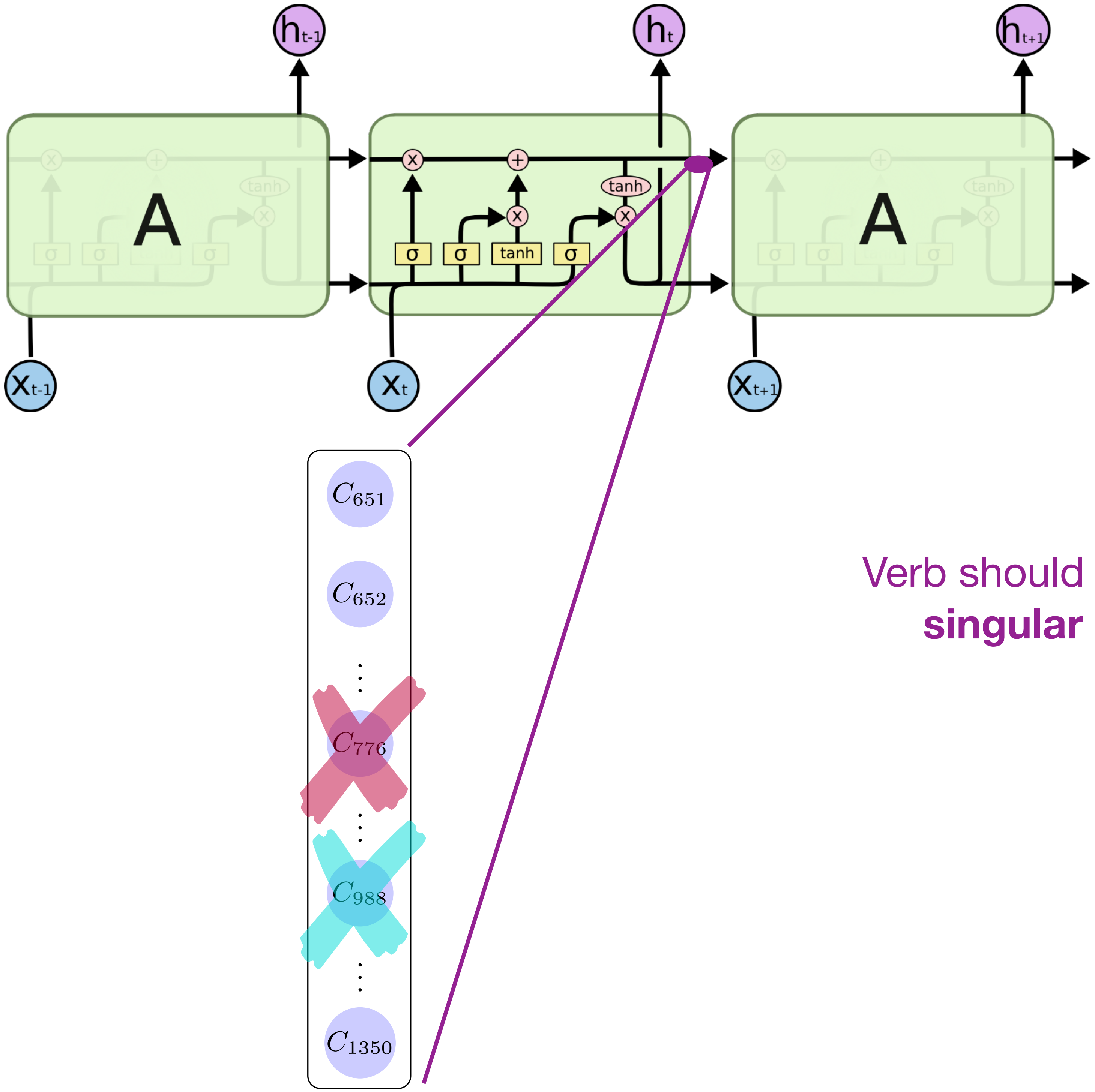
LSTM learned circuitry for S–V agreement



NA task	C	Ablated 776	Ablated 988	Full
Simple	S	-	-	100
Adv	S	-	-	100
2Adv	S	-	-	99.9
CoAdv	S	-	82	98.7
namePP	SS	-	-	99.3
nounPP	SS	-	-	99.2
nounPP	SP	-	54.2	87.2
nounPPAdv	SS	-	-	99.5
nounPPAdv	SP	-	54.0	91.2
Simple	P	-	-	100
Adv	P	-	-	99.6
2Adv	P	-	-	99.3
CoAdv	P	79.2	-	99.3
namePP	PS	39.9	-	68.9
nounPP	PS	48.0	-	92.0
nounPP	PP	78.3	-	99.0
nounPPAdv	PS	63.7	-	99.2
nounPPAdv	PP	-	-	99.8

(Lakretz et al., 2019)

LSTM learned circuitry for S–V agreement

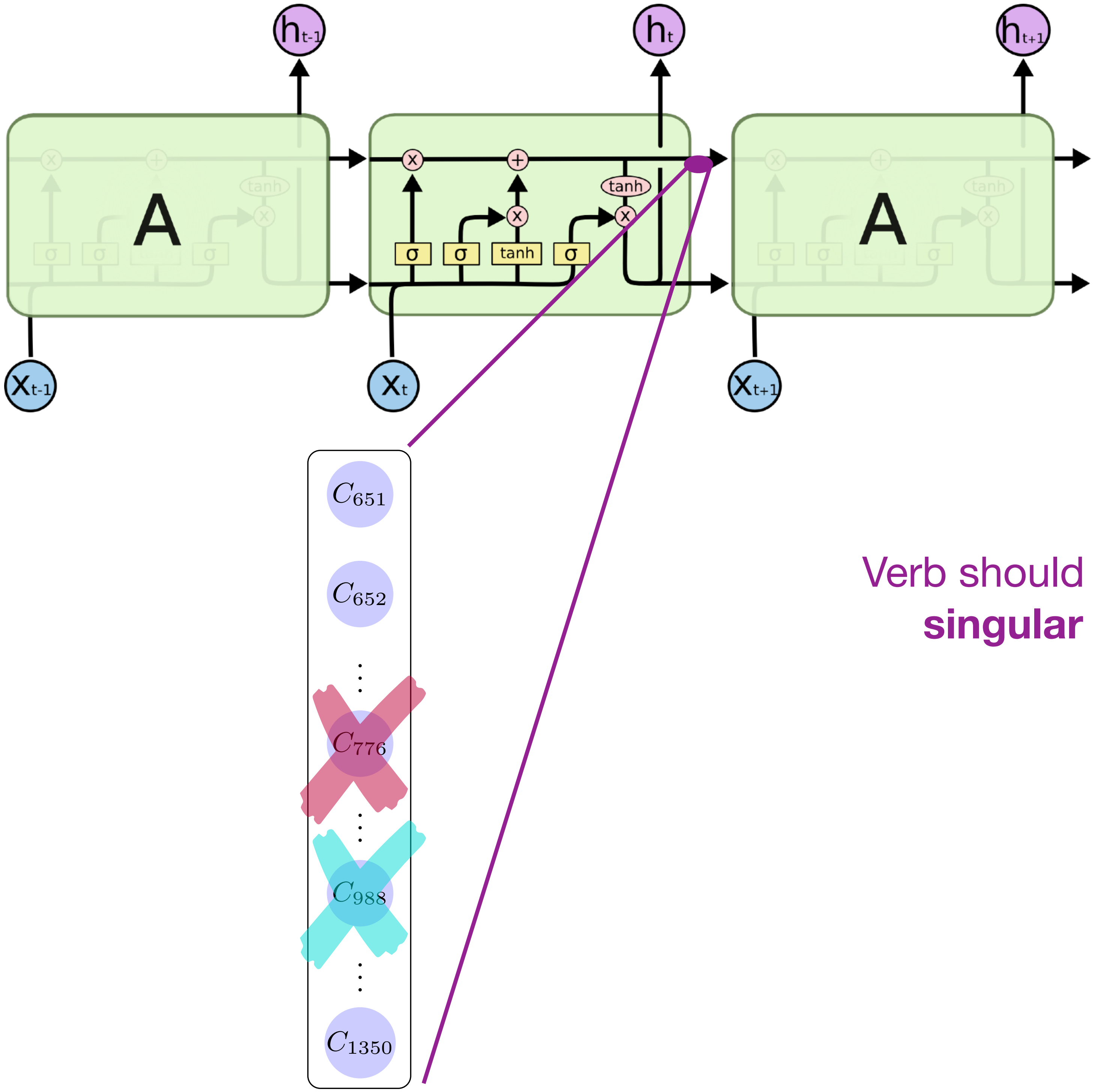


Verb should be **singular**

NA task	C	Ablated 776	Ablated 988	Full
Simple	S	-	-	100
Adv	S	-	-	100
2Adv	S	-	-	99.9
CoAdv	S	-	82	98.7
namePP	SS	-	-	99.3
nounPP	SS	-	-	99.2
nounPP	SP	-	54.2	87.2
nounPPAdv	SS	-	-	99.5
nounPPAdv	SP	-	54.0	91.2
Simple	P	-	-	100
Adv	P	-	-	99.6
2Adv	P	-	-	99.3
CoAdv	P	79.2	-	99.3
namePP	PS	39.9	-	68.9
nounPP	PS	48.0	-	92.0
nounPP	PP	78.3	-	99.0
nounPPAdv	PS	63.7	-	99.2
nounPPAdv	PP	-	-	99.8

(Lakretz et al., 2019)

LSTM learned circuitry for S–V agreement

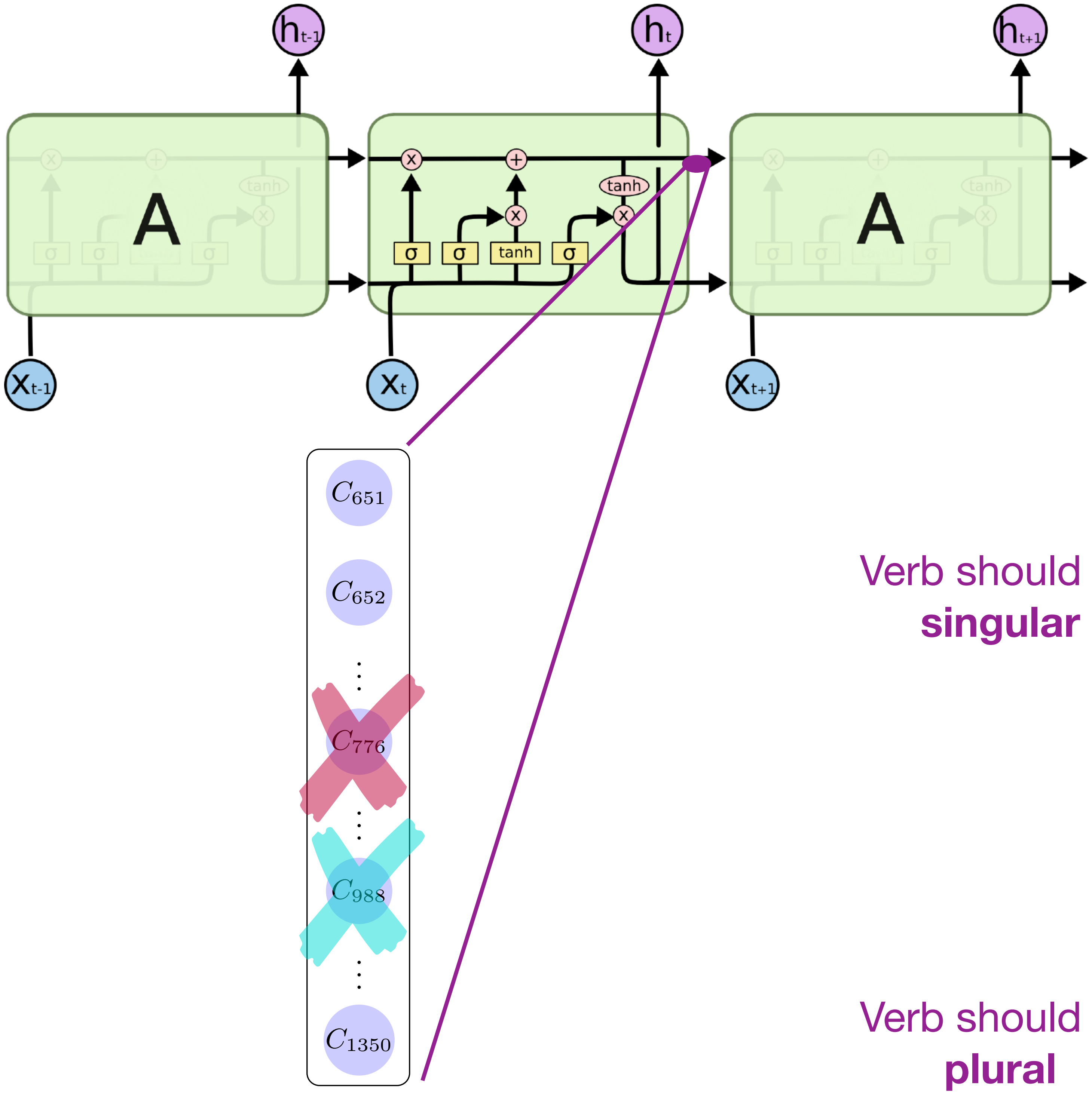


Verb should be
singular

NA task	C	Ablated 776	Ablated 988	Full
Simple	S	-	-	100
Adv	S	-	-	100
2Adv	S	-	-	99.9
CoAdv	S	-	82	98.7
namePP	SS	-	-	99.3
nounPP	SS	-	-	99.2
nounPP	SP	-	54.2	87.2
nounPPAdv	SS	-	-	99.5
nounPPAdv	SP	-	54.0	91.2
Simple	P	-	-	100
Adv	P	-	-	99.6
2Adv	P	-	-	99.3
CoAdv	P	79.2	-	99.3
namePP	PS	39.9	-	68.9
nounPP	PS	48.0	-	92.0
nounPP	PP	78.3	-	99.0
nounPPAdv	PS	63.7	-	99.2
nounPPAdv	PP	-	-	99.8

(Lakretz et al., 2019)

LSTM learned circuitry for S–V agreement



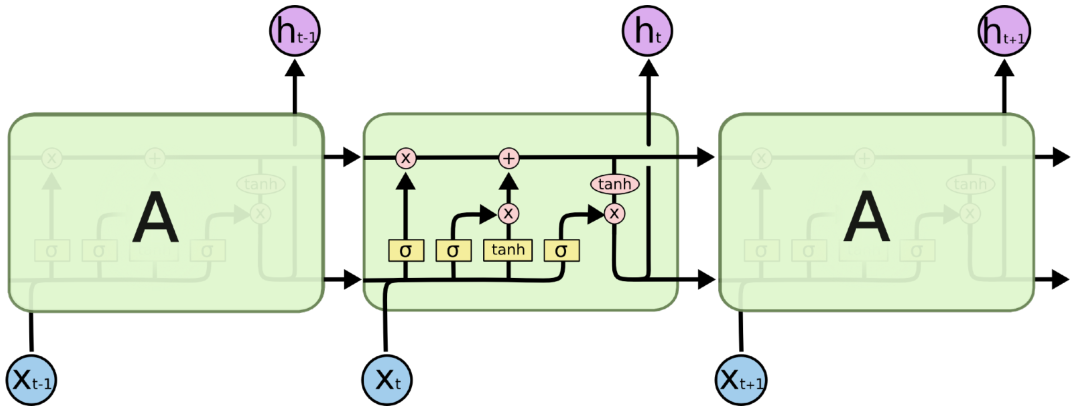
Verb should be **singular**

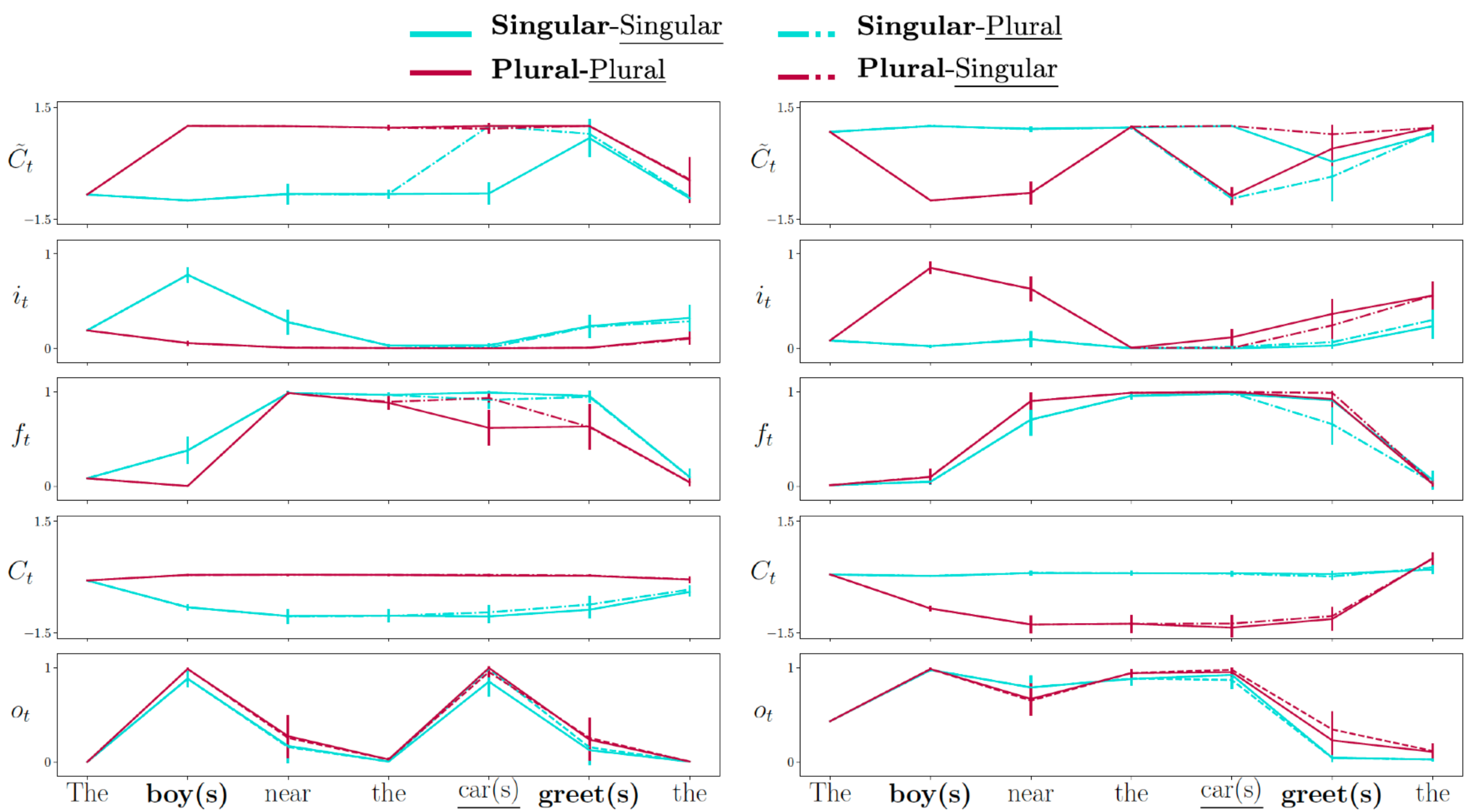
Verb should be **plural**

NA task	C	Ablated 776	Ablated 988	Full
Simple	S	-	-	100
Adv	S	-	-	100
2Adv	S	-	-	99.9
CoAdv	S	-	82	98.7
namePP	SS	-	-	99.3
nounPP	SS	-	-	99.2
nounPP	SP	-	54.2	87.2
nounPPAdv	SS	-	-	99.5
nounPPAdv	SP	-	54.0	91.2
Simple	P	-	-	100
Adv	P	-	-	99.6
2Adv	P	-	-	99.3
CoAdv	P	79.2	-	99.3
namePP	PS	39.9	-	68.9
nounPP	PS	48.0	-	92.0
nounPP	PP	78.3	-	99.0
nounPPAdv	PS	63.7	-	99.2
nounPPAdv	PP	-	-	99.8

(Lakretz et al., 2019)

LSTM learned circuitry for S–V agreement





(a) 988 (singular)

(b) 776 (plural)

What we covered today

- Impressionistic assessment of language model text
- Perplexity-based evaluation
- Targeted grammatical evaluation: subject–verb agreement
  - Left-to-right prediction paradigm
  - Psycholinguistics of subject–verb agreeement
  - Evaluation on "colorless green" (*nonce*) sentences
  - Controlled stimuli
  - Ablation tests to reveal circuit-level processing in models

Next time: more targeted evaluation of neural language models' grammatical capabilities, and implications for learnability of natural language grammar

References

Bock, K., & Miller, C. A. (1991). Broken agreement. Cognitive Psychology, 23, 45–93. Chelba, C., Mikolov, T., Schuster, M., Ge, Q., Brants, T., Koehn, P., & Robinson, T. (2013). One billion word benchmark for measuring progress in statistical language modeling.

Eberhard, K. M., Cutting, J. C., & Bock, K. (2005). Making syntax of sense: Number agreement in sentence production. Psychological Review, 112(3), 531–559.

Franck, J., Vigliocco, G., & Nicol, J. (2002). Subject-verb agreement errors in French and English: The role of syntactic hierarchy. Language & Cognitive Processes, 17 (4), 371–404.

Futrell, R., Wilcox, E., Morita, T., Qian, P., Ballesteros, M., & Levy, R. (2019). Neural language models as psycholinguistic subjects: Representations of syntactic state. In Proceedings of the 18th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 32–42).

Gulordava, K., Bojanowski, P., Grave, E., Linzen, T., & Baroni, M. (2018). Colorless green recurrent networks dream hierarchically. In Proceedings of the 17th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Hu- man Language Technologies (pp. 1195–1205). New Orleans, Louisiana: Association for Computational Linguistics.

Jozefowicz, R., Vinyals, O., Schuster, M., Shazeer, N., & Wu, Y. (2016). Exploring the limits of language modeling. arXiv preprint arXiv:1602.02410.

Lakretz, Y., Kruszewski, G., Desbordes, T., Hupkes, D., Dehaene, S., & Baroni, M. (2019). The emergence of number and syntax units in LSTM language models. In Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers) (pp. 11– 20).

Linzen, T., Dupoux, E., & Goldberg, Y. (2016). Assessing the ability of LSTMs to learn syntax-sensitive dependencies. Transactions of the Association for Computational Linguistics.

Marvin, R., & Linzen, T. (2018). Targeted syntactic evaluation of language models. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (pp. 1192–1202). Brussels, Belgium: Association for Computational Linguistics.

Wagers, M. W., Lau, E. F., & Phillips, C. (2009). Agreement attraction in comprehension: Representations and processes. Journal of Memory and Language, 61, 206–237.

Wilcox, E., Levy, R. P., Morita, T., & Futrell, R. (2018). What do RNN language models learn about filler–gap dependencies? In Proceedings of the workshop on analyzing and interpreting neural networks for NLP.