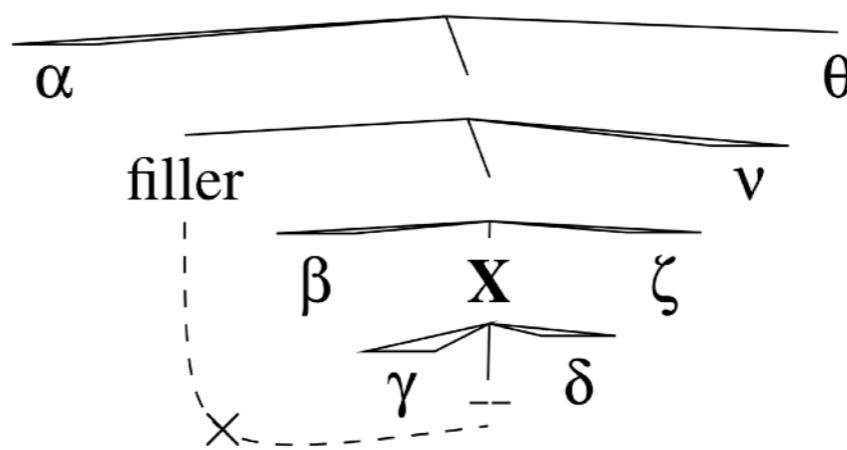
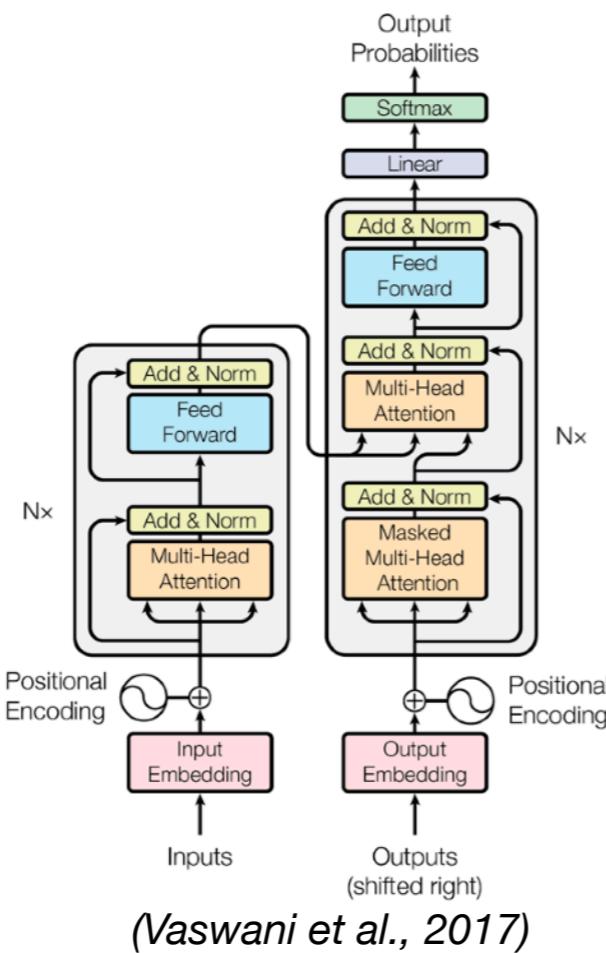


Transformer language models, targeted syntactic evaluation, and learnability



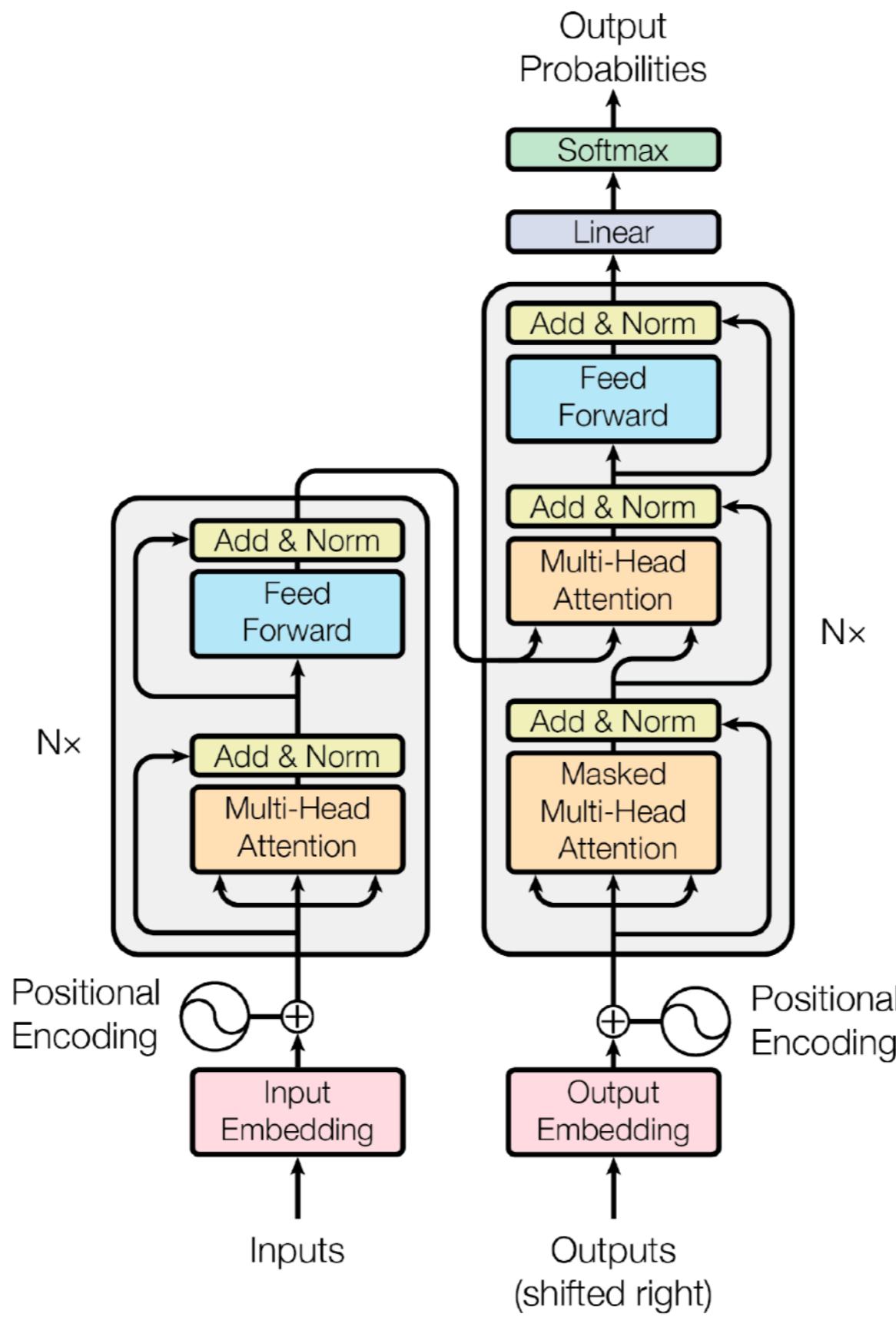
(Wilcox et al., 2019)

Roger Levy
9.19: Computational Psycholinguistics
15 November 2021

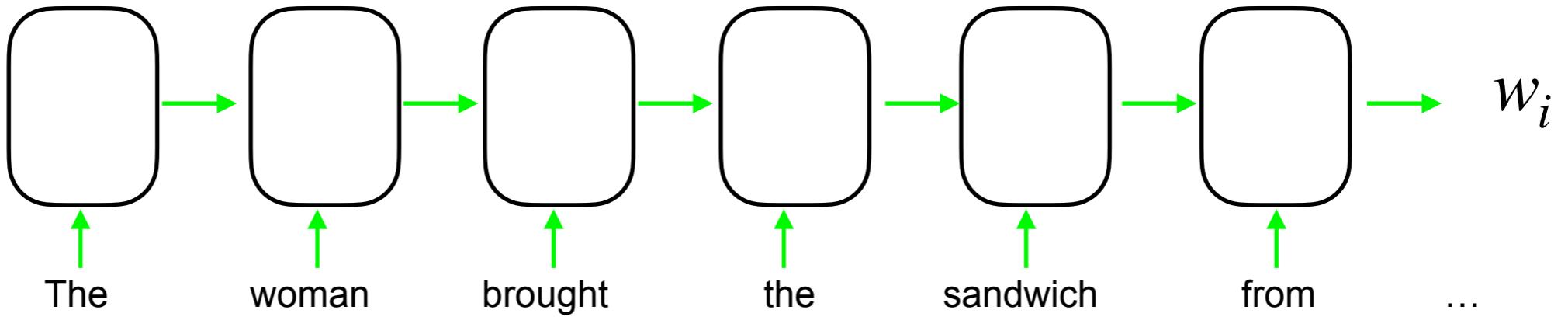
Agenda for today

- The Transformer
- Targeted syntactic testing: filler–gap dependencies
- Learnability: syntactic **islands**

The Transformer model

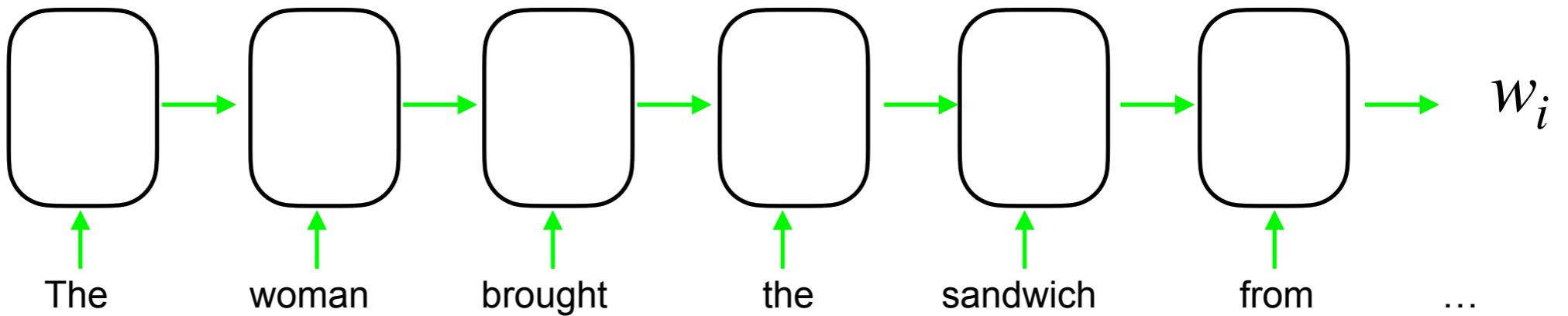


Motivating the Transformer model



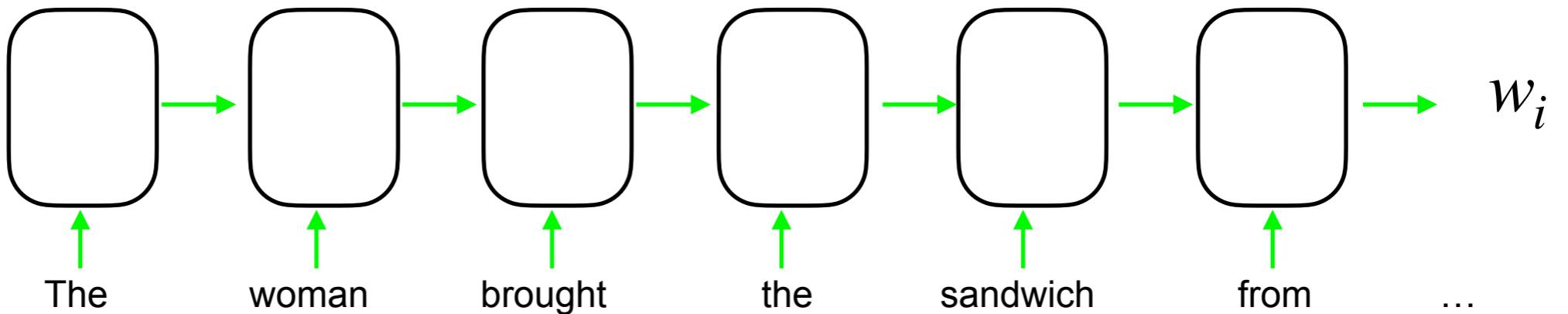
Motivating the Transformer model

- With RNNs, a fixed-dimension model could propagate information indefinitely into the future...but it's hard!



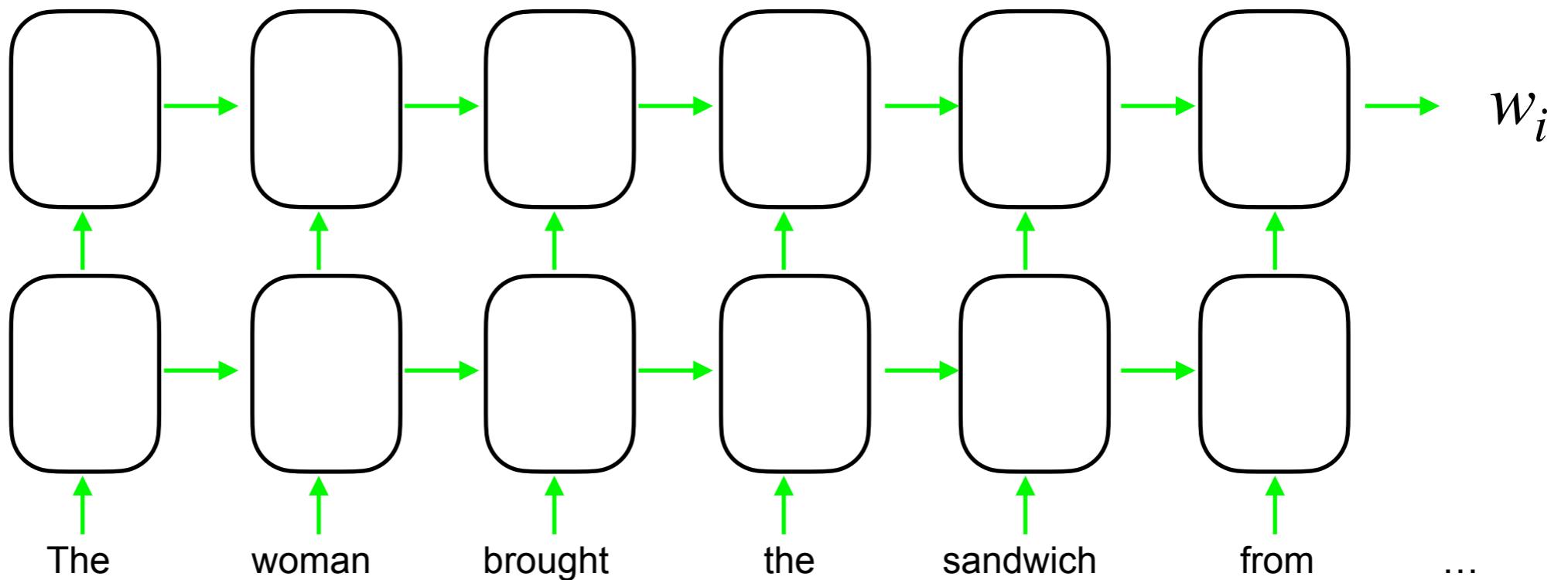
Motivating the Transformer model

- With RNNs, a fixed-dimension model could propagate information indefinitely into the future...but it's hard!
- We can make RNNs **deep** by stacking them...



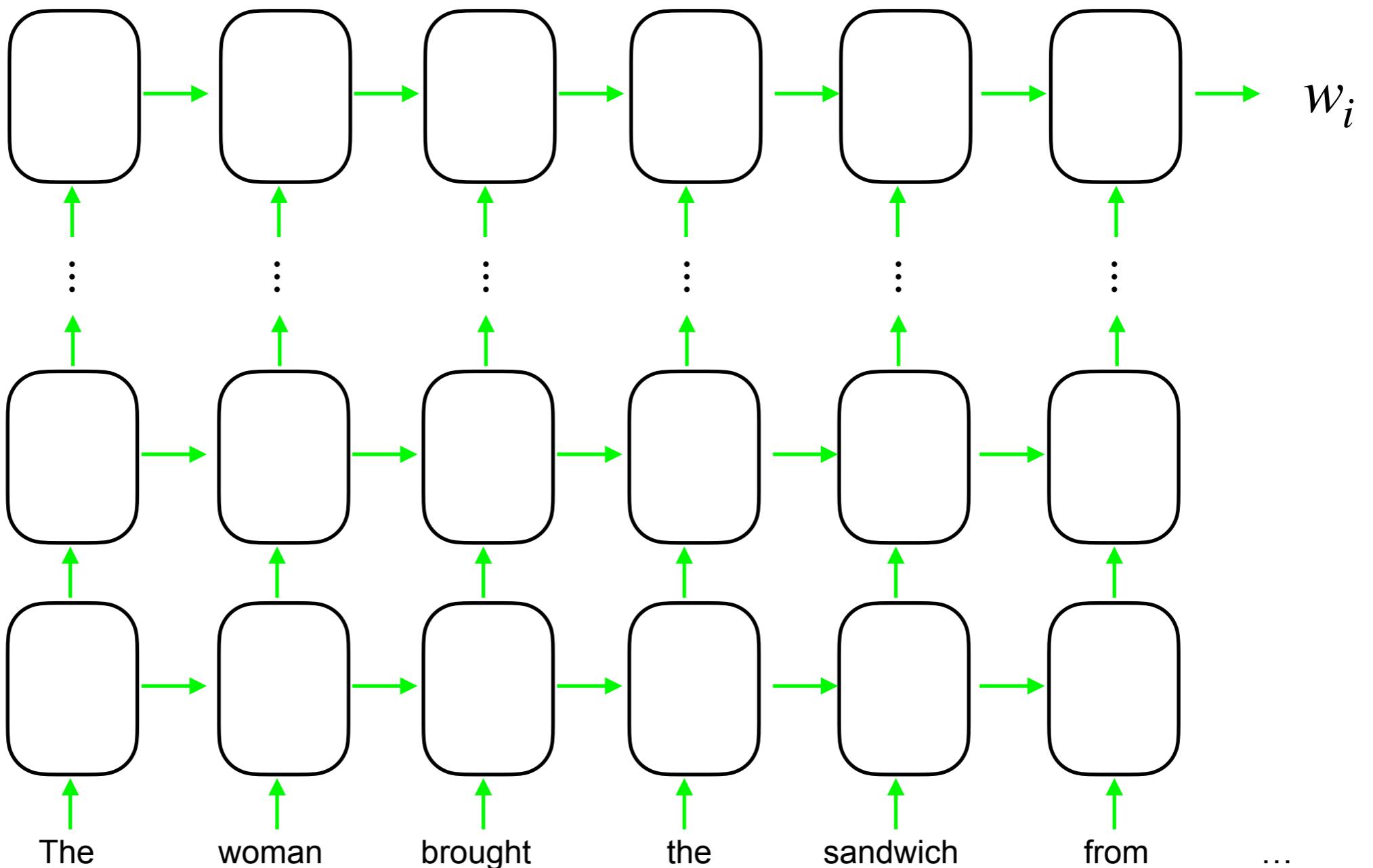
Motivating the Transformer model

- With RNNs, a fixed-dimension model could propagate information indefinitely into the future...but it's hard!
- We can make RNNs **deep** by stacking them...



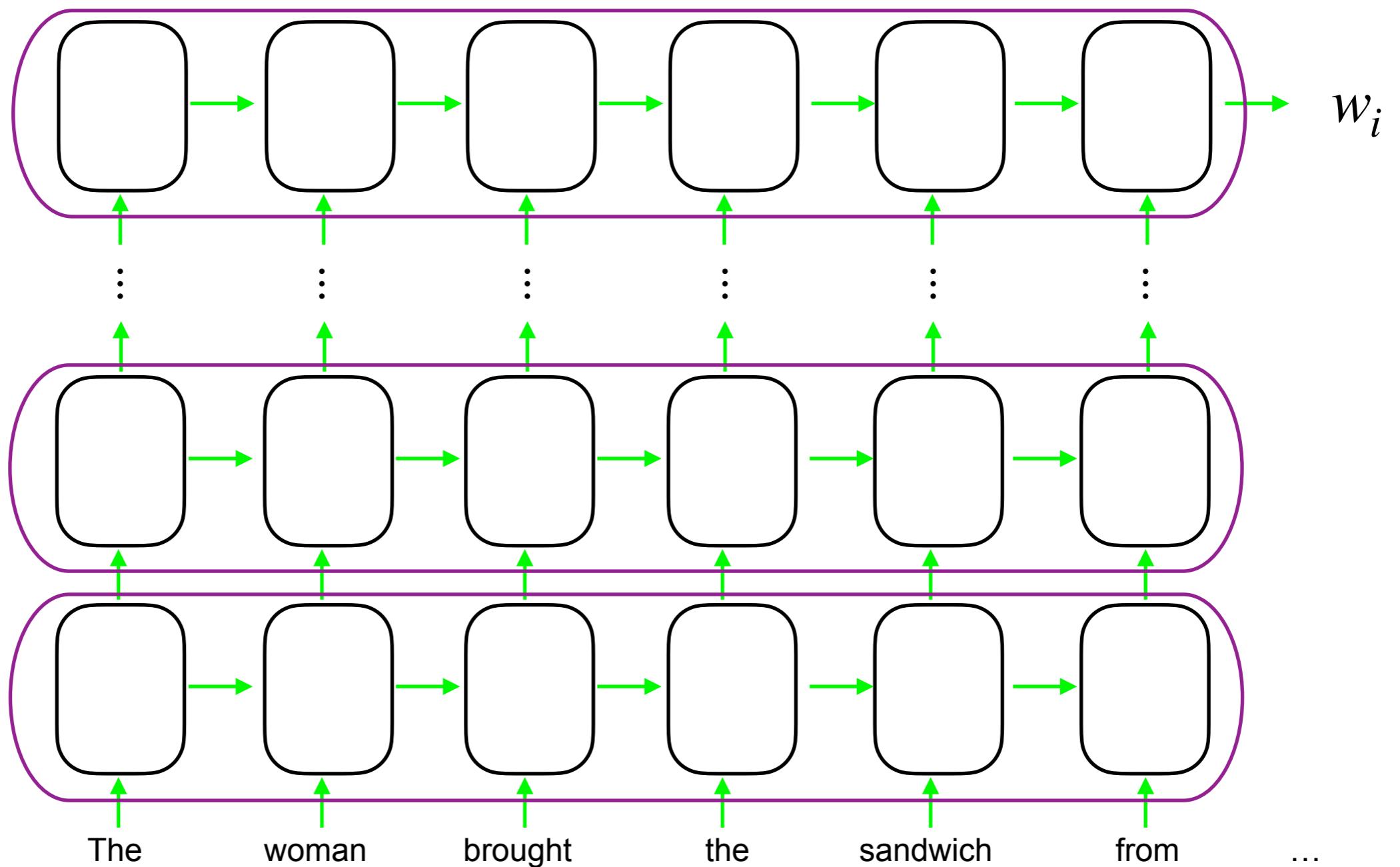
Motivating the Transformer model

- With RNNs, a fixed-dimension model could propagate information indefinitely into the future...but it's hard!
- We can make RNNs **deep** by stacking them...

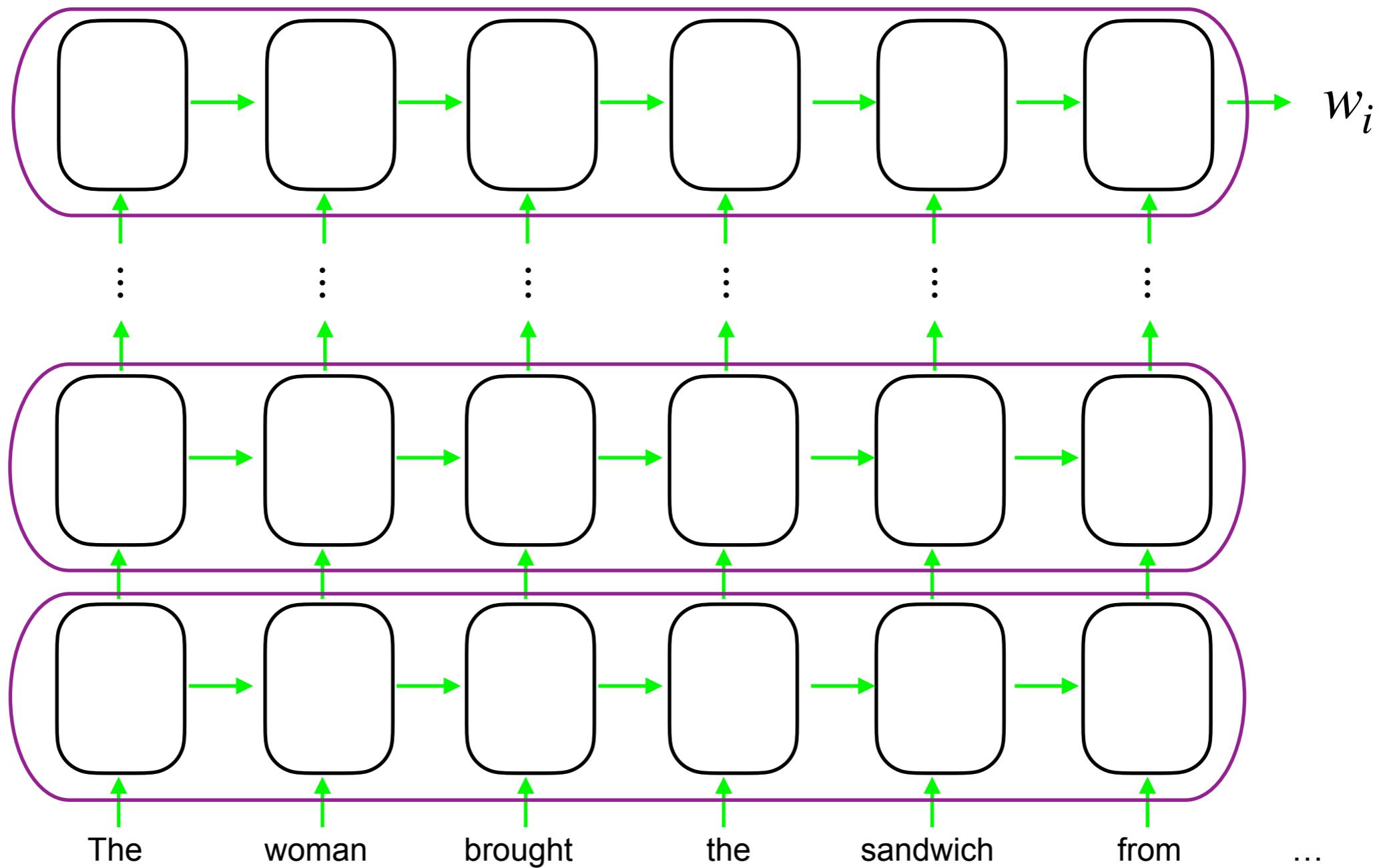


Motivating the Transformer model

- With RNNs, a fixed-dimension model could propagate information indefinitely into the future...but it's hard!
- We can make RNNs **deep** by stacking them...

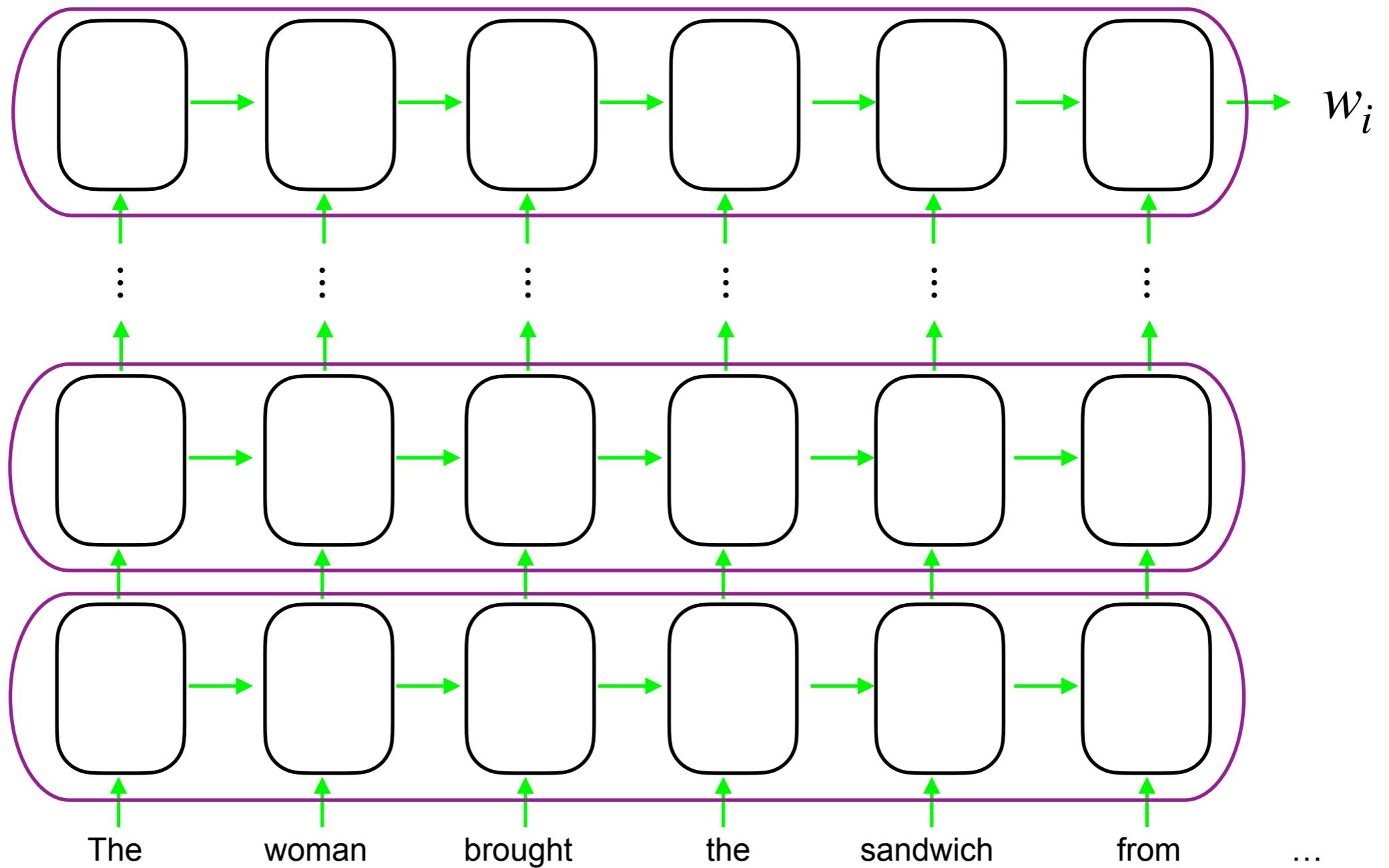


Motivating the Transformer model



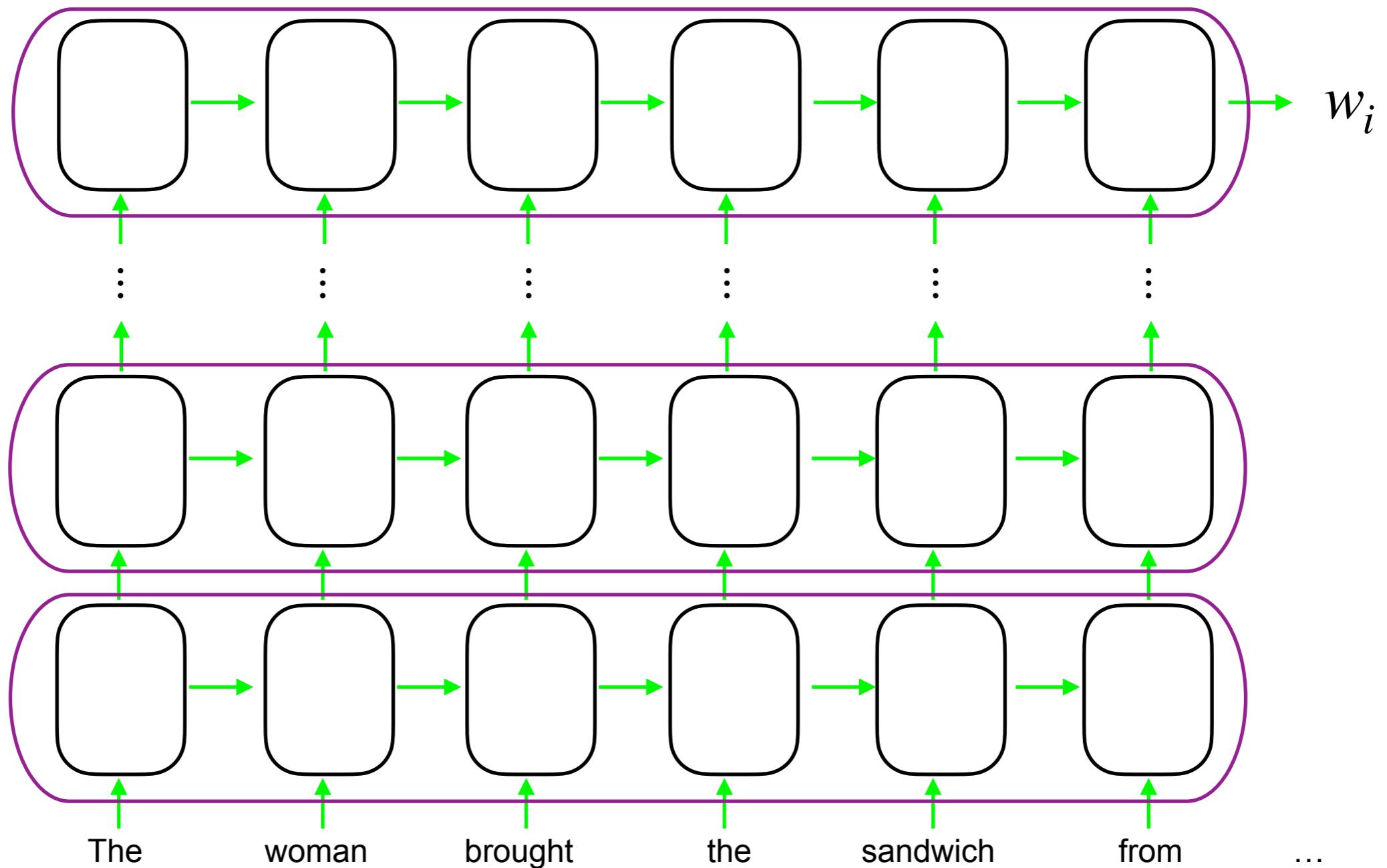
Motivating the Transformer model

- ...but input distant in the context is still far away.



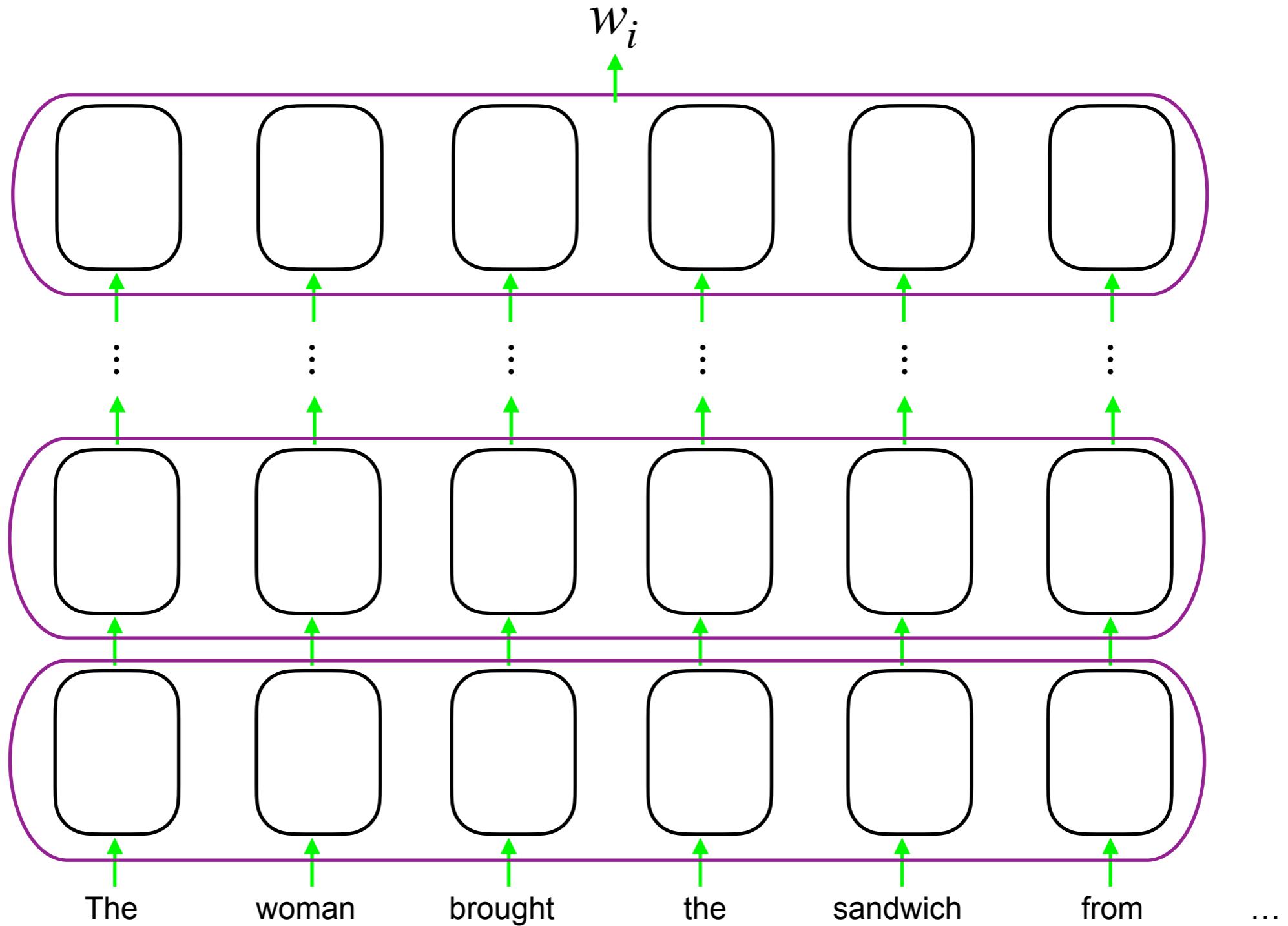
Motivating the Transformer model

- ...but input distant in the context is still far away.
- Solution: make all context words equally distant from w_i !



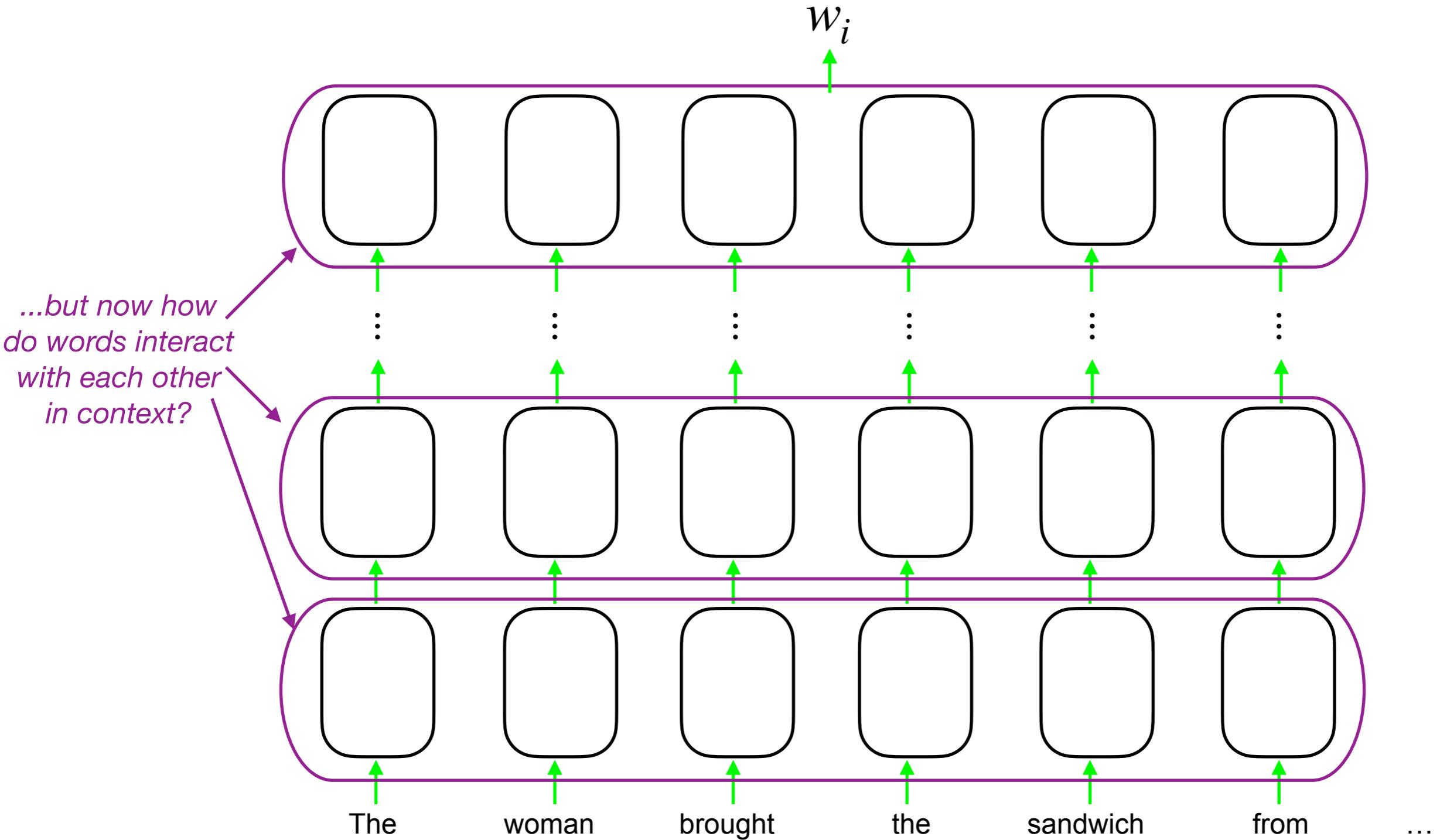
Motivating the Transformer model

- ...but input distant in the context is still far away.
- Solution: make all context words equally distant from w_i !



Motivating the Transformer model

- ...but input distant in the context is still far away.
- Solution: make all context words equally distant from w_i !



Input + Positional Embedding

the

dog

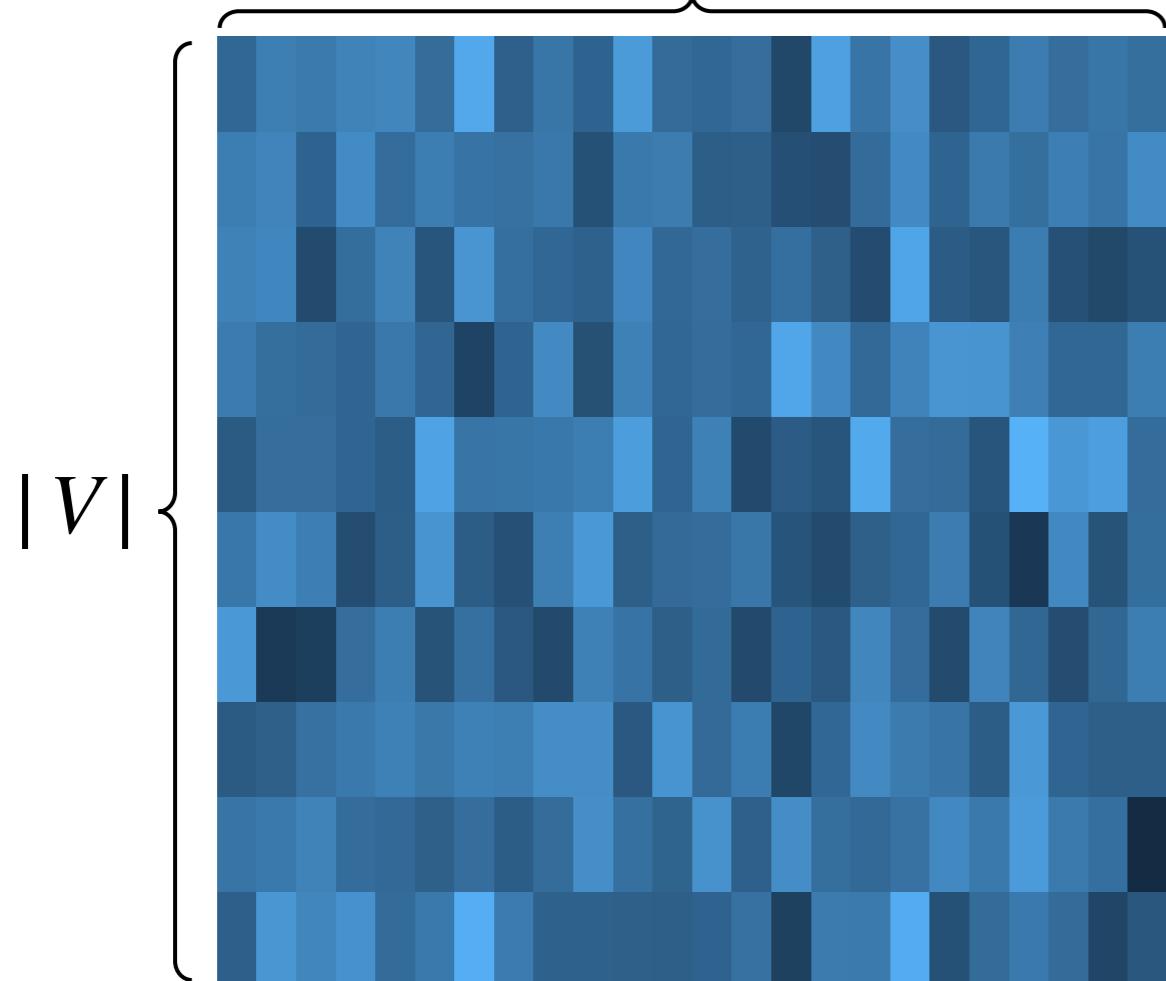
ate

the

...

Input + Positional Embedding

Word embedding matrix: d



the

dog

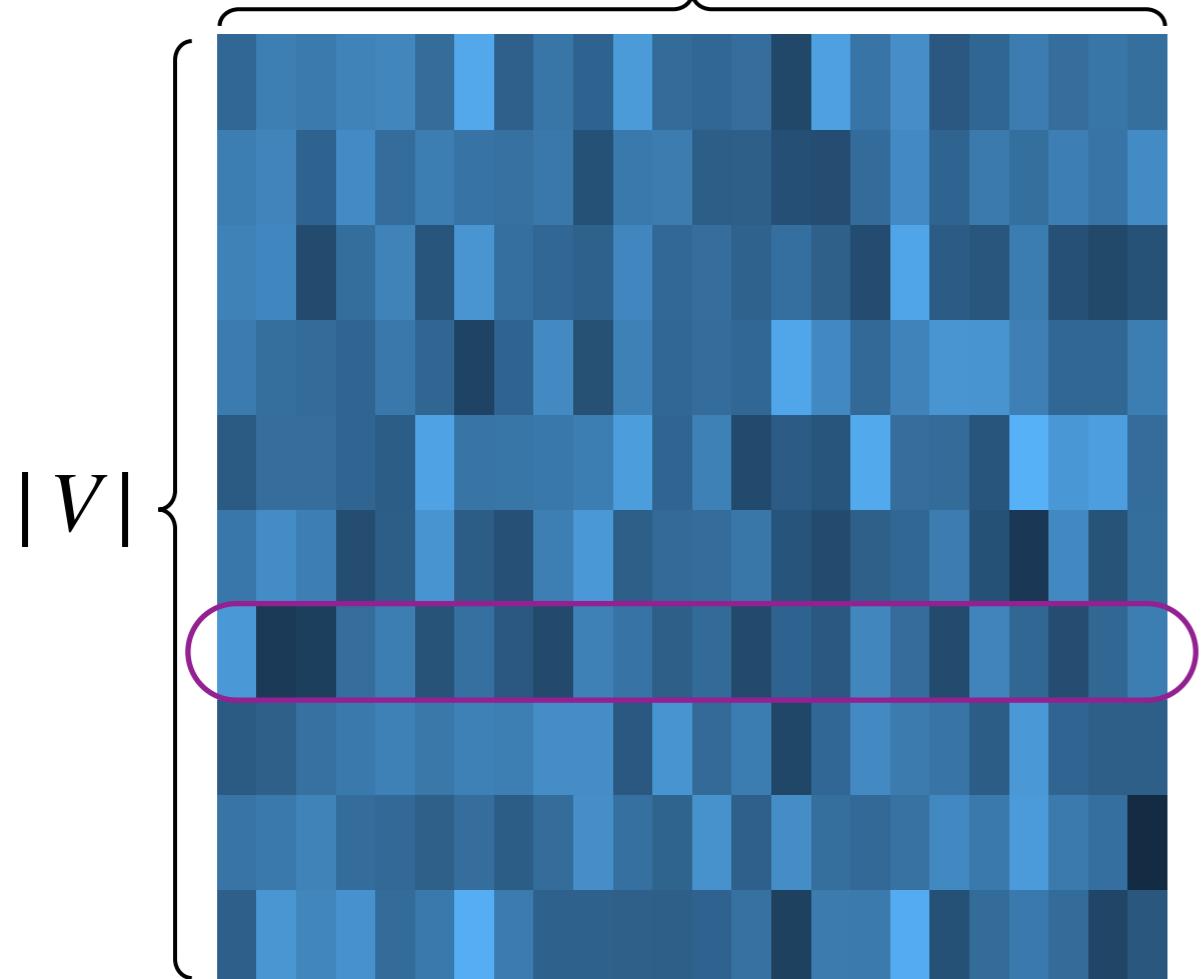
ate

the

...

Input + Positional Embedding

Word embedding matrix: d



the

dog

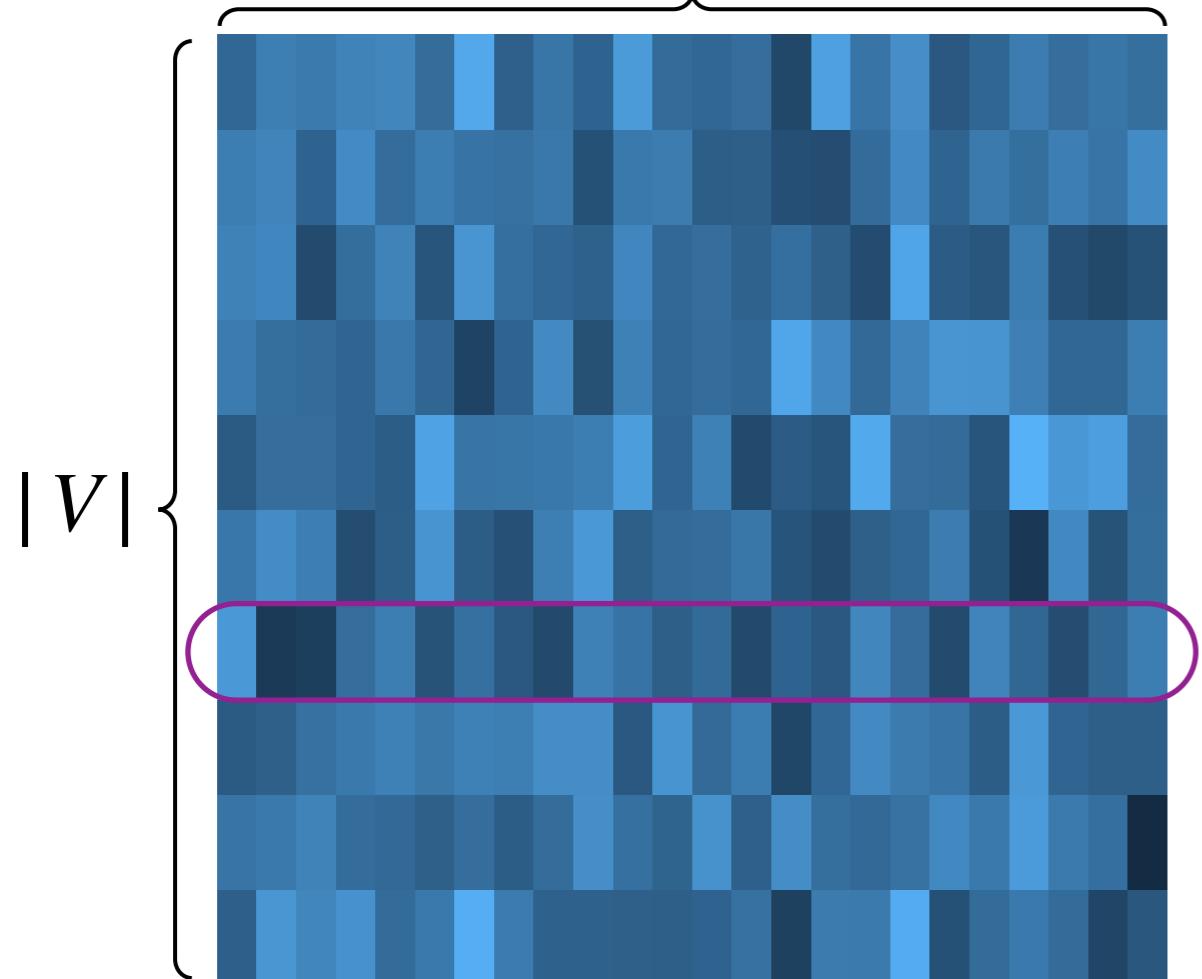
ate

the

...

Input + Positional Embedding

Word embedding matrix: d



the

dog

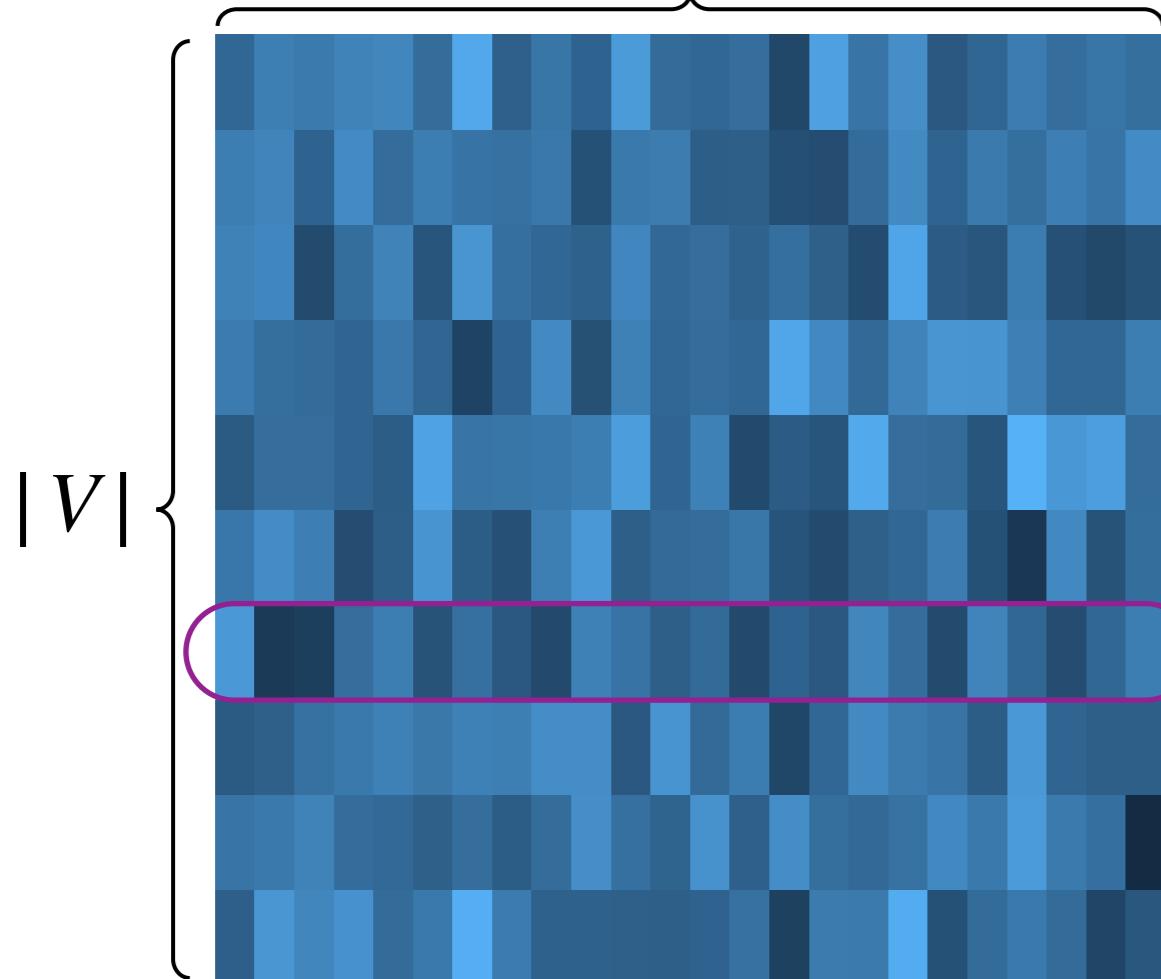


the

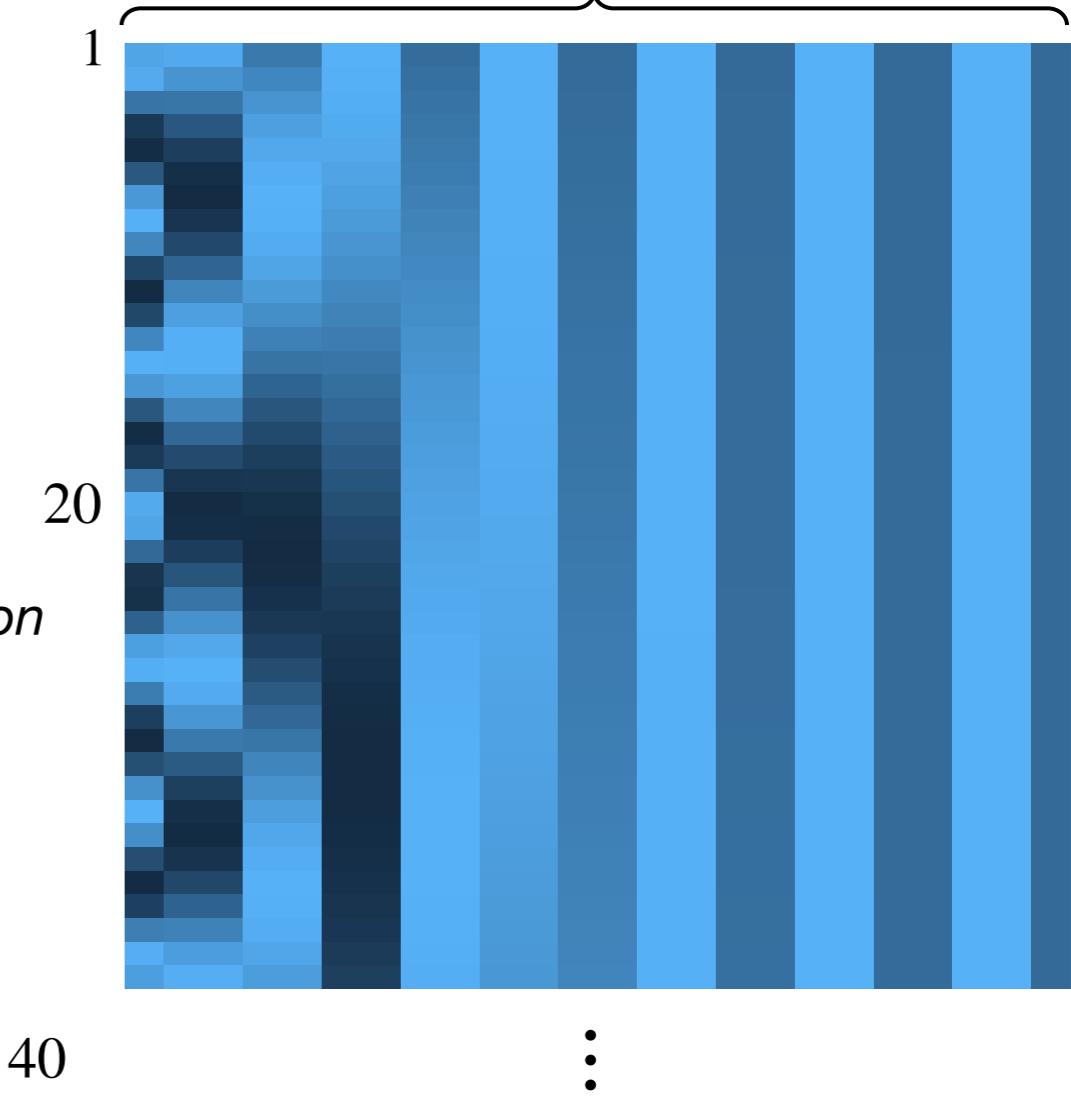
...

Input + Positional Embedding

Word embedding matrix: d



Position embedding matrix: d



the

dog

ate

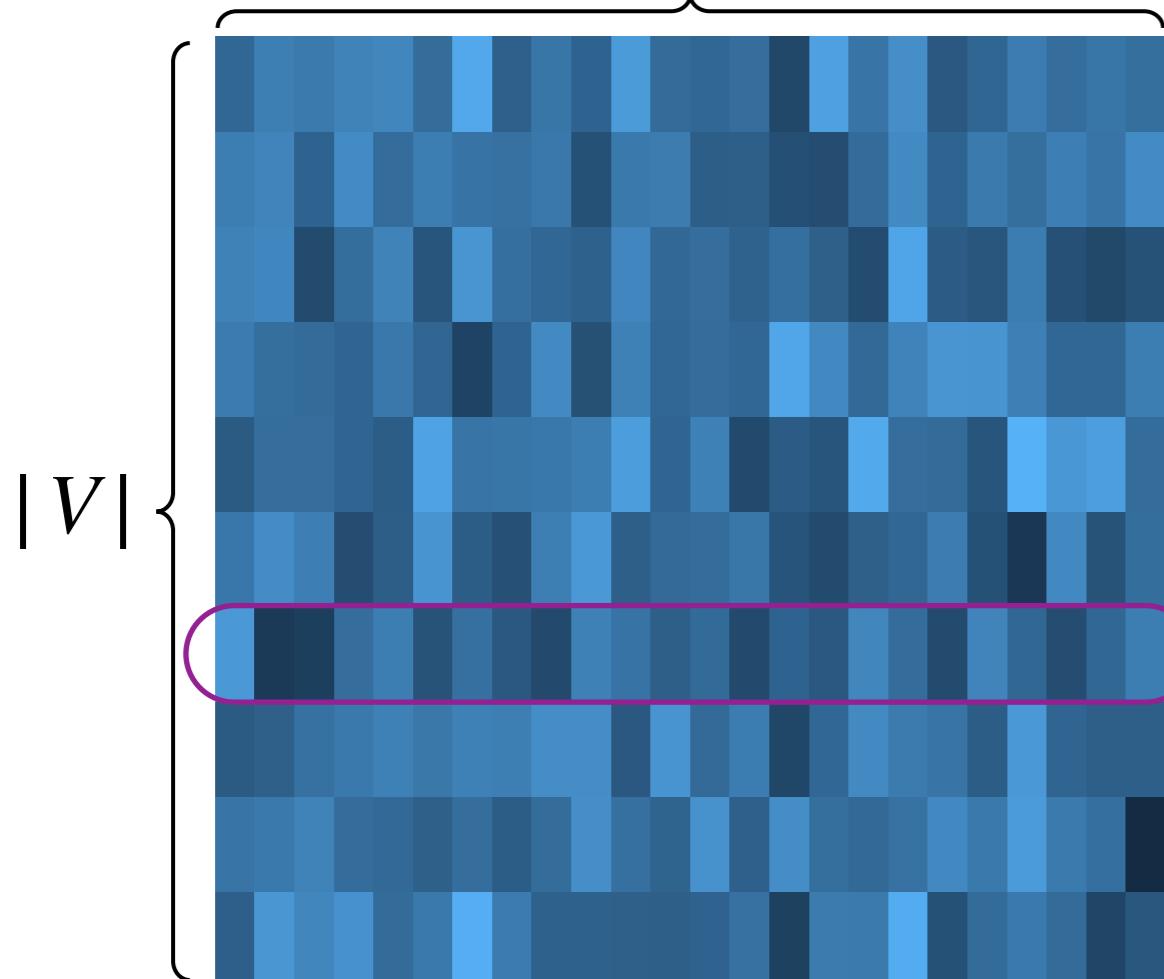


the

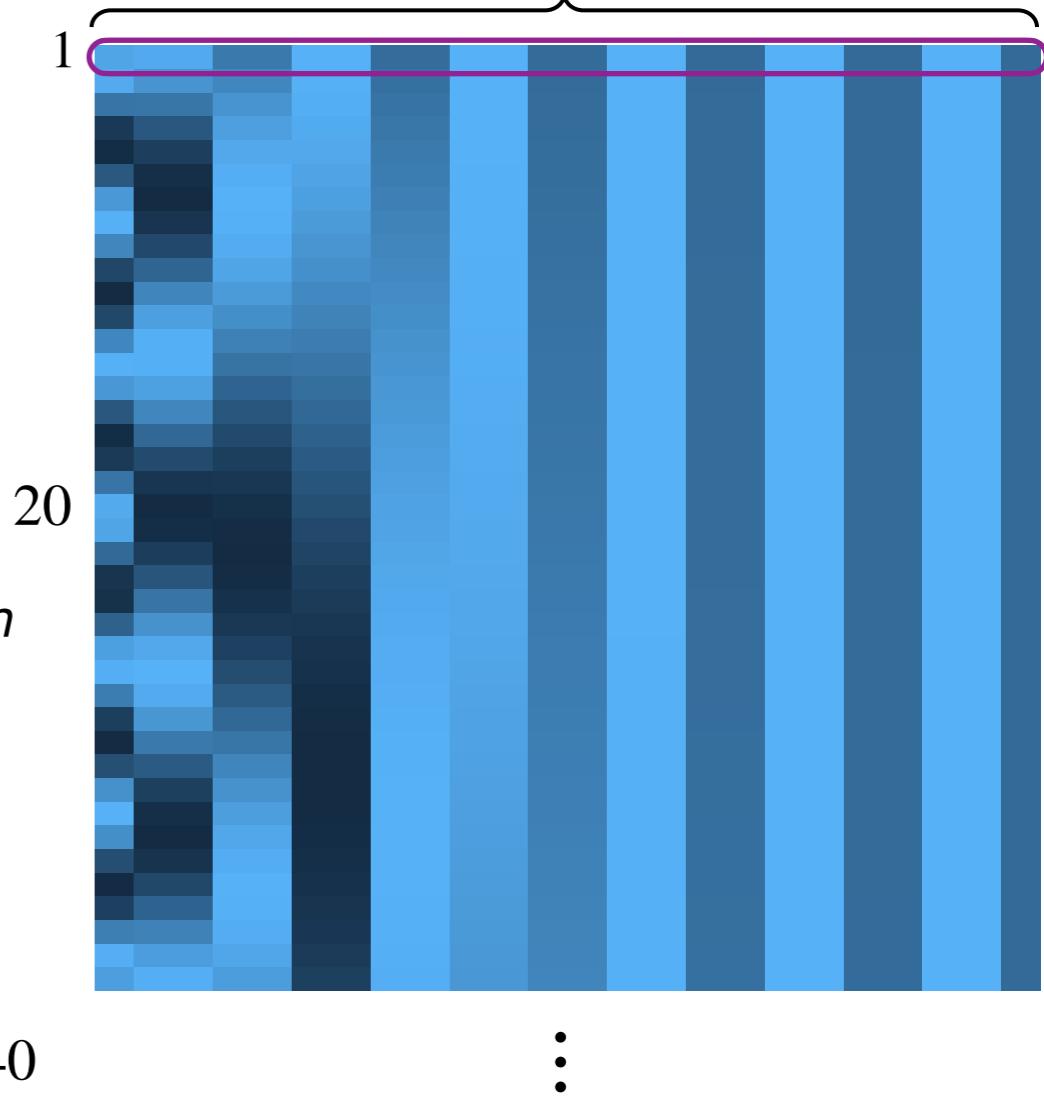
...

Input + Positional Embedding

Word embedding matrix: d



Position embedding matrix: d



the

dog

ate

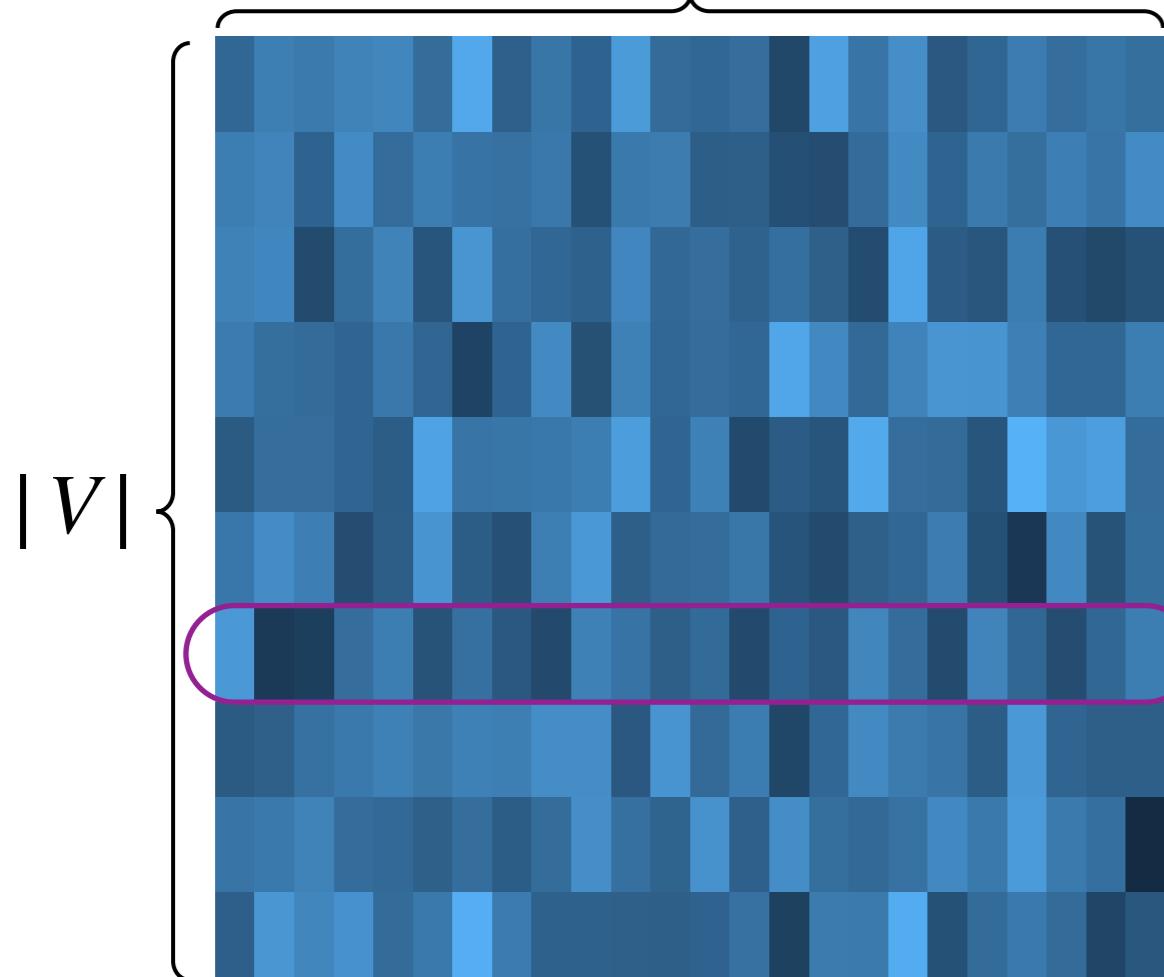


the

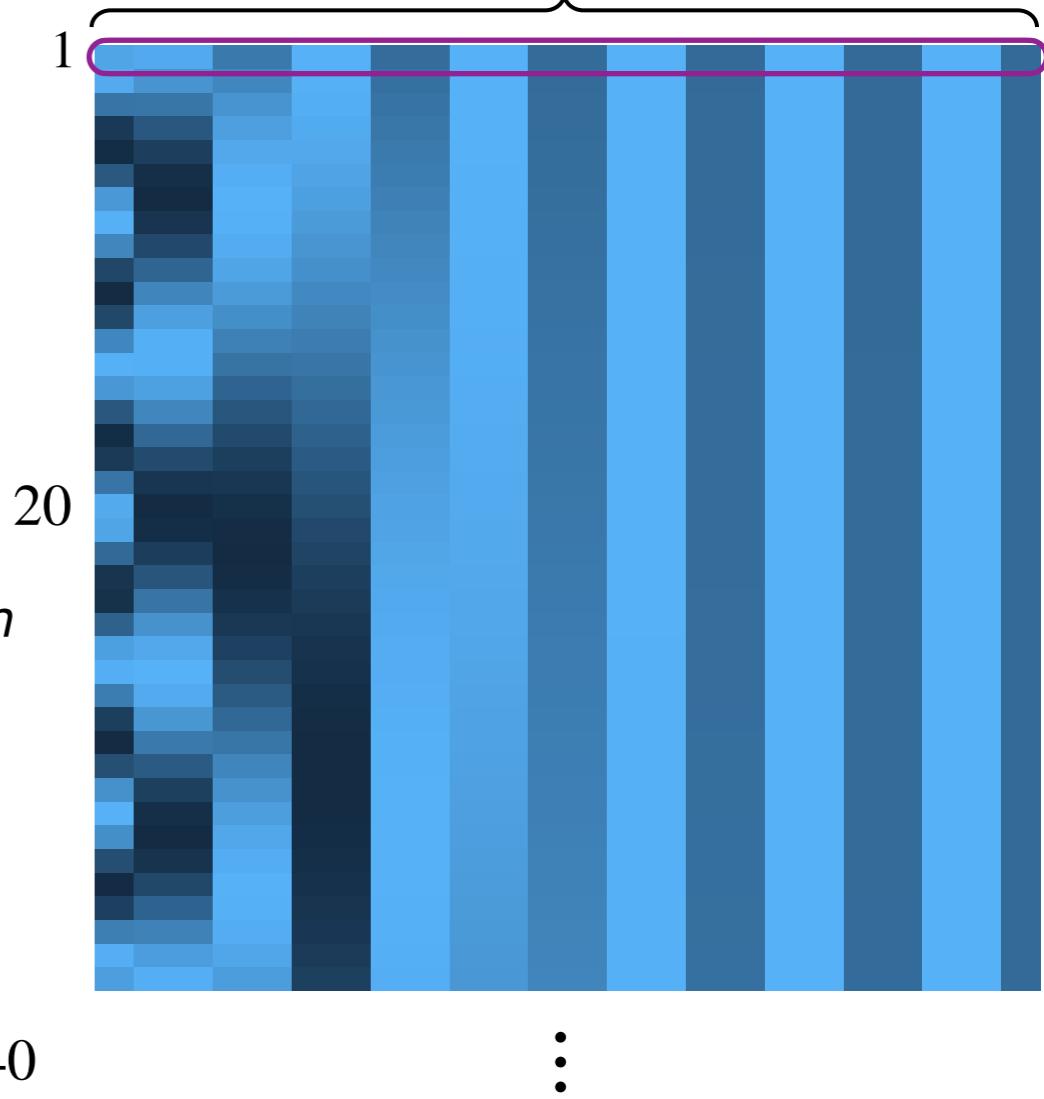
...

Input + Positional Embedding

Word embedding matrix: d



Position embedding matrix: d



||



the

dog

ate

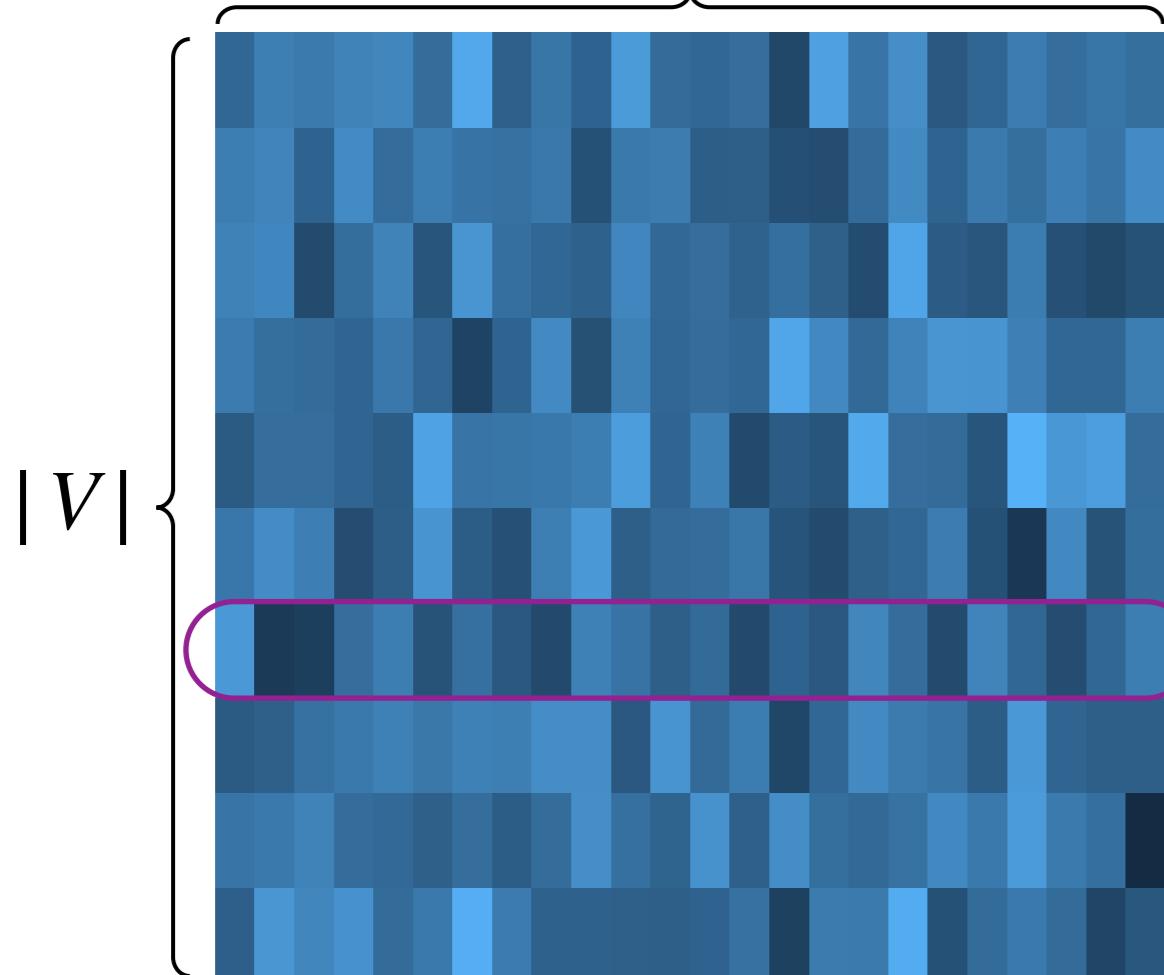


the

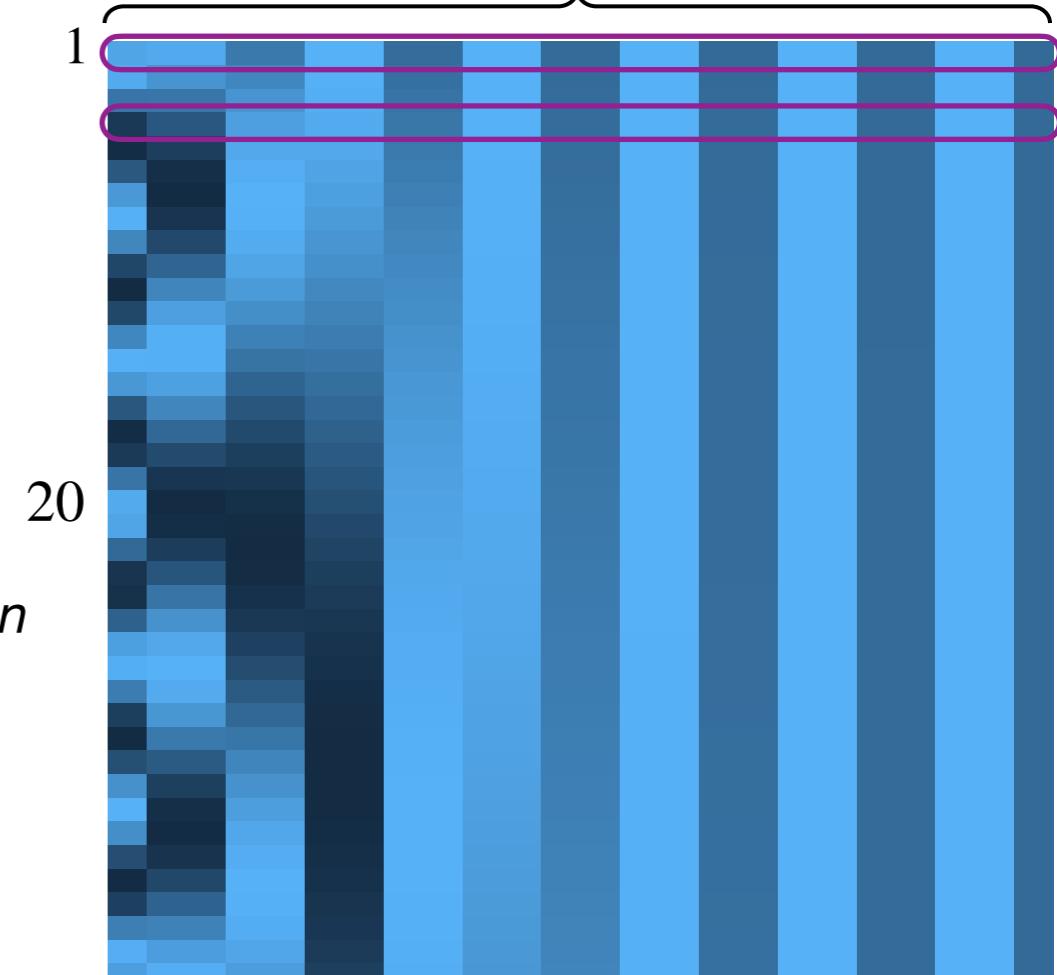
...

Input + Positional Embedding

Word embedding matrix: d



Position embedding matrix: d



||



the

dog

ate

40

⋮

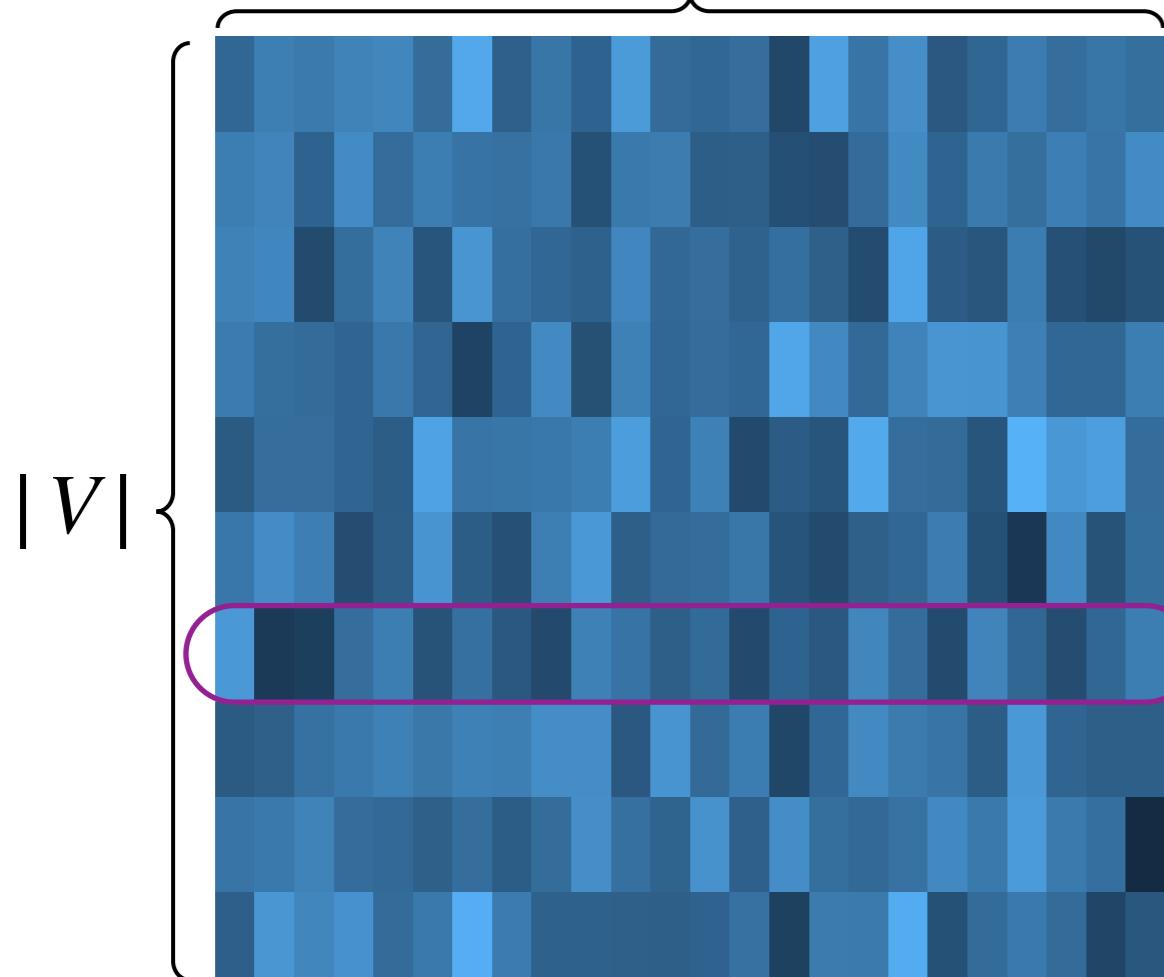


the

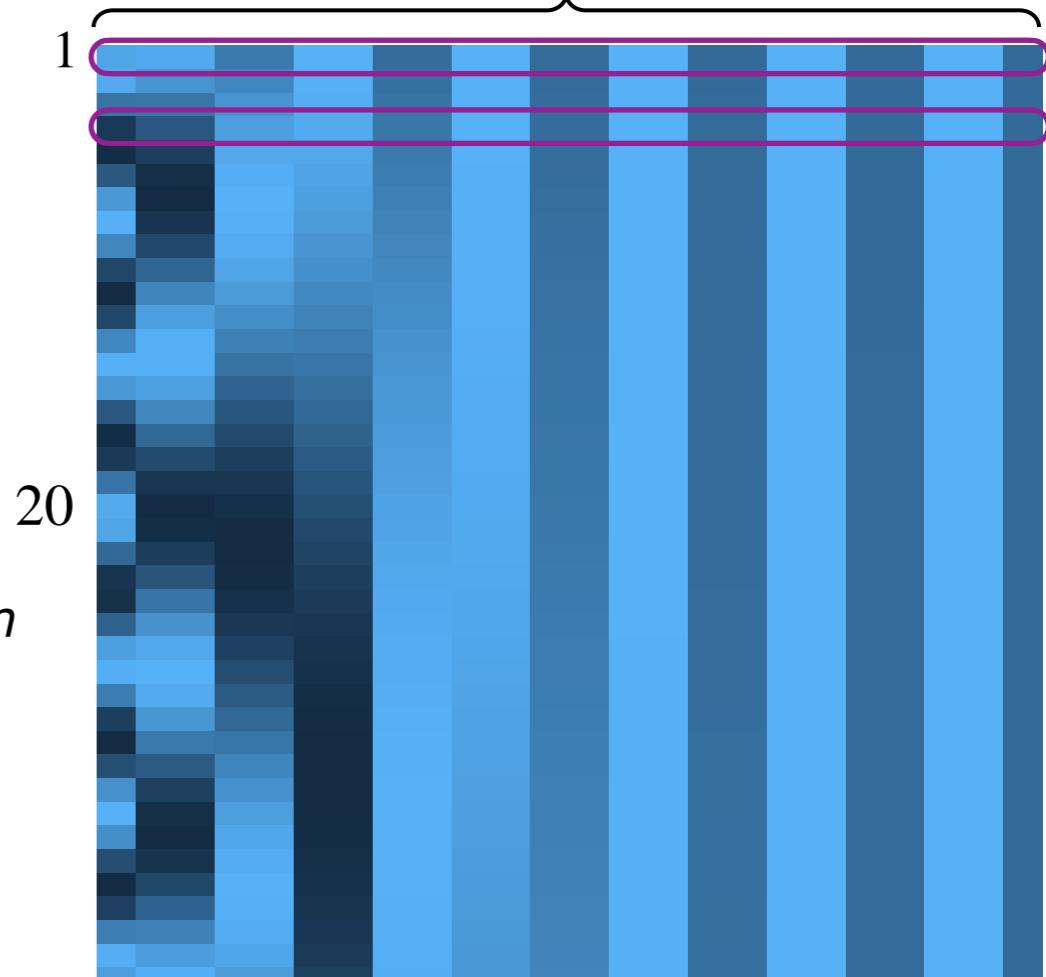
...

Input + Positional Embedding

Word embedding matrix: d



Position embedding matrix: d



||



⊕



the

dog

ate



||



⊕



the

...

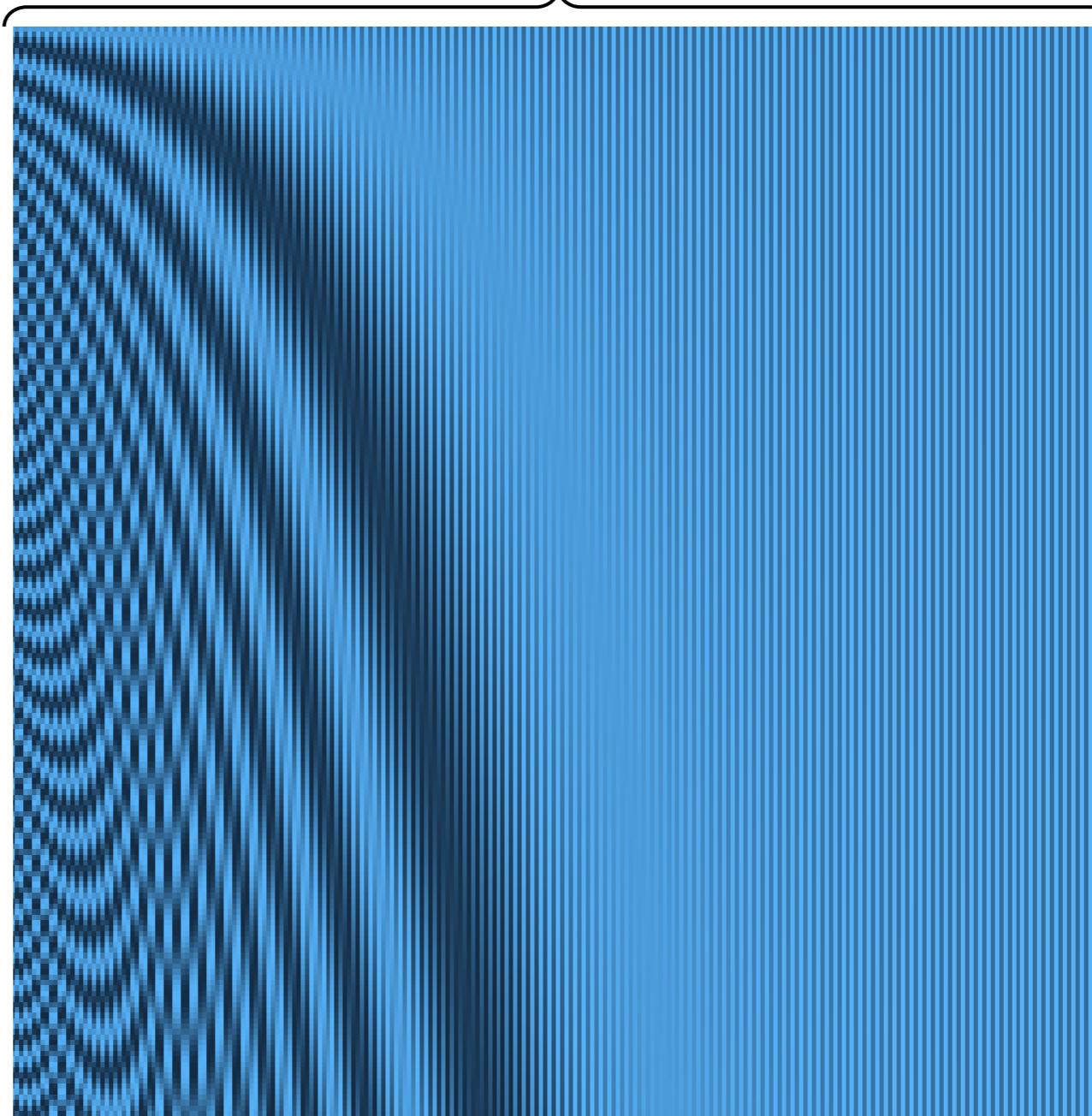
The positional embedding function

$$PE(pos, 2i) = \sin\left(\frac{pos}{10000^{\frac{2i}{d}}}\right)$$

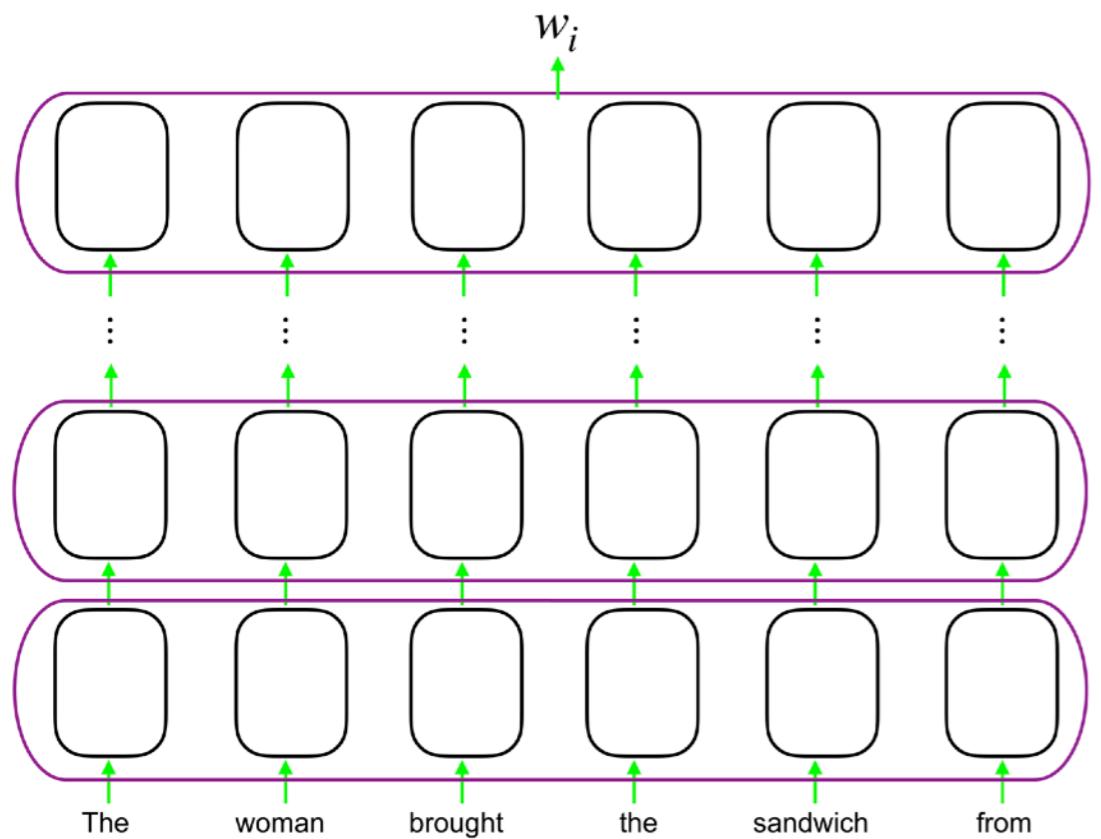
$$PE(pos, 2i + 1) = \cos\left(\frac{pos}{10000^{\frac{2i}{d}}}\right)$$

$d = 512$

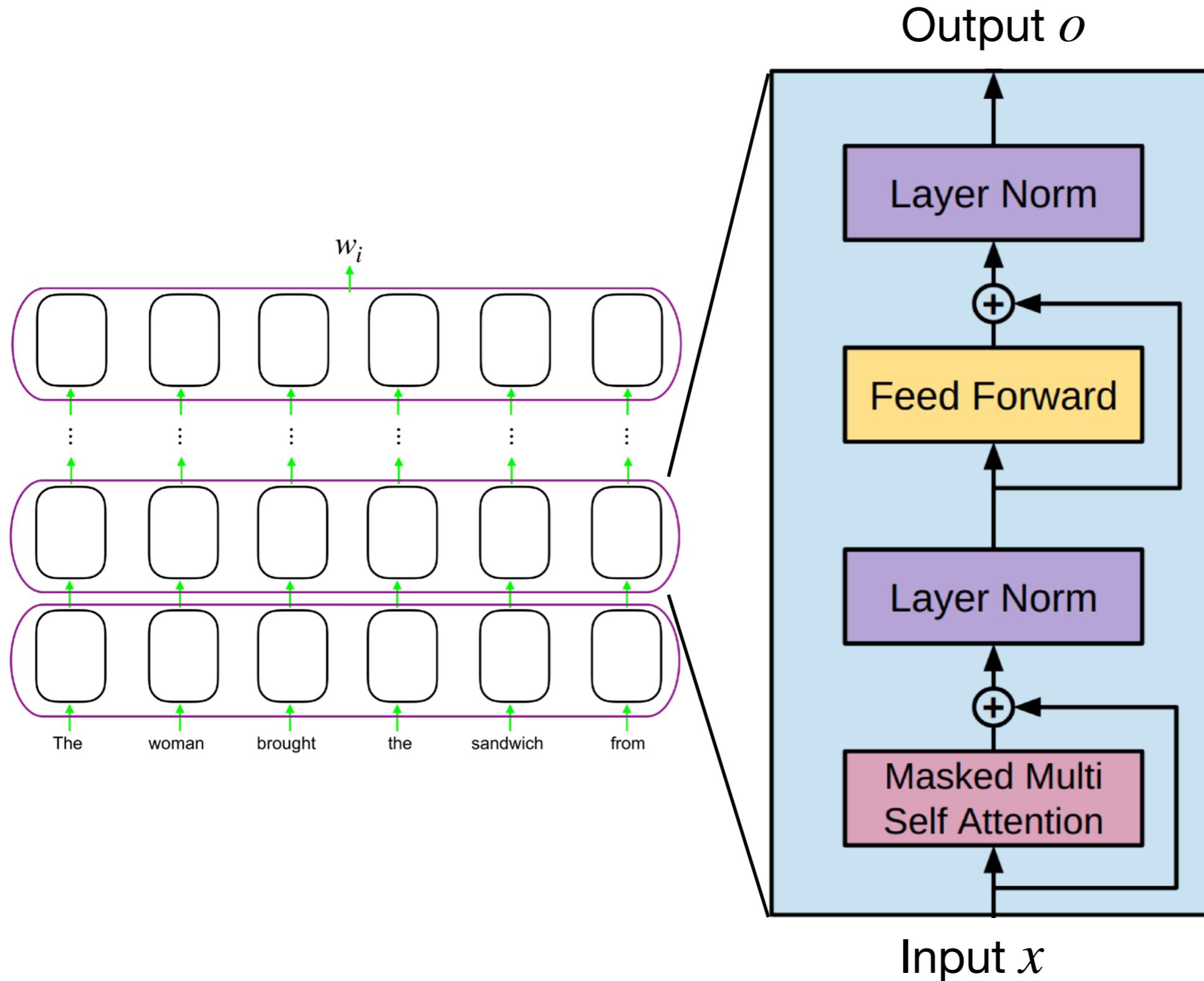
*word
position*



The Transformer unit

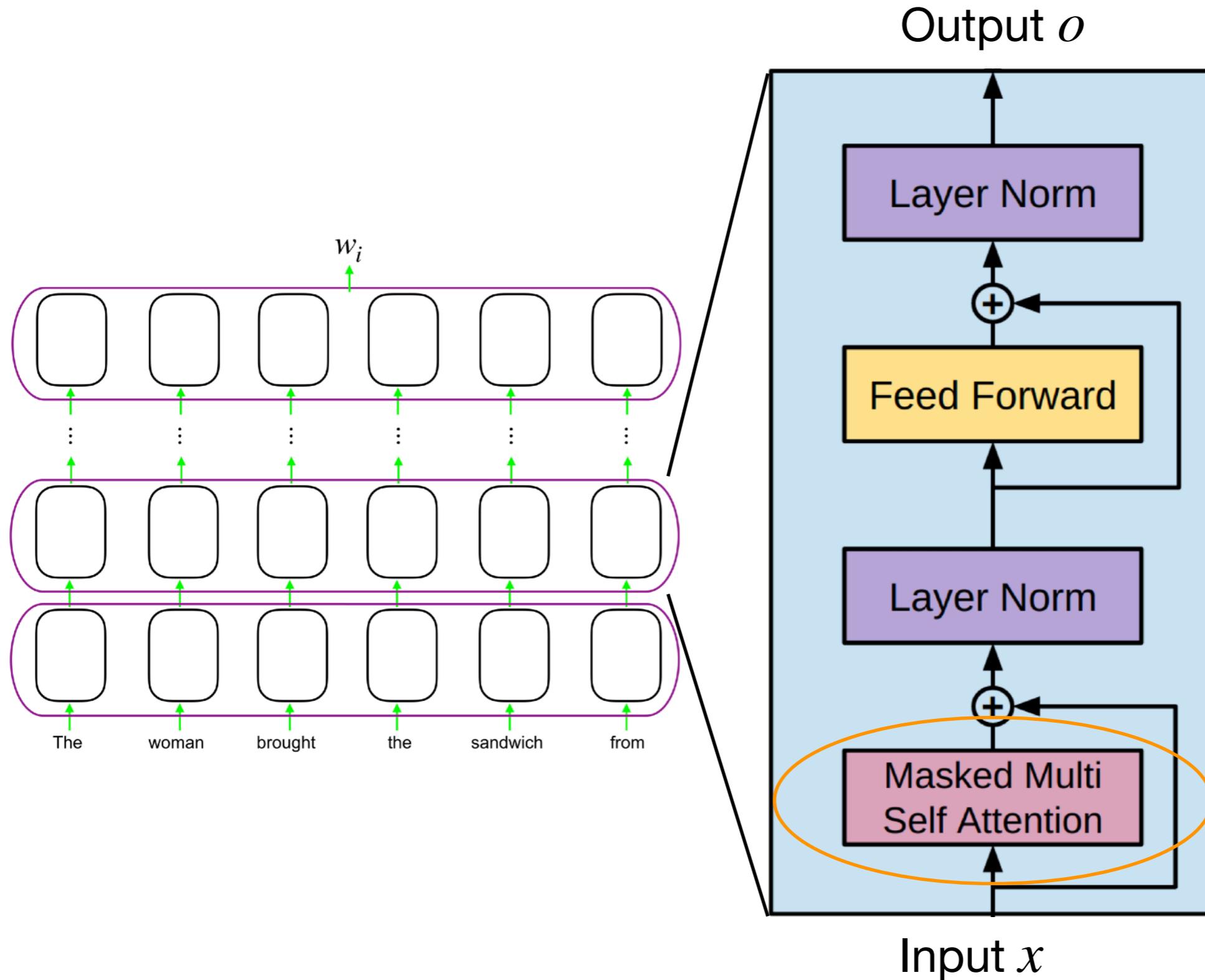


The Transformer unit



(Figure from Radford et al., 2018)

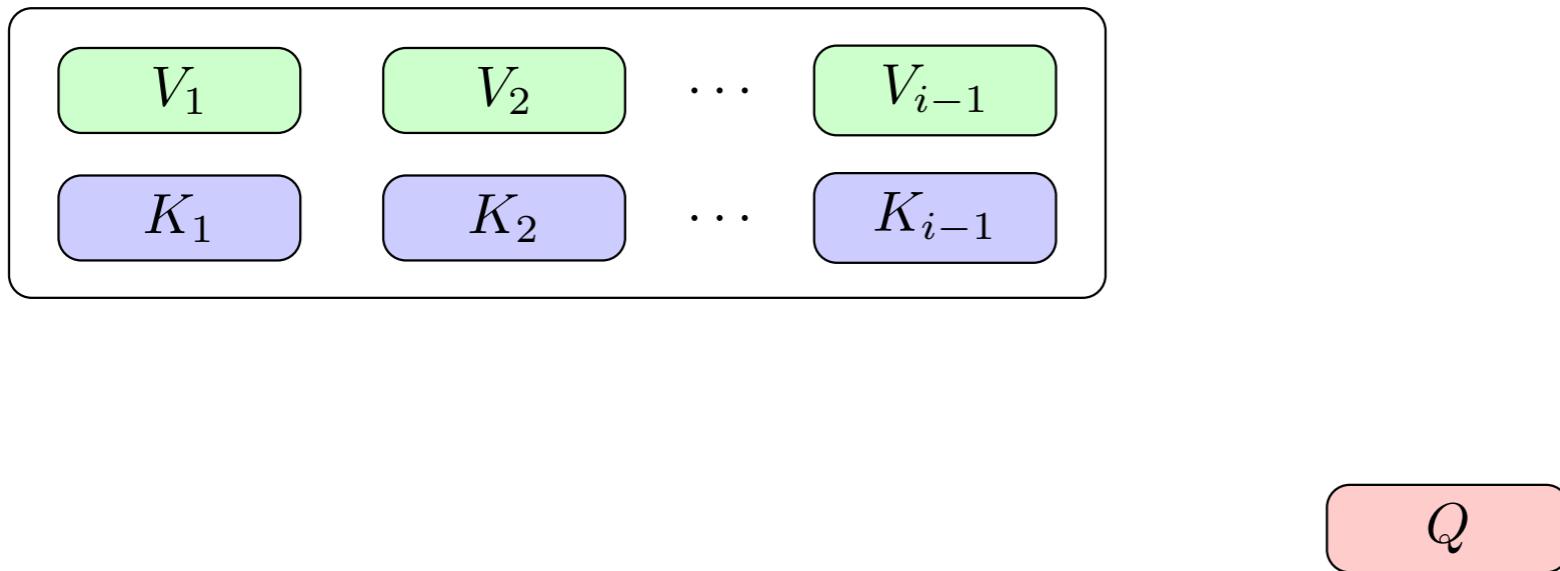
The Transformer unit



(Figure from Radford et al., 2018)

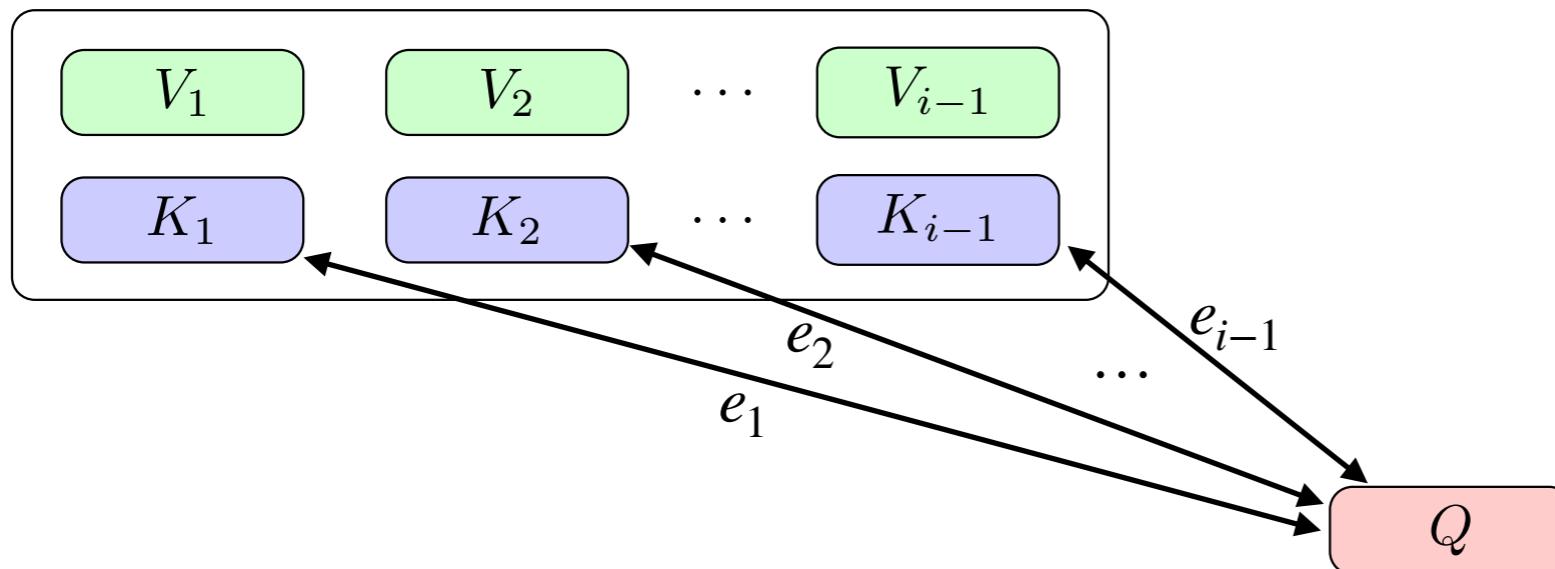
Neural Attention

Query, Key, and Value



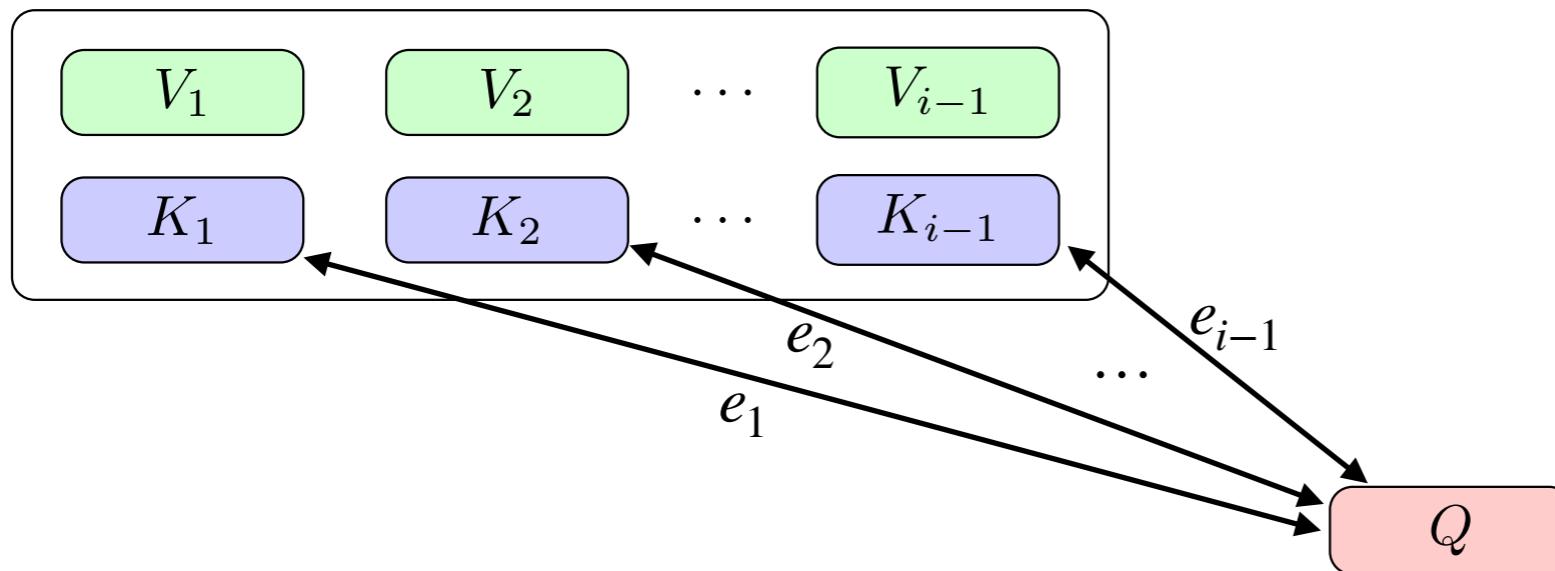
Neural Attention

Query, Key, and Value



Neural Attention

Query, Key, and Value

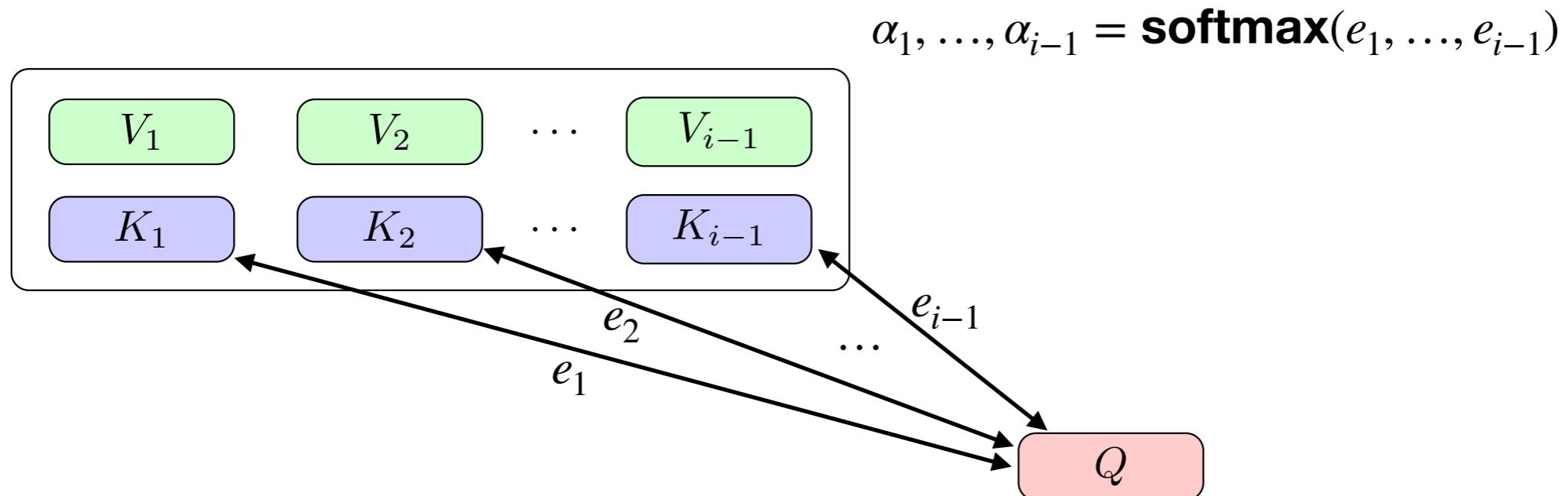


Attention function options:

$$e_j = \begin{cases} v \tanh [W_Q Q + W_K K_j] & \text{(Bahdanau et al., 2014)} \\ Q^T W K_j & \text{(Luong et al., 2015)} \\ \frac{Q^T K}{\sqrt{|K|}} & \text{(Vaswani et al., 2017)} \end{cases}$$

Neural Attention

Query, Key, and Value

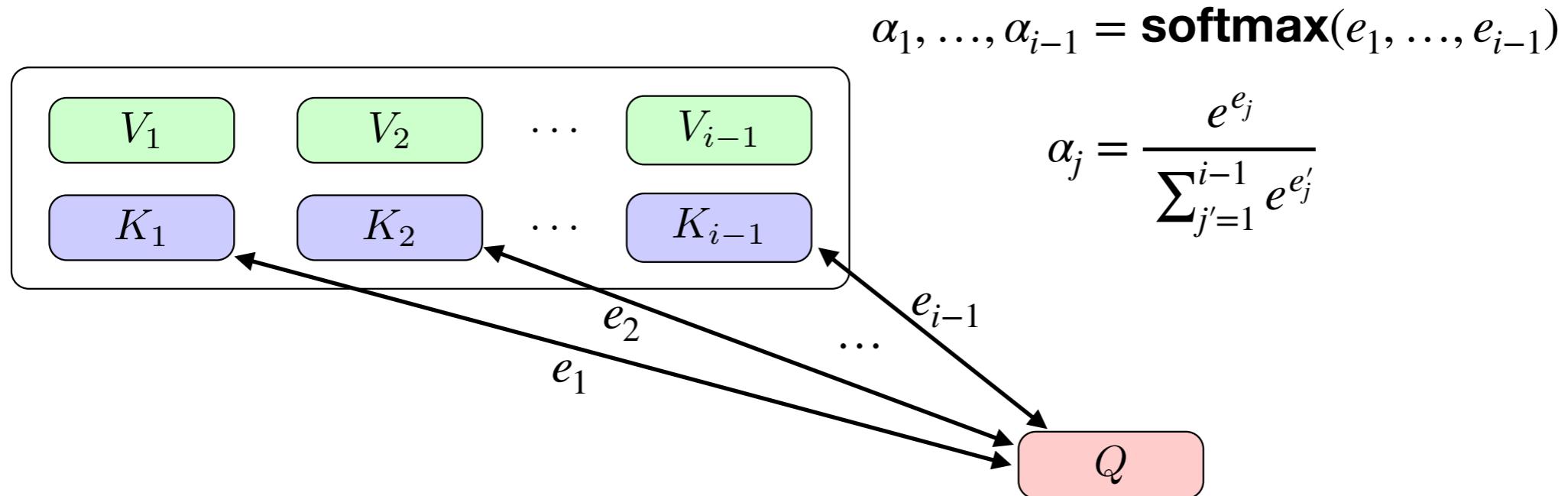


Attention function options:

$$e_j = \begin{cases} v \tanh [W_Q Q + W_K K_j] & \text{(Bahdanau et al., 2014)} \\ Q^T W K_j & \text{(Luong et al., 2015)} \\ \frac{Q^T K}{\sqrt{|K|}} & \text{(Vaswani et al., 2017)} \end{cases}$$

Neural Attention

Query, Key, and Value

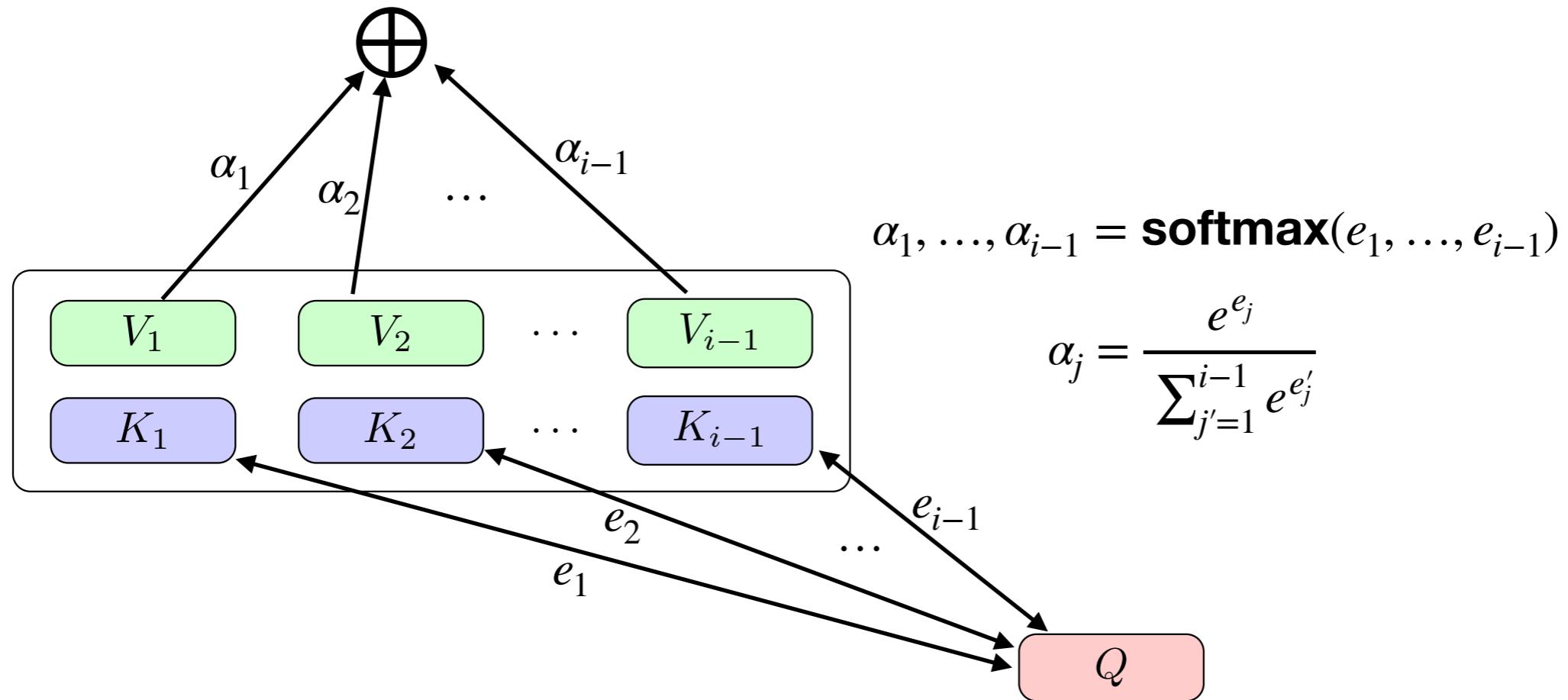


Attention function options:

$$e_j = \begin{cases} v \tanh [W_Q Q + W_K K_j] & \text{(Bahdanau et al., 2014)} \\ Q^T W K_j & \text{(Luong et al., 2015)} \\ \frac{Q^T K}{\sqrt{|K|}} & \text{(Vaswani et al., 2017)} \end{cases}$$

Neural Attention

Query, Key, and Value

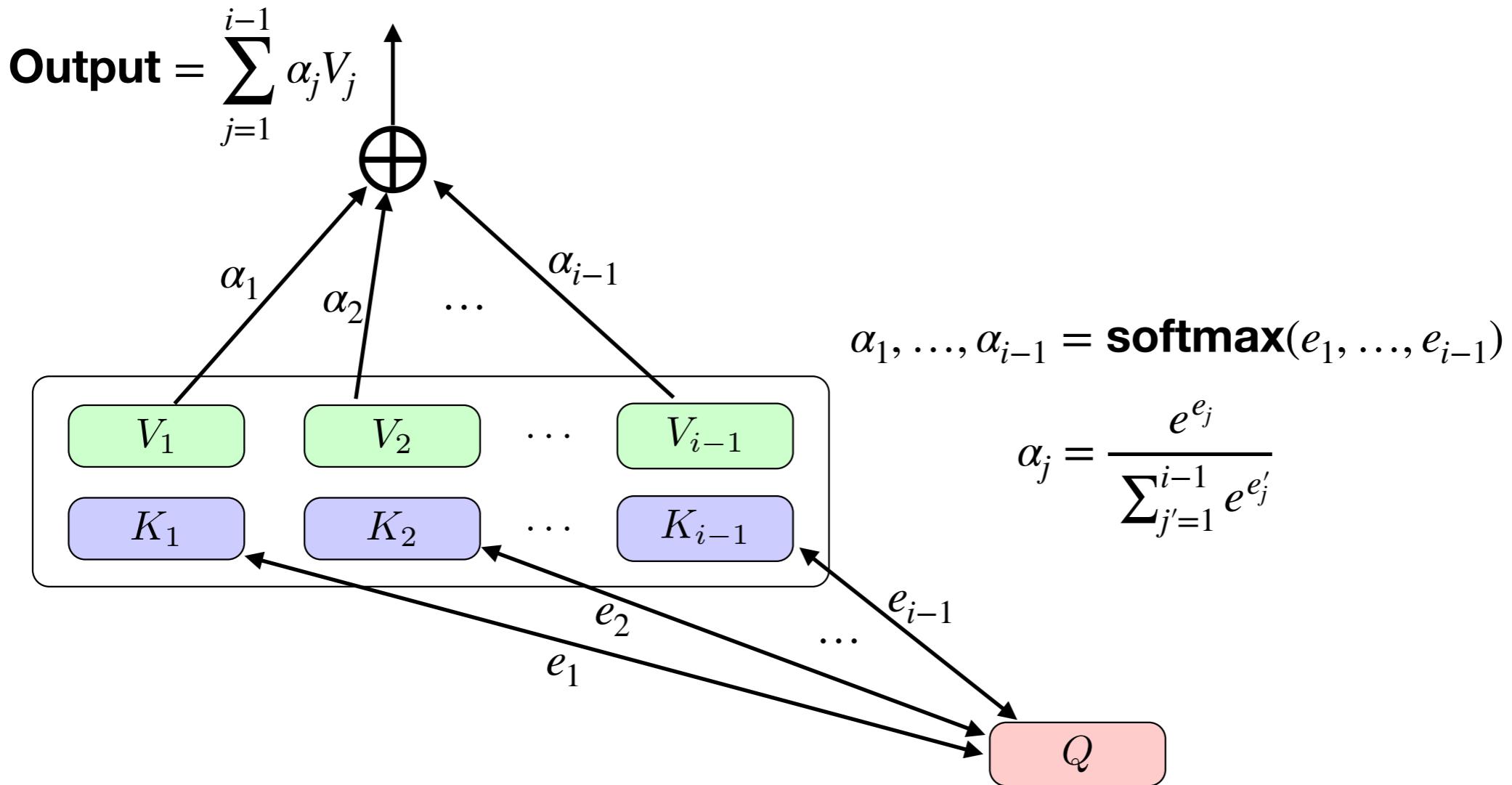


Attention function options:

$$e_j = \begin{cases} v \tanh [W_Q Q + W_K K_j] & \text{(Bahdanau et al., 2014)} \\ Q^T W K_j & \text{(Luong et al., 2015)} \\ \frac{Q^T K}{\sqrt{|K|}} & \text{(Vaswani et al., 2017)} \end{cases}$$

Neural Attention

Query, Key, and Value



Attention function options:

$$e_j = \begin{cases} v \tanh [W_Q Q + W_K K_j] & \text{(Bahdanau et al., 2014)} \\ Q^T W K_j & \text{(Luong et al., 2015)} \\ \frac{Q^T K}{\sqrt{|K|}} & \text{(Vaswani et al., 2017)} \end{cases}$$

A single masked attention "head"

x_1

x_2

x_3

w_1

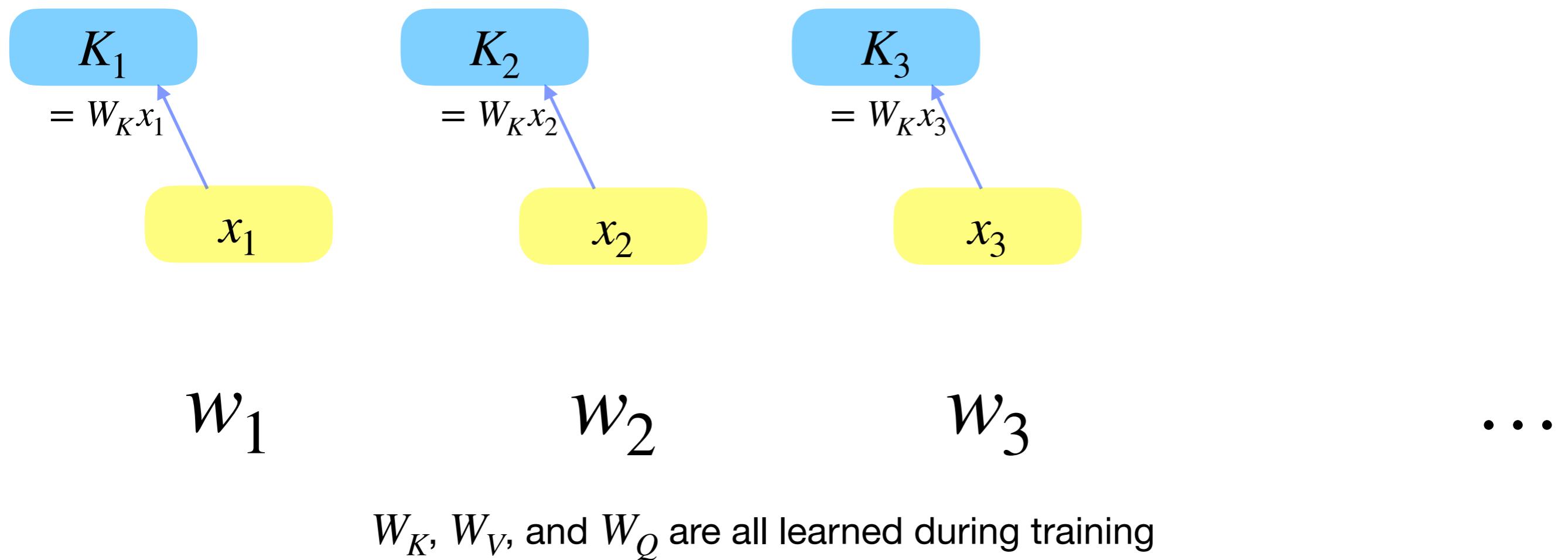
w_2

w_3

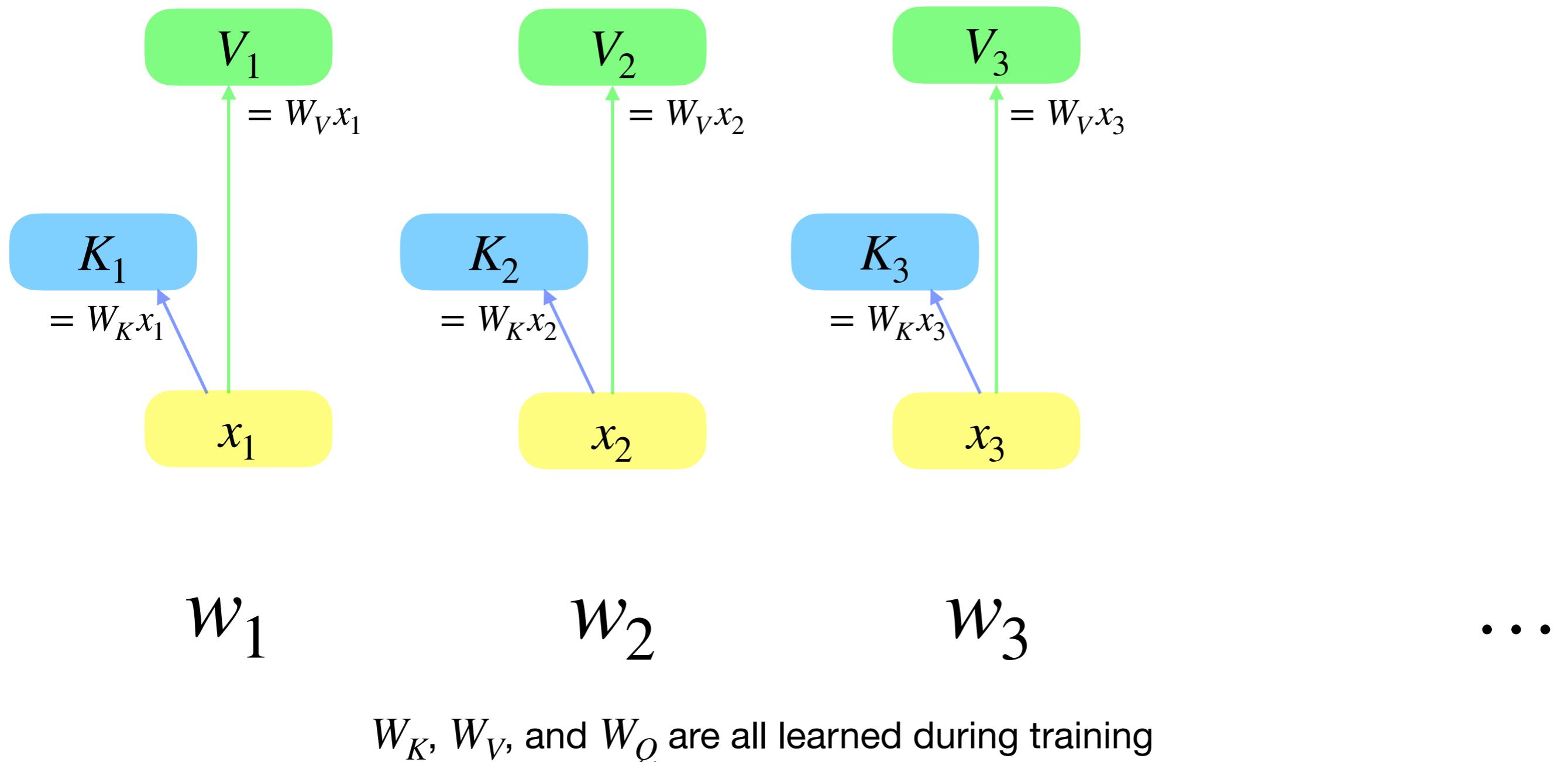
...

W_K , W_V , and W_Q are all learned during training

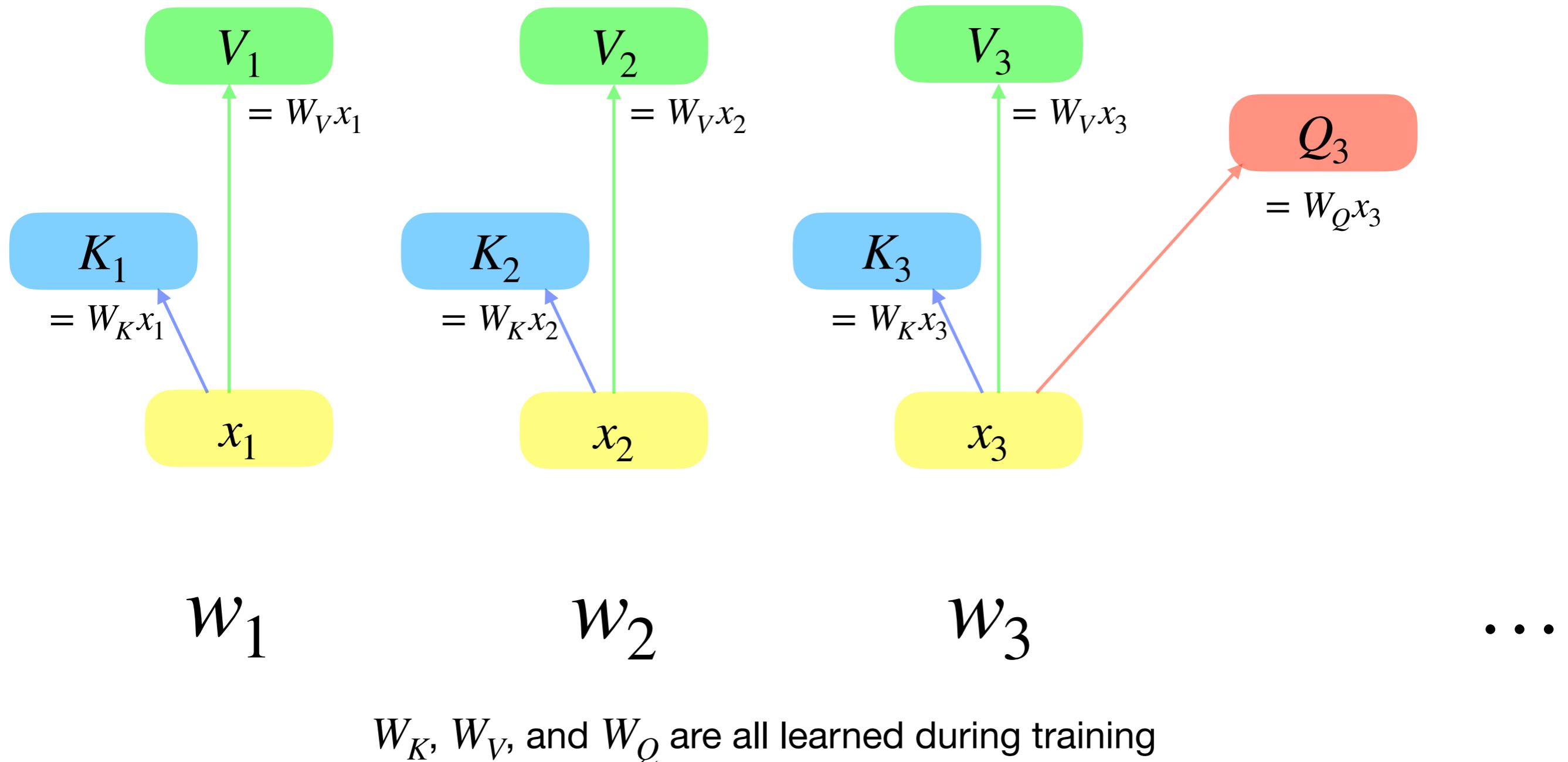
A single masked attention "head"



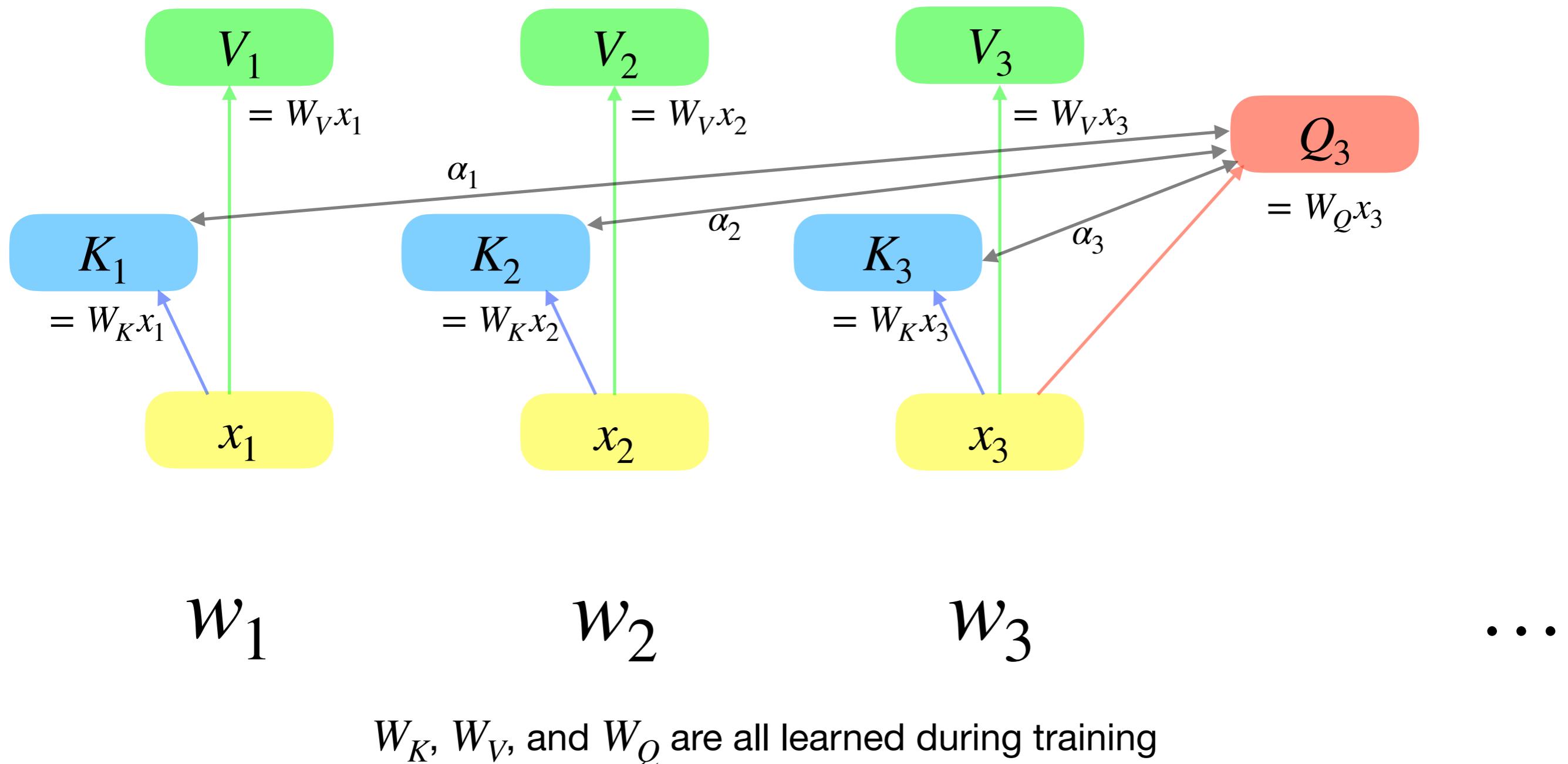
A single masked attention "head"



A single masked attention "head"

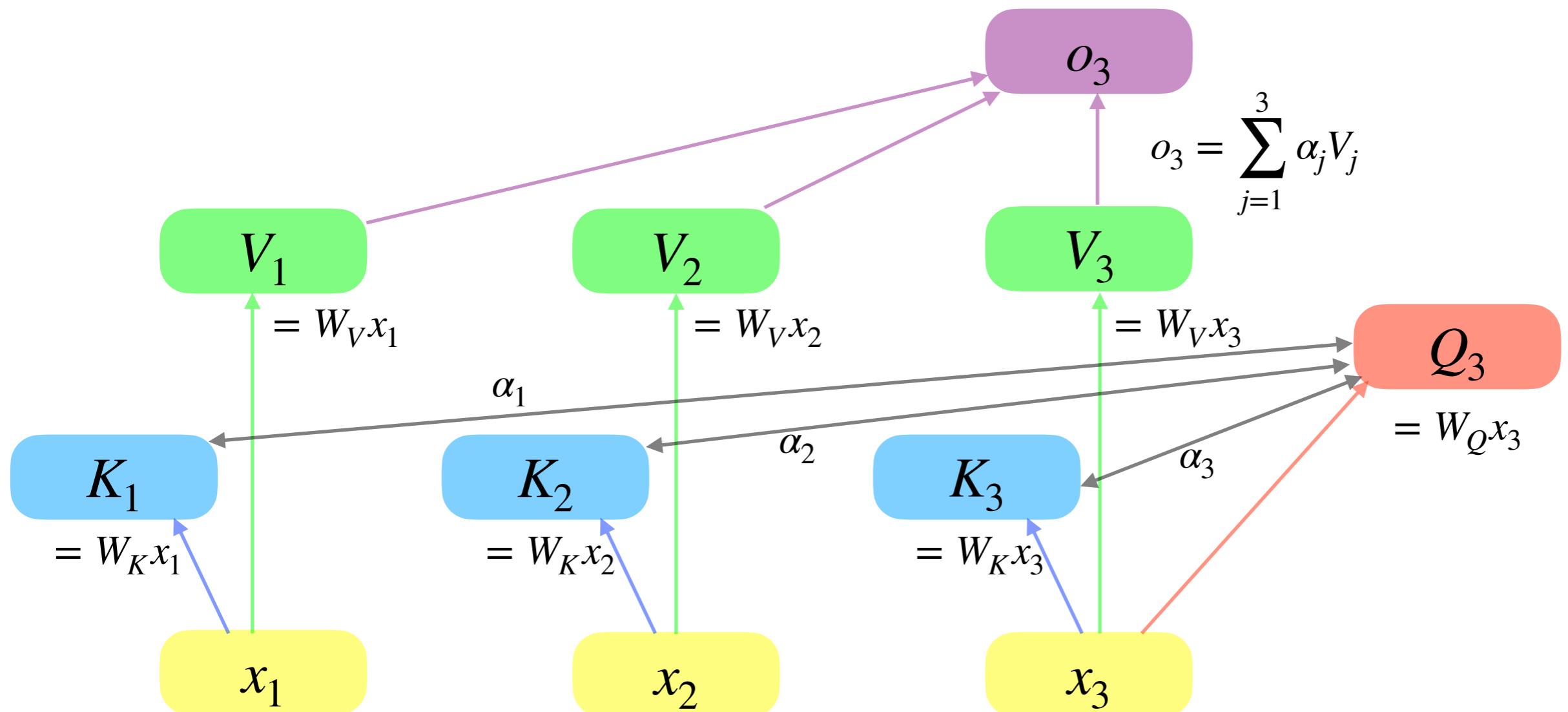


A single masked attention "head"



W_K , W_V , and W_Q are all learned during training

A single masked attention "head"



W_1

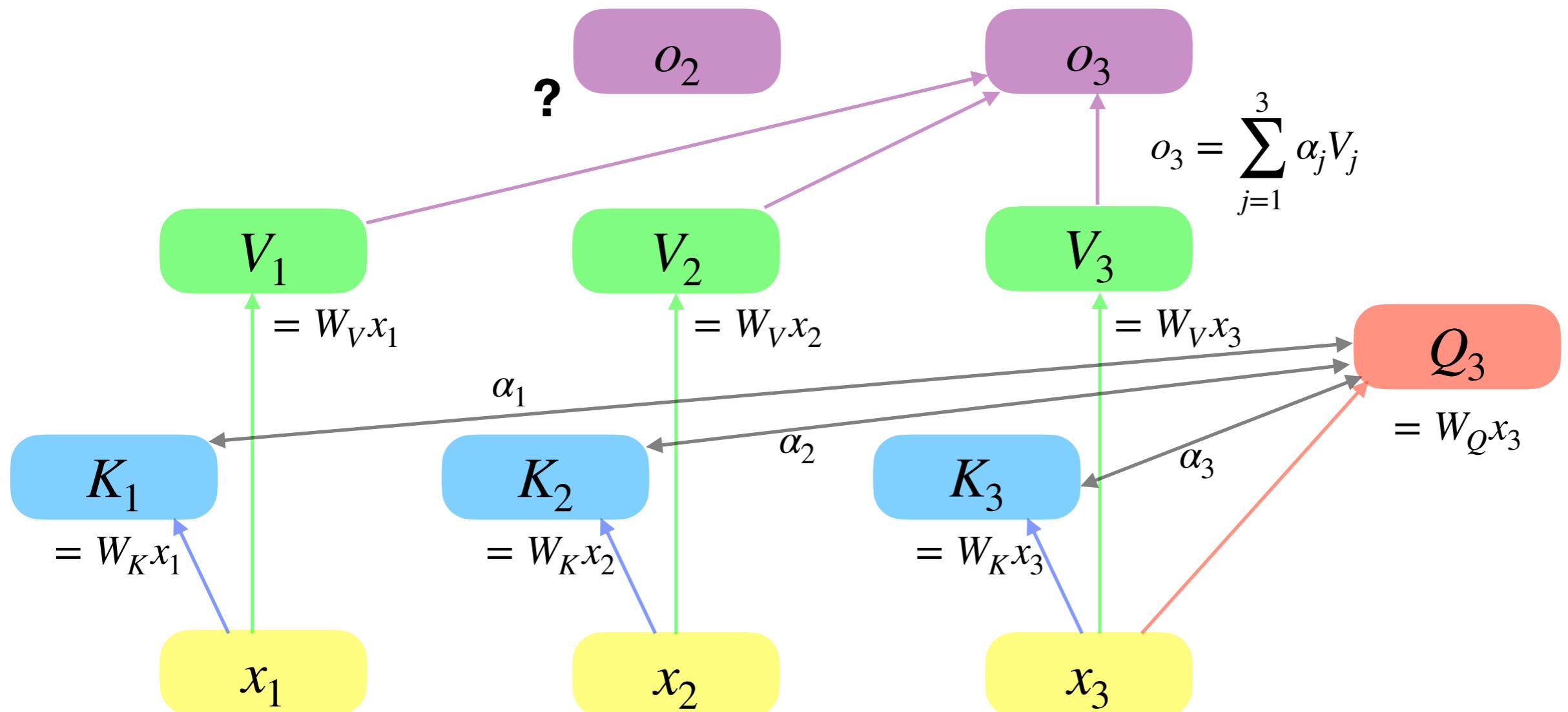
W_2

W_3

...

W_K , W_V , and W_Q are all learned during training

A single masked attention "head"



w_1

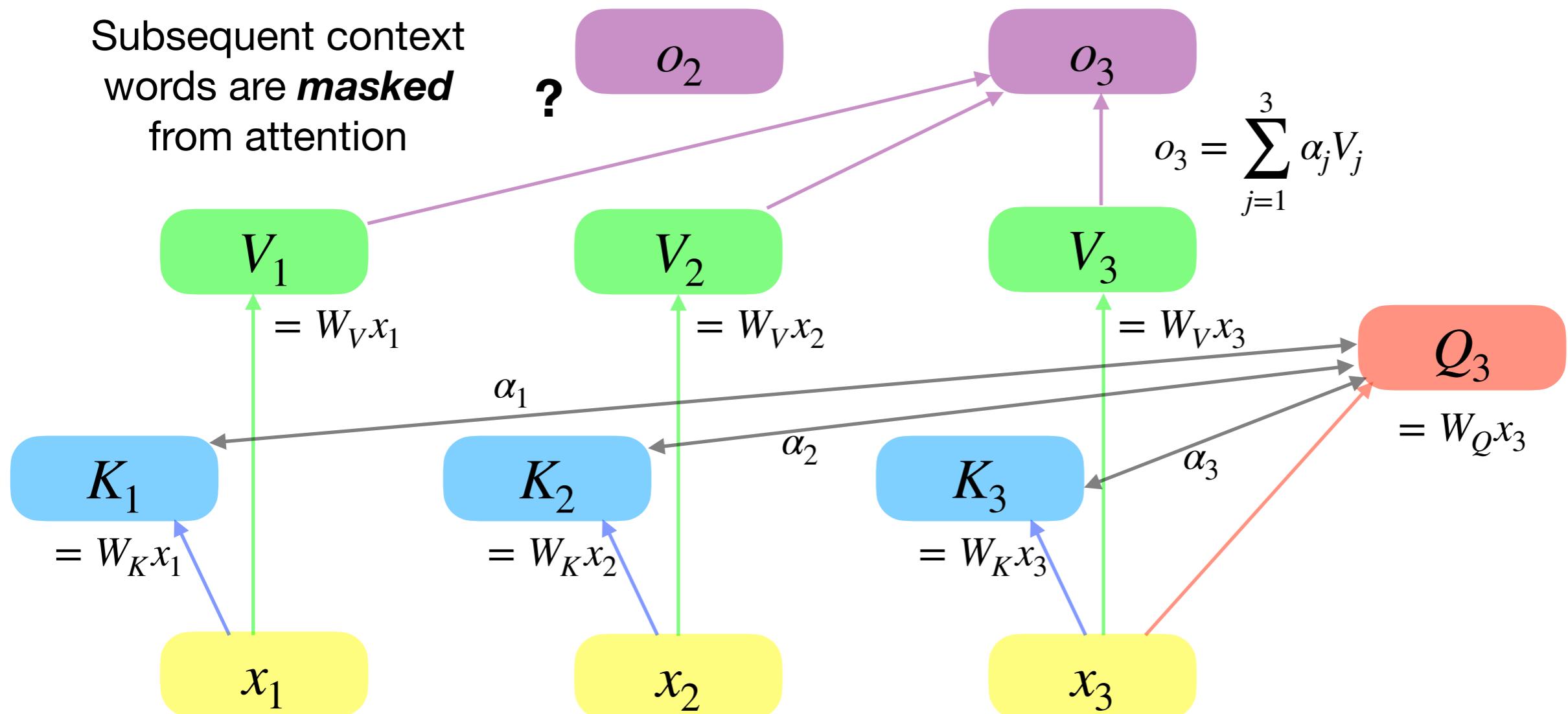
w_2

w_3

...

A single masked attention "head"

Subsequent context words are **masked** from attention



w_1

w_2

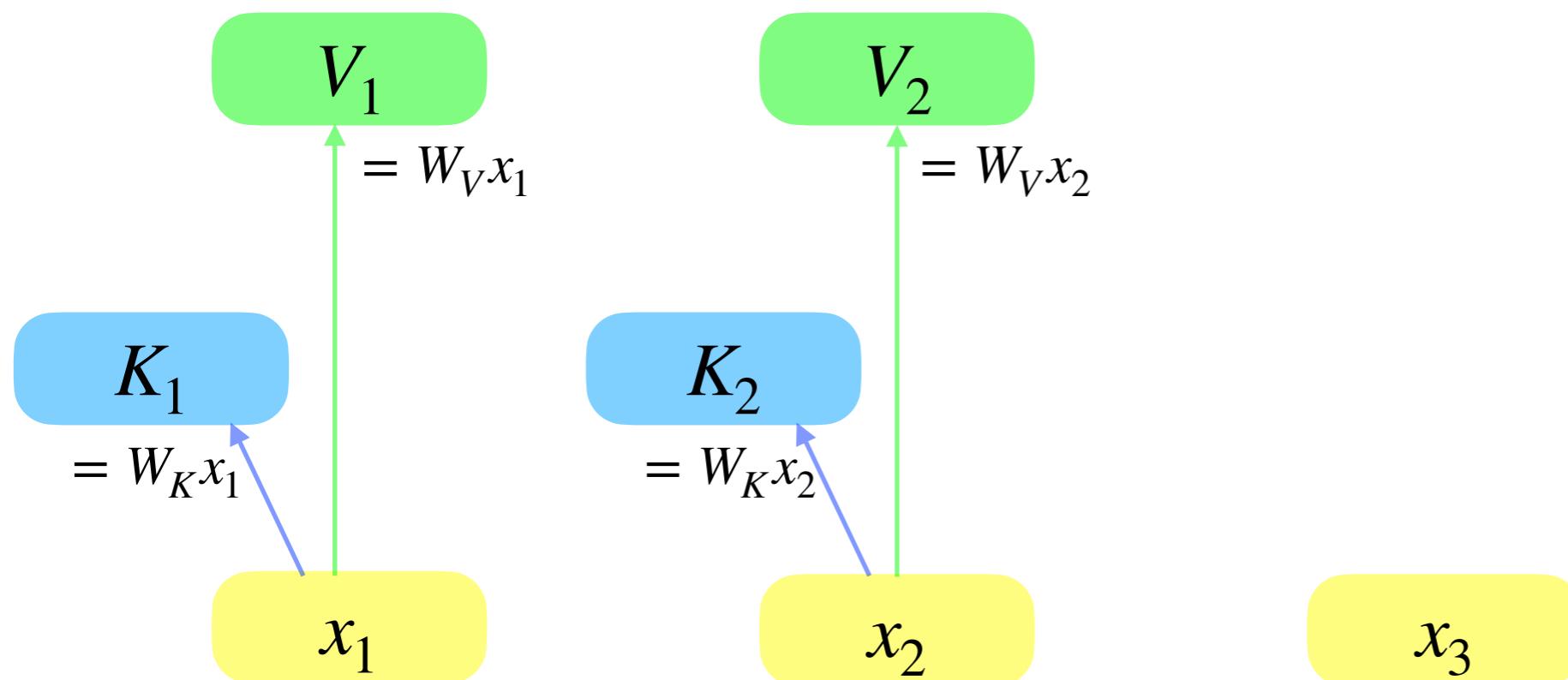
w_3

...

A single masked attention "head"

Subsequent context words are **masked** from attention

? o₂



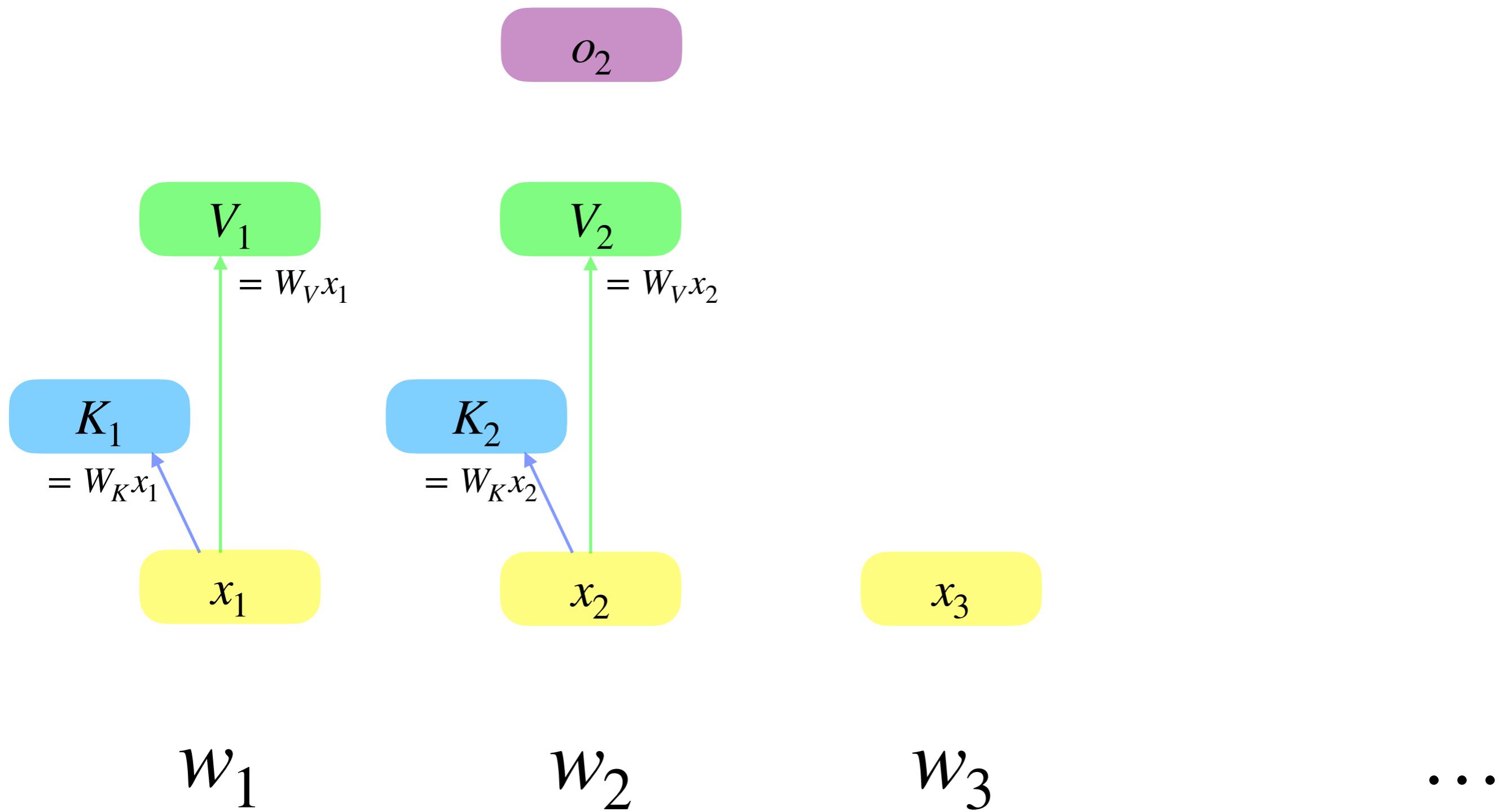
w_1

w_2

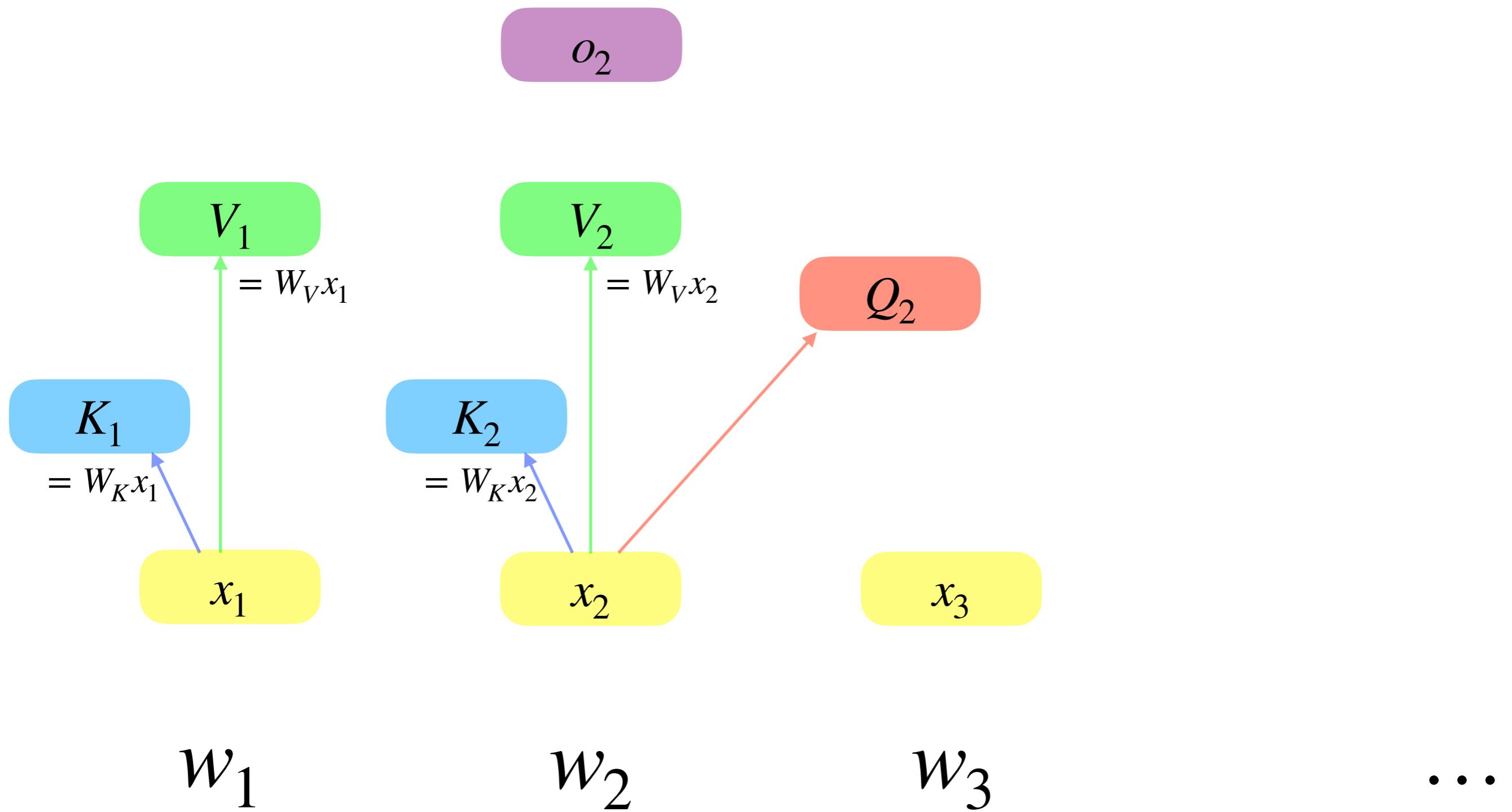
w_3

...

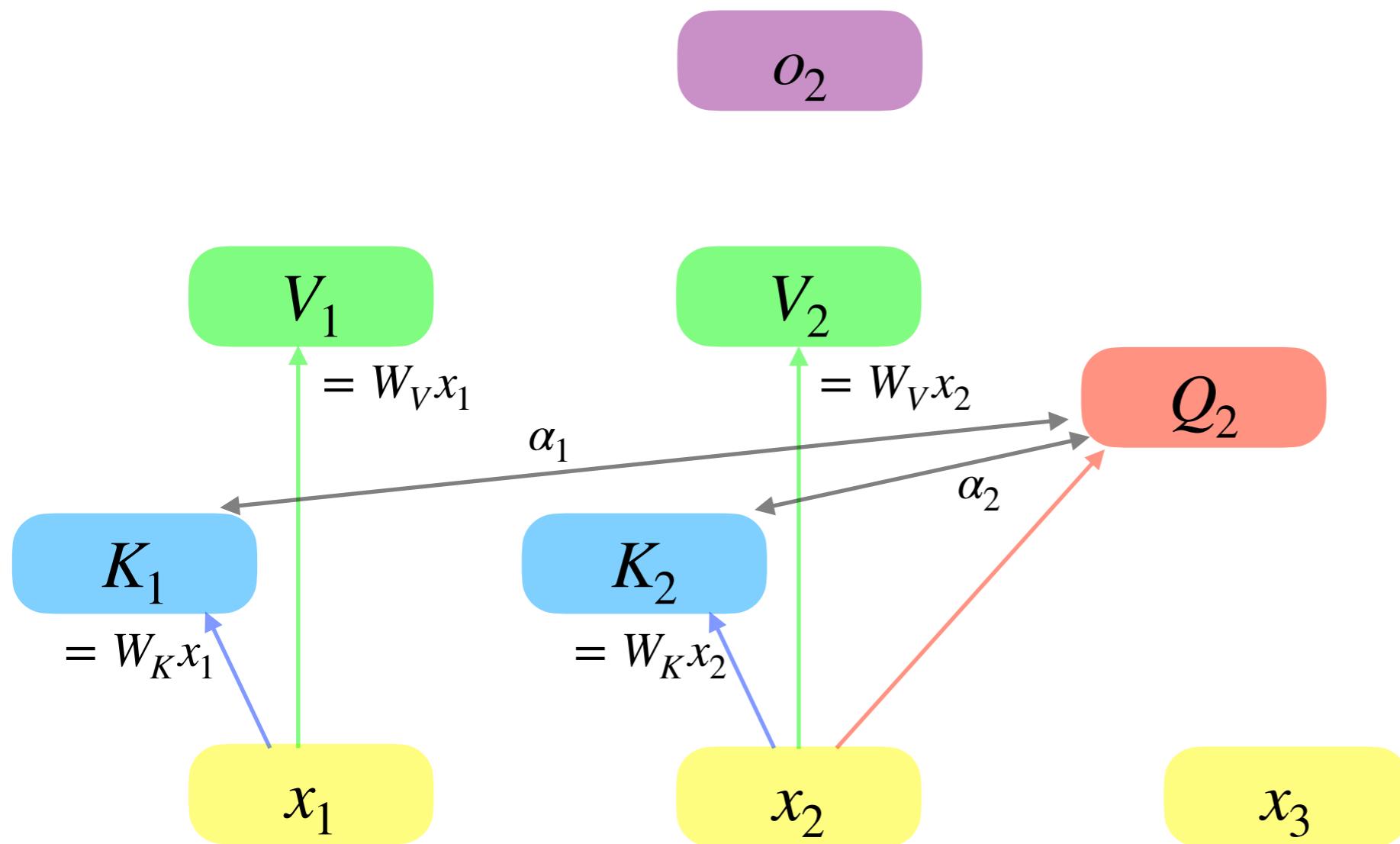
A single masked attention "head"



A single masked attention "head"



A single masked attention "head"



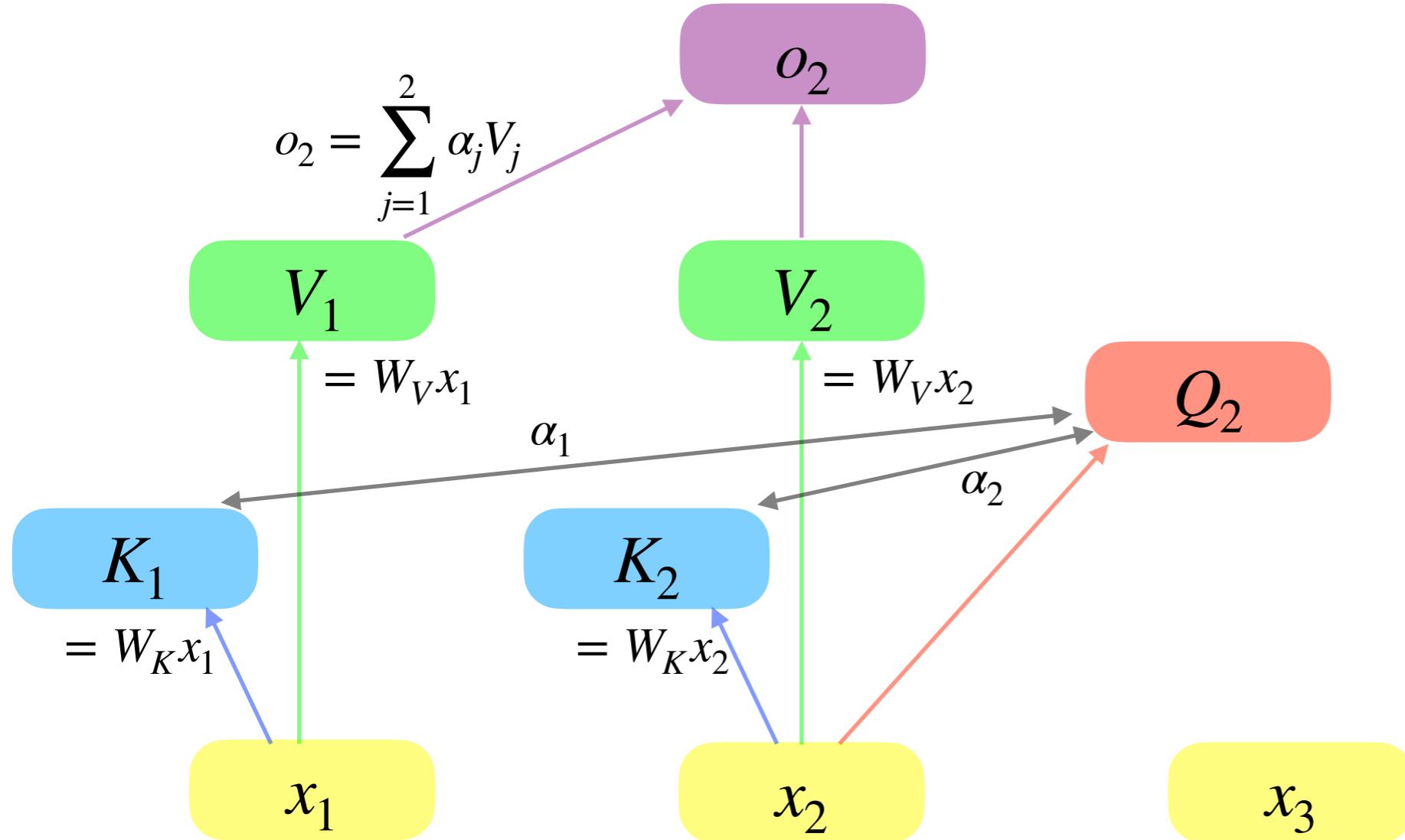
w_1

w_2

w_3

\dots

A single masked attention "head"



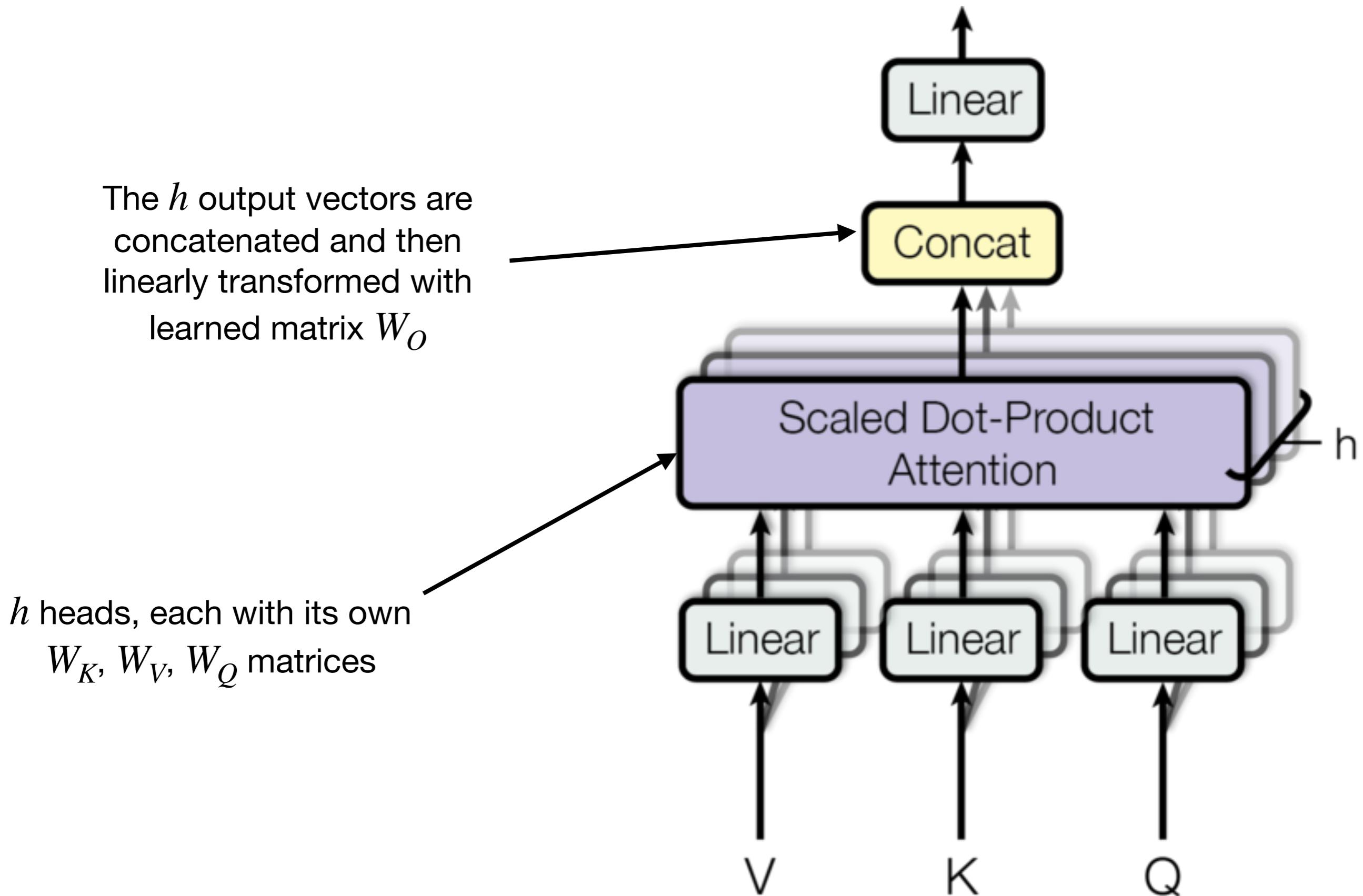
w_1

w_2

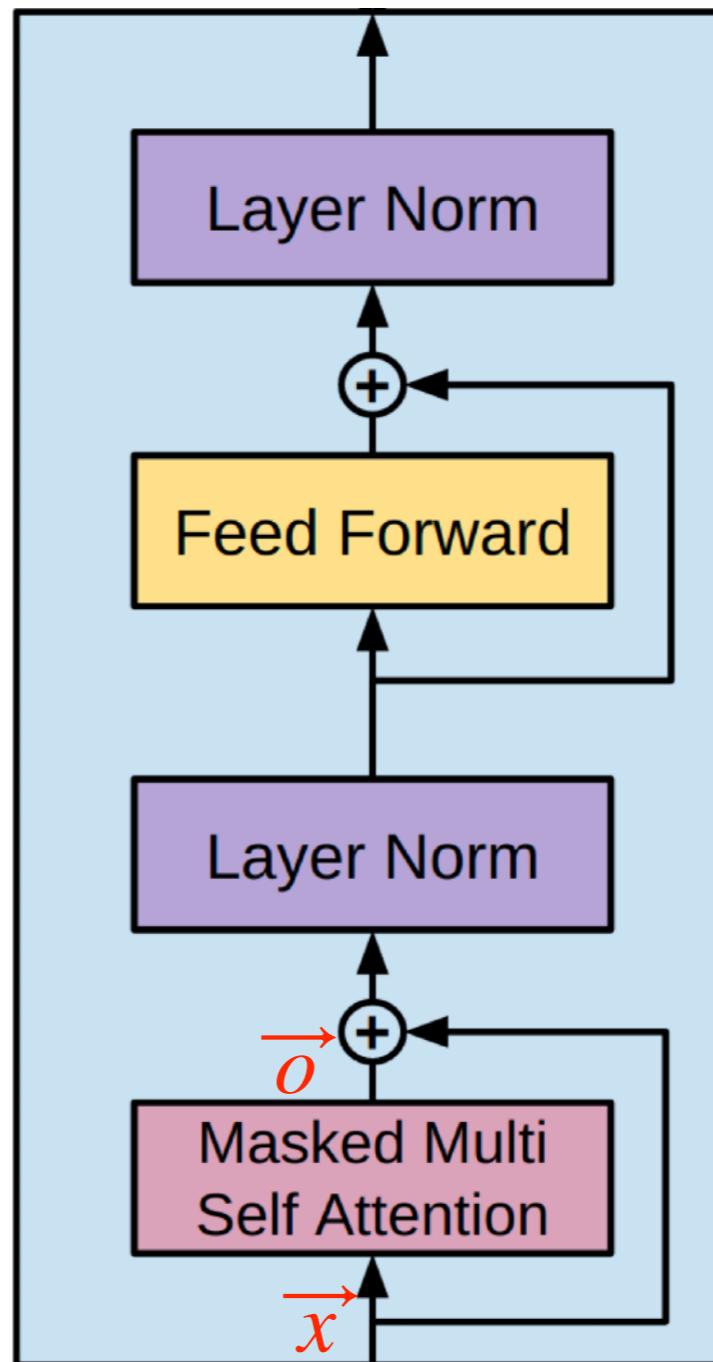
w_3

...

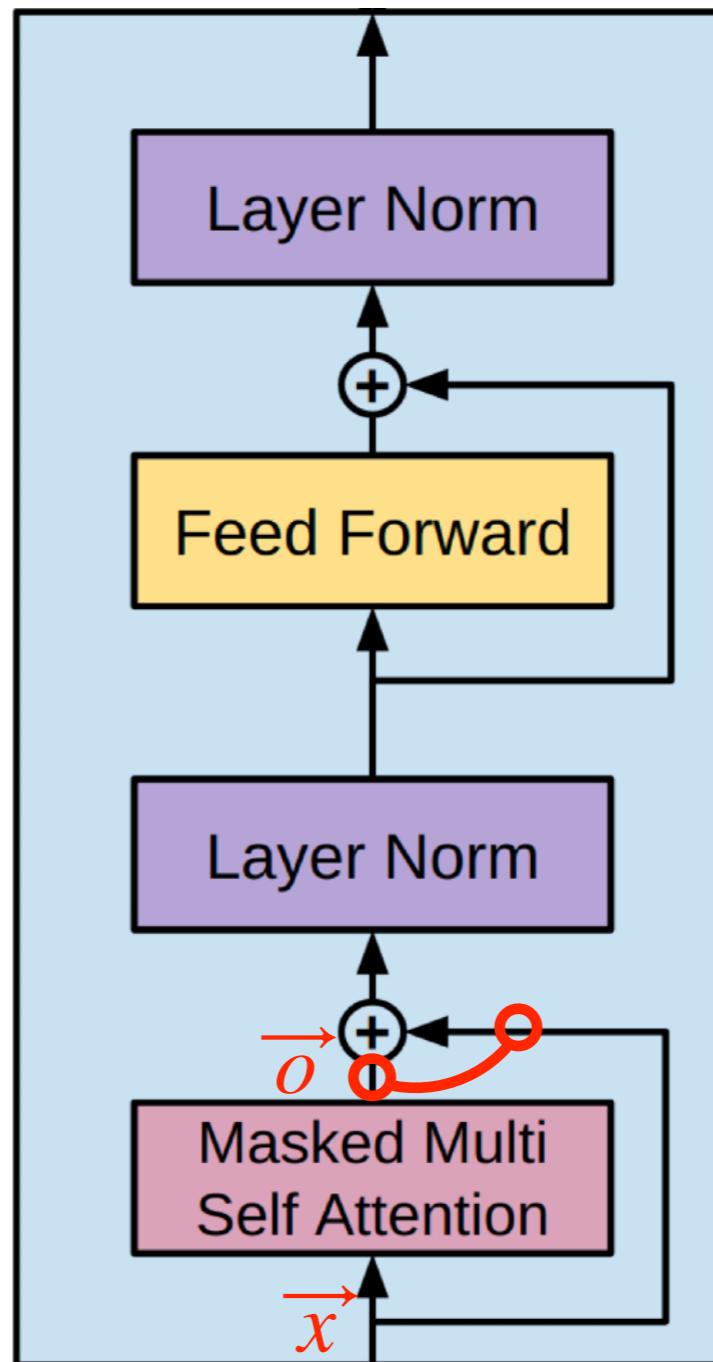
Multi-headed attention



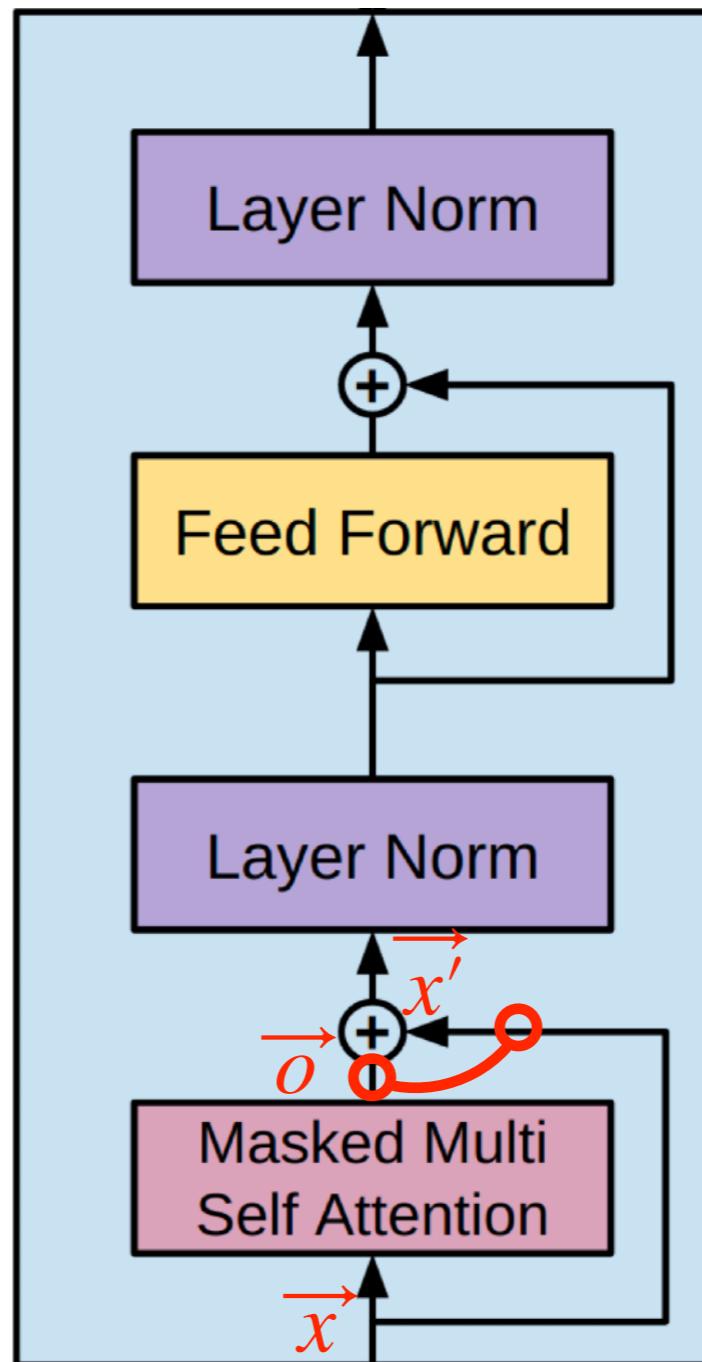
Residual connection & layer normalization



Residual connection & layer normalization

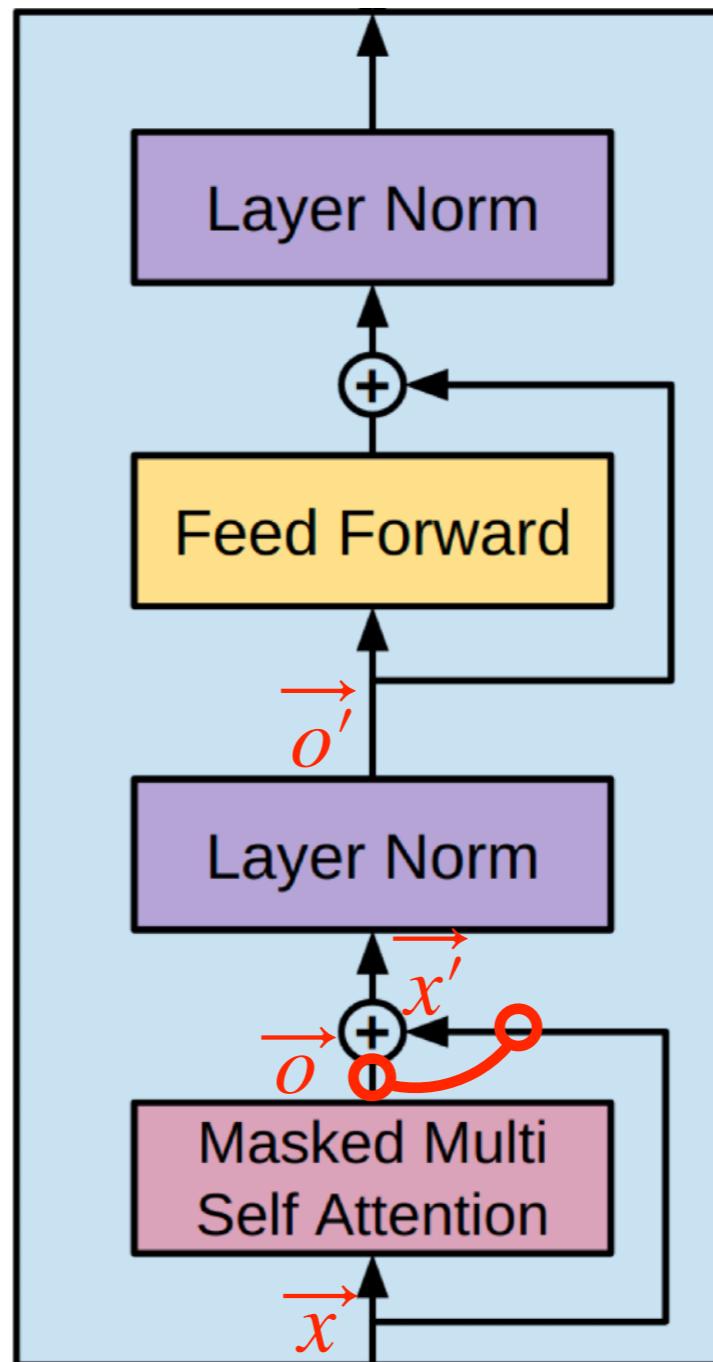


Residual connection & layer normalization



$$\vec{x}' = \vec{x} + \vec{o}$$

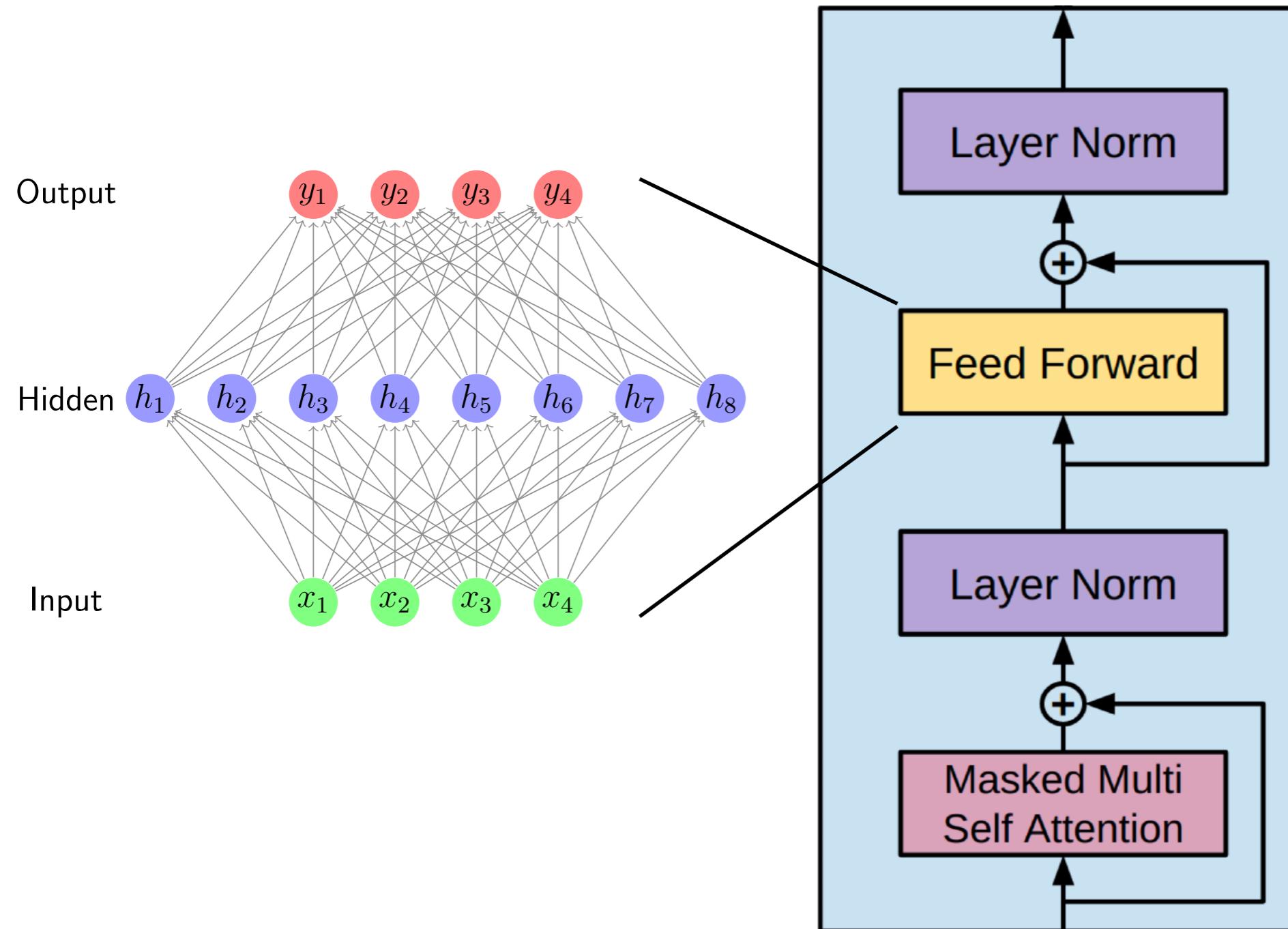
Residual connection & layer normalization



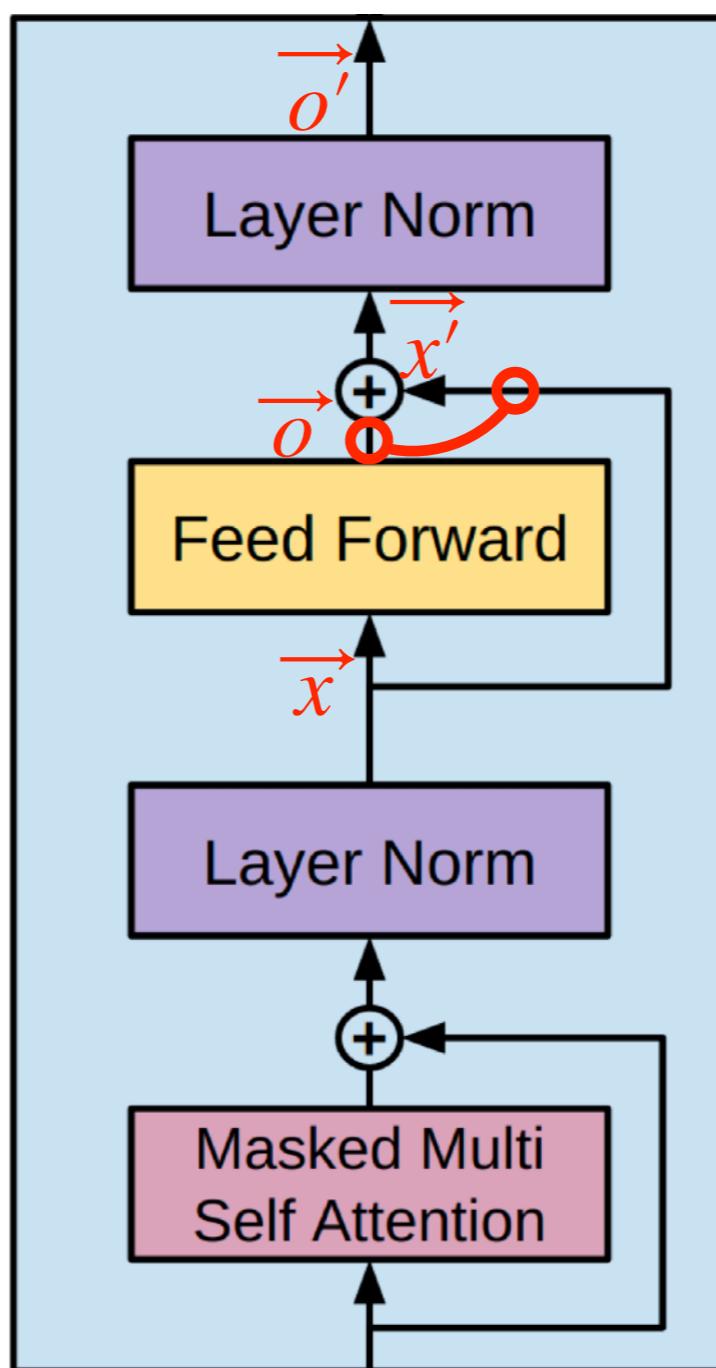
$$\vec{o}' = \frac{\vec{x}' - \text{Mean}(\vec{x}')}{\text{StdDev}(\vec{x}')}}$$

$$\vec{x}' = \vec{x} + \vec{o}$$

Feed-forward layer



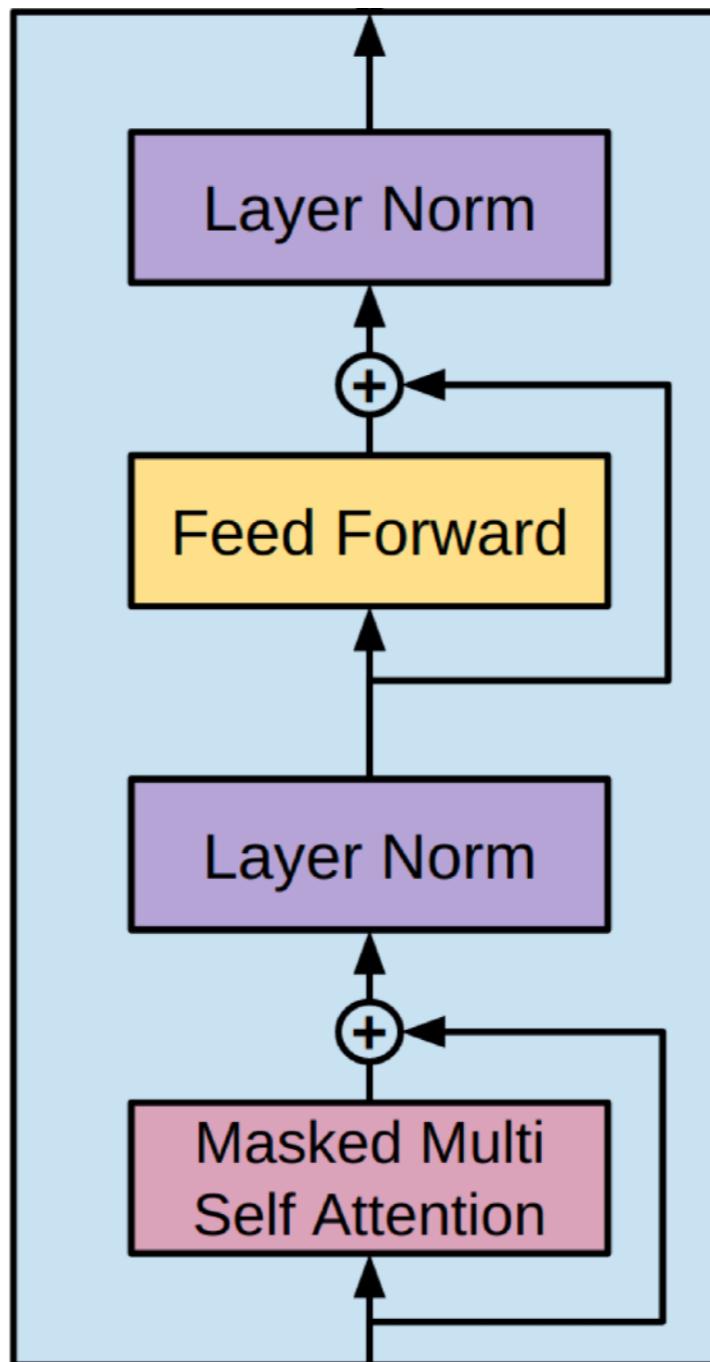
Res. connection & layer norm. (again)



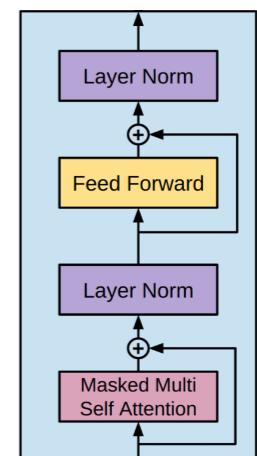
$$\vec{o}' = \frac{\vec{x}' - \text{Mean}(\vec{x}')}{\text{StdDev}(\vec{x}')}$$

$$\vec{x}' = \vec{x} + \vec{o}$$

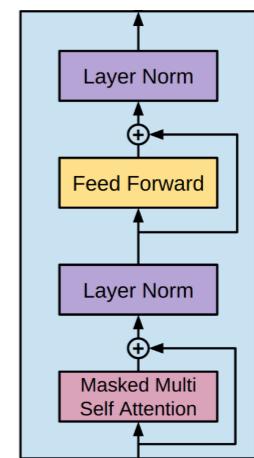
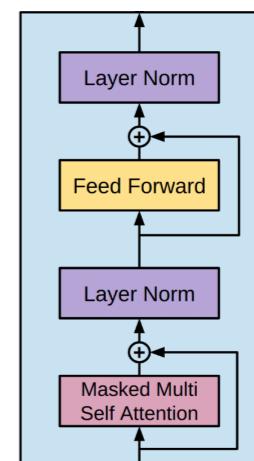
Res. connection & layer norm. (again)



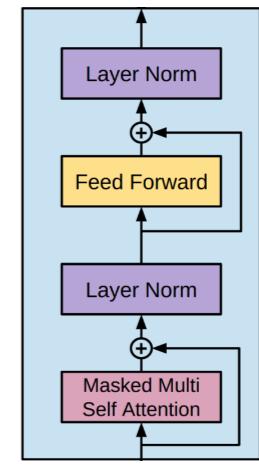
Res. connection & layer norm. (again)



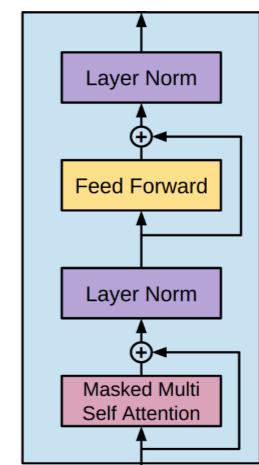
Res. connection & layer norm. (again)



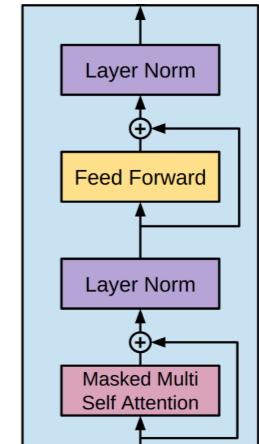
Res. connection & layer norm. (again)



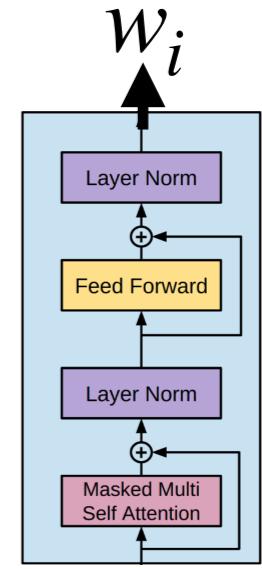
⋮



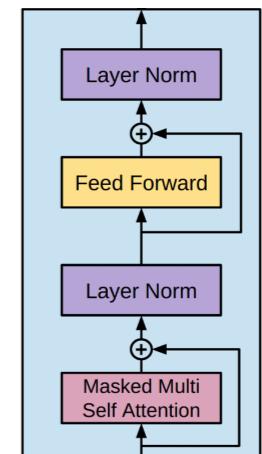
⋮



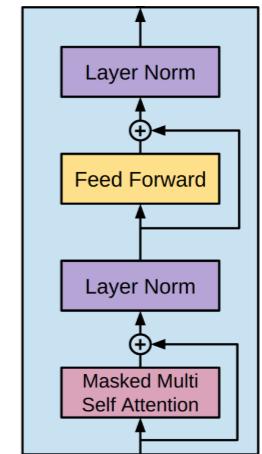
Res. connection & layer norm. (again)



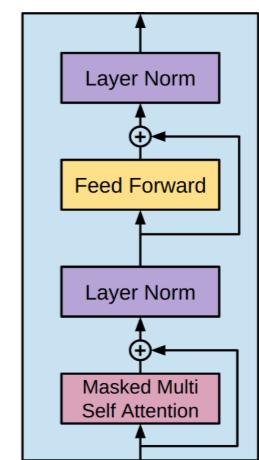
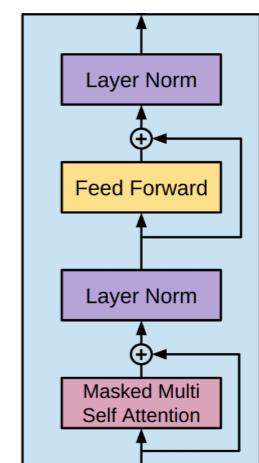
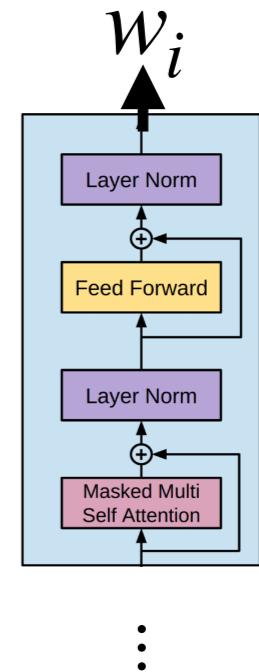
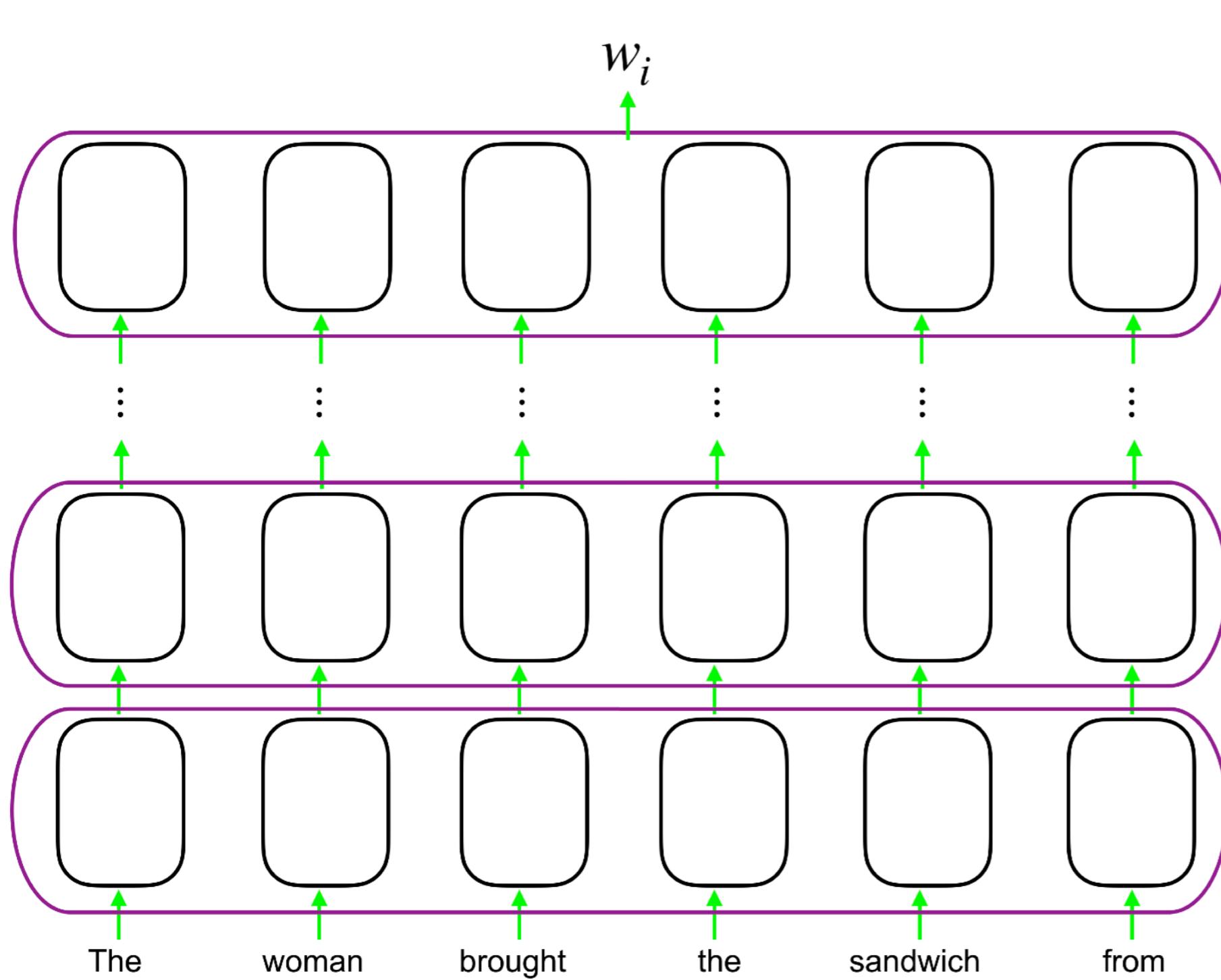
⋮



⋮



Res. connection & layer norm. (again)



Transformer + a huge corpus = ...?

New AI fake text generator may be too dangerous to release, say creators

► The Guardian

- OpenAI text-generating tool GPT2 won't be released for fear of misuse

► Business Insider

 [View Full Coverage](#)



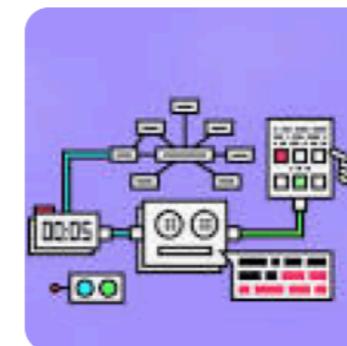
Feb 14, 2019 ▾

▼ The Verge

OpenAI has published the text-generating AI it said was too dangerous to share

GPT-2 is part of a new breed of text-generation systems that have impressed experts with their ability to generate coherent text from minimal ...

Nov 7, 2019



Write With Transformer |

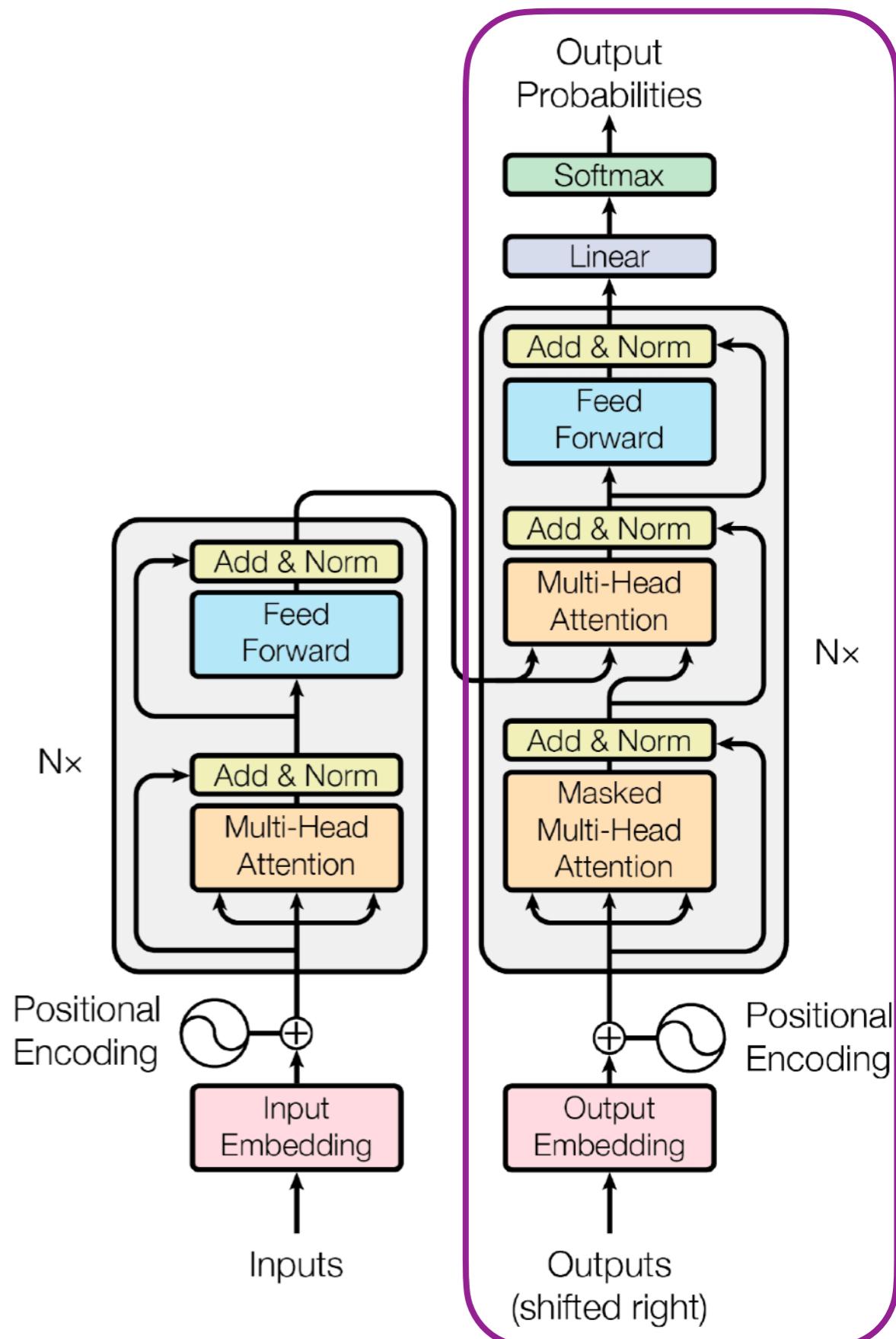
transformer.huggingface.co

Giant language model testing room: <http://gltr.io/dist/index.html>

Papers to read to understand GPT-2

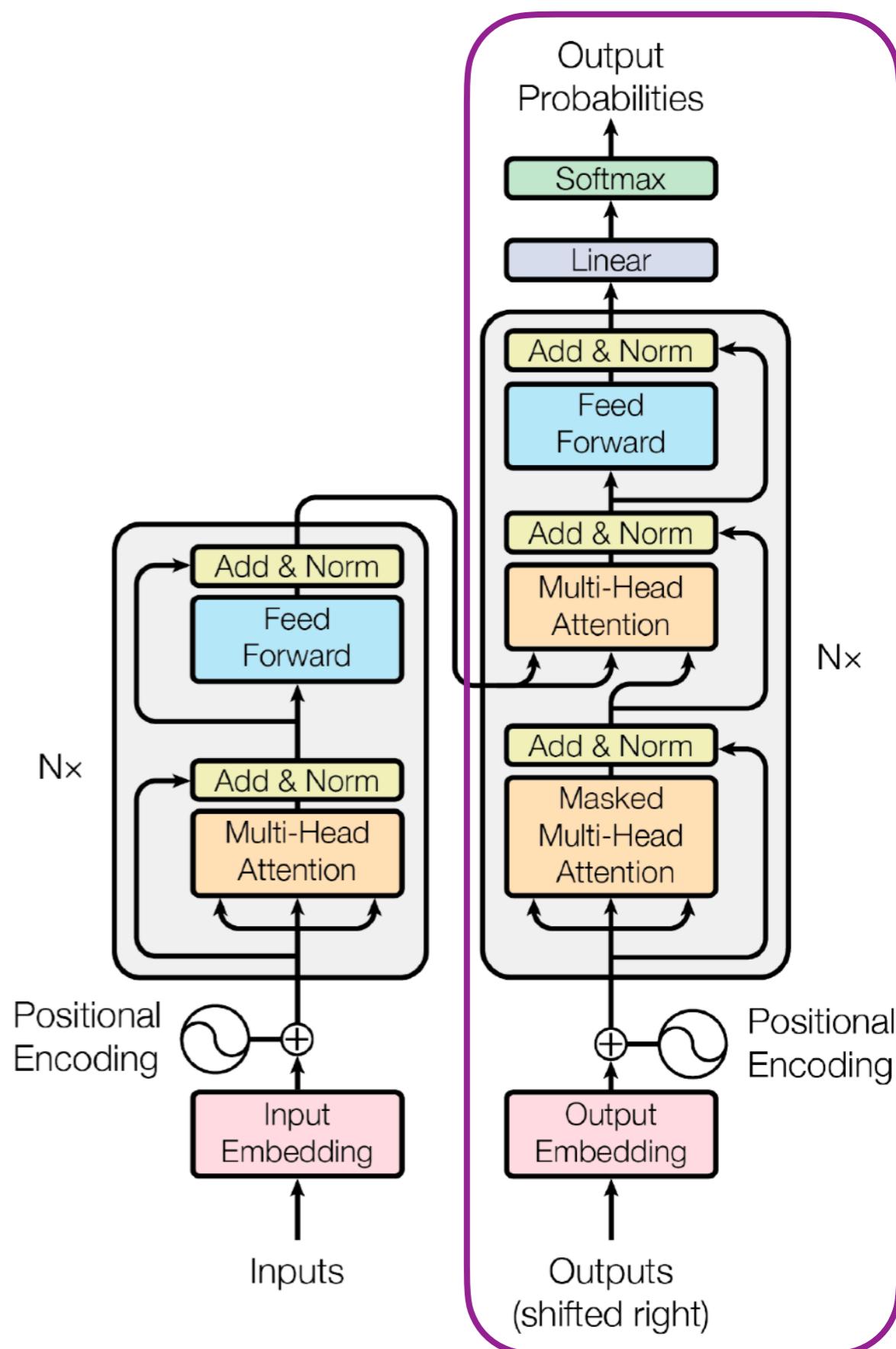
- Radford et al. (2019): the GPT-2 paper itself
- Radford et al. (2018): the GPT architecture, mostly shared by GPT-2
- Liu et al. (2018): the Transformer decoder
- Vaswani et al. (2017): the original Transformer paper
- Ba et al. (2016): layer normalization

The full Transformer model



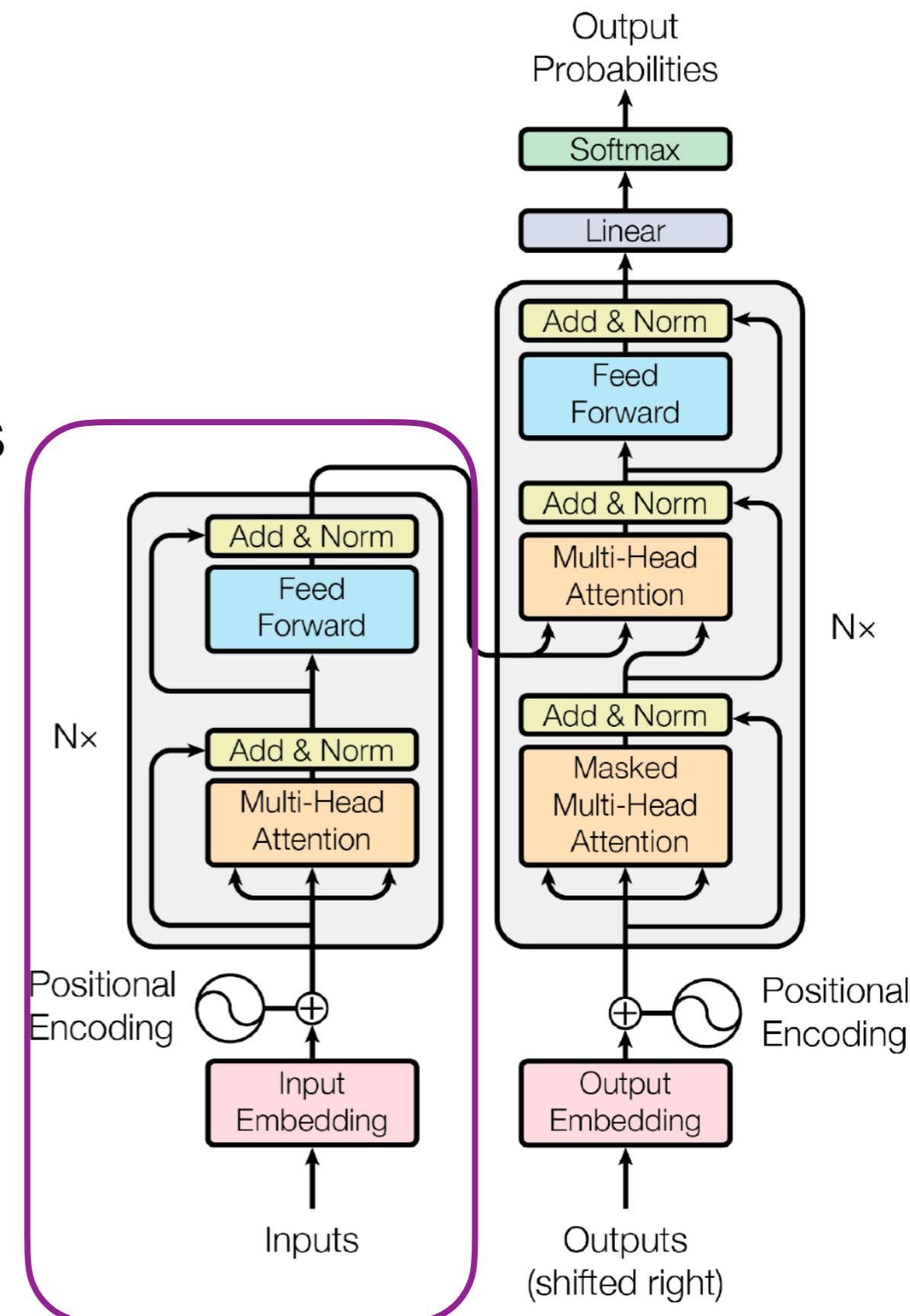
The full Transformer model

- In ML/NLP, the model we just studied is called the ***Transformer decoder***



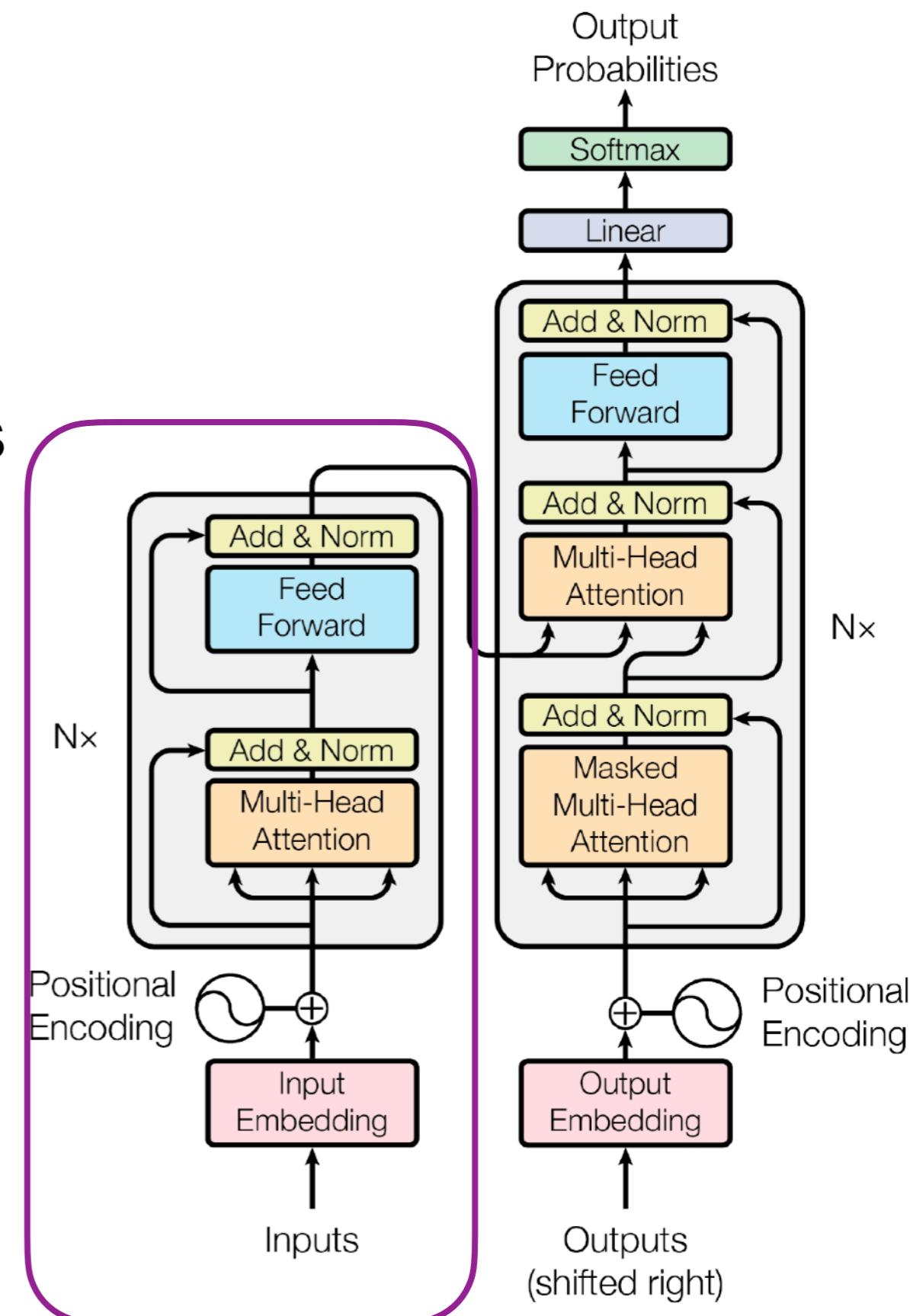
The full Transformer model

- In ML/NLP, the model we just studied is called the ***Transformer decoder***
- Sometimes, the Transformer is conditioned on a string that doesn't itself get predicted—this is called the ***encoder***



The full Transformer model

- In ML/NLP, the model we just studied is called the ***Transformer decoder***
- Sometimes, the Transformer is conditioned on a string that doesn't itself get predicted—this is called the ***encoder***
- Only difference: in encoder, attention is over the ***entire string***, not just words to the left



The full Transformer model

- In ML/NLP, the model we just studied is called the ***Transformer decoder***
- Sometimes, the Transformer is conditioned on a string that doesn't itself get predicted—this is called the ***encoder***
- Only difference: in encoder, attention is over the ***entire string***, not just words to the left
- BERT = Transformer encoder!

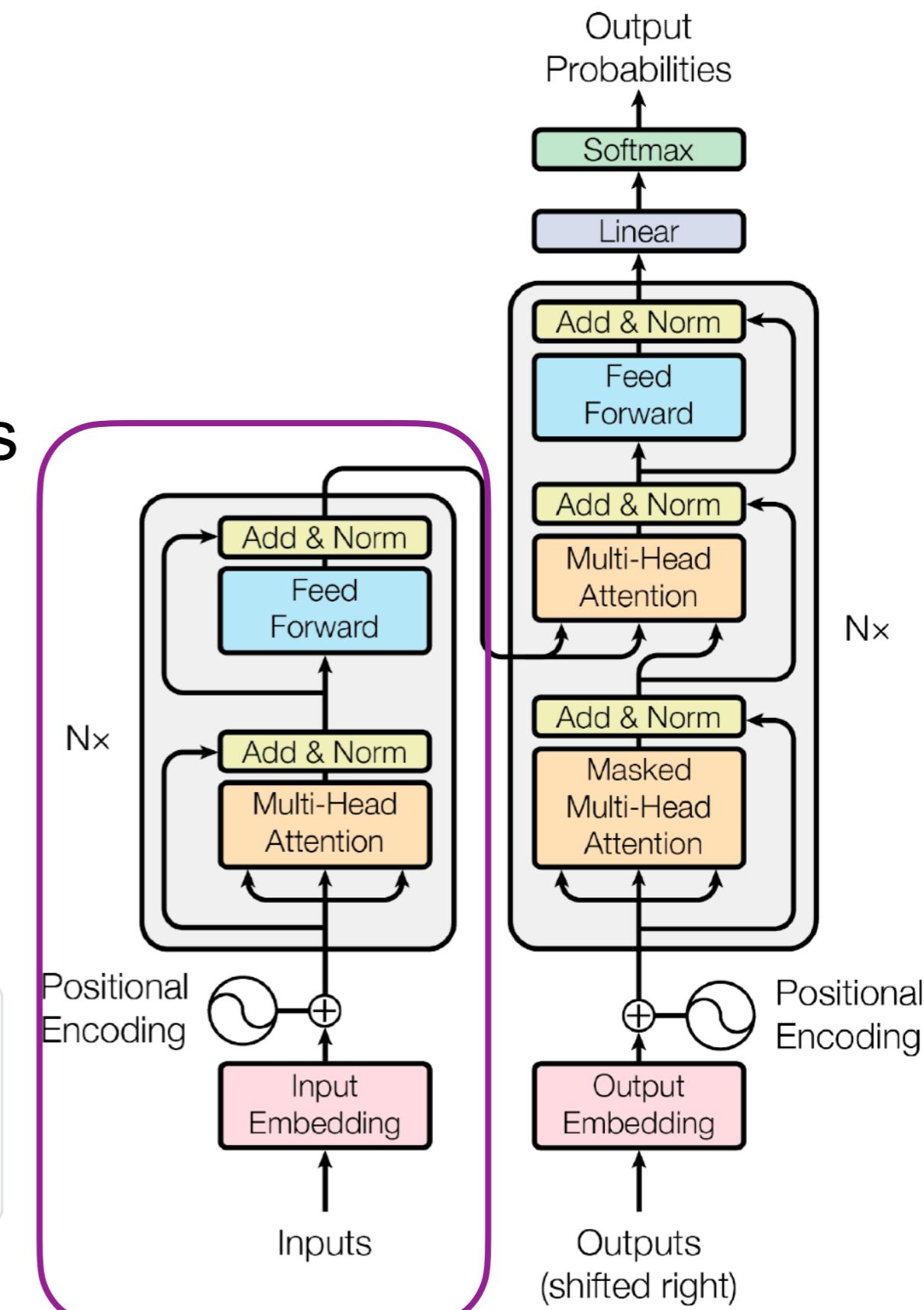
Google has updated its search algorithm: Say hello to BERT

SmartCompany.com.au · Nov 4



(Devlin et al., 2018)

(Vaswani et al., 2017)



GPT-2 on targeted syntax testing

syntaxgym.org

The screenshot shows the homepage of SyntaxGym. On the left is a blue sidebar with the title "SYNTAXGYM" and icons for Dashboard, Test suites, Language models, Visualizations, About, and Documentation. A "See more" button is at the bottom. The main content area features the SyntaxGym logo (a kettlebell with a tree diagram) and the text: "SyntaxGym is a unified platform where language and NLP researchers can design psycholinguistic tests and visualize the performance of language models. Our goal is to make psycholinguistic assessment of language models more standardized, reproducible, and accessible to a wide variety of researchers. The project is run out of the [MIT Computational Psycholinguistics Laboratory](#)". Below this are three cards: "TEST SUITES" (interested in viewing or designing psycholinguistic test suites? Create a new test suite online or upload one as a `.json` file. [See more →](#)), "LANGUAGE MODELS" (Have a model you want to evaluate? Add a model as a Docker container, and it will automatically be evaluated on existing test suites. [See more →](#)), and "VISUALIZATIONS" (Want to compare the results of different models across test suites? Visualize model performance through interactive charts. [See more →](#)). At the bottom, a note says: "Not sure where to start? [Read more](#) or take a look at the [documentation](#)."

Filler–gap dependencies



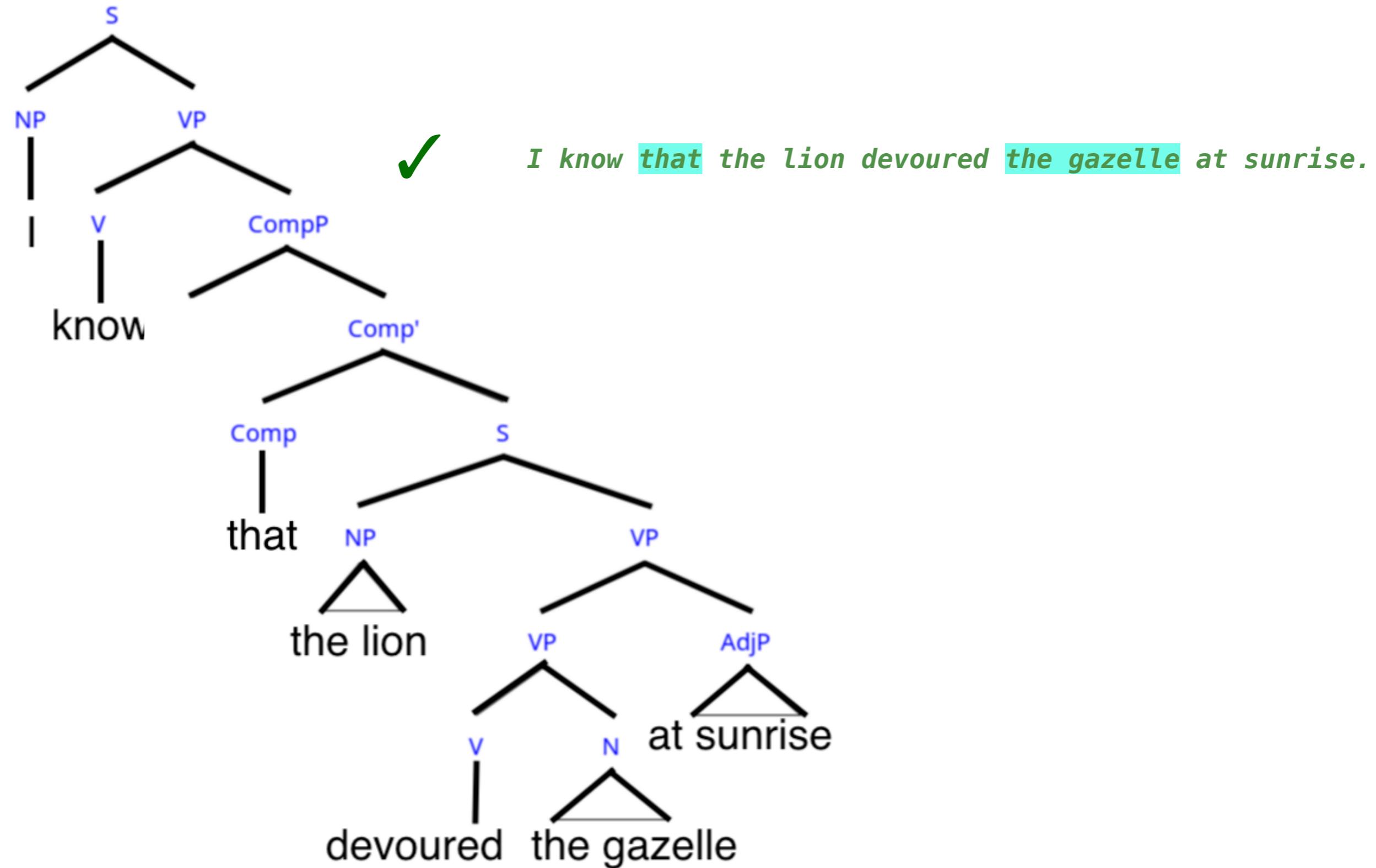
I know that the lion devoured the gazelle at sunrise.

Filler–gap dependencies

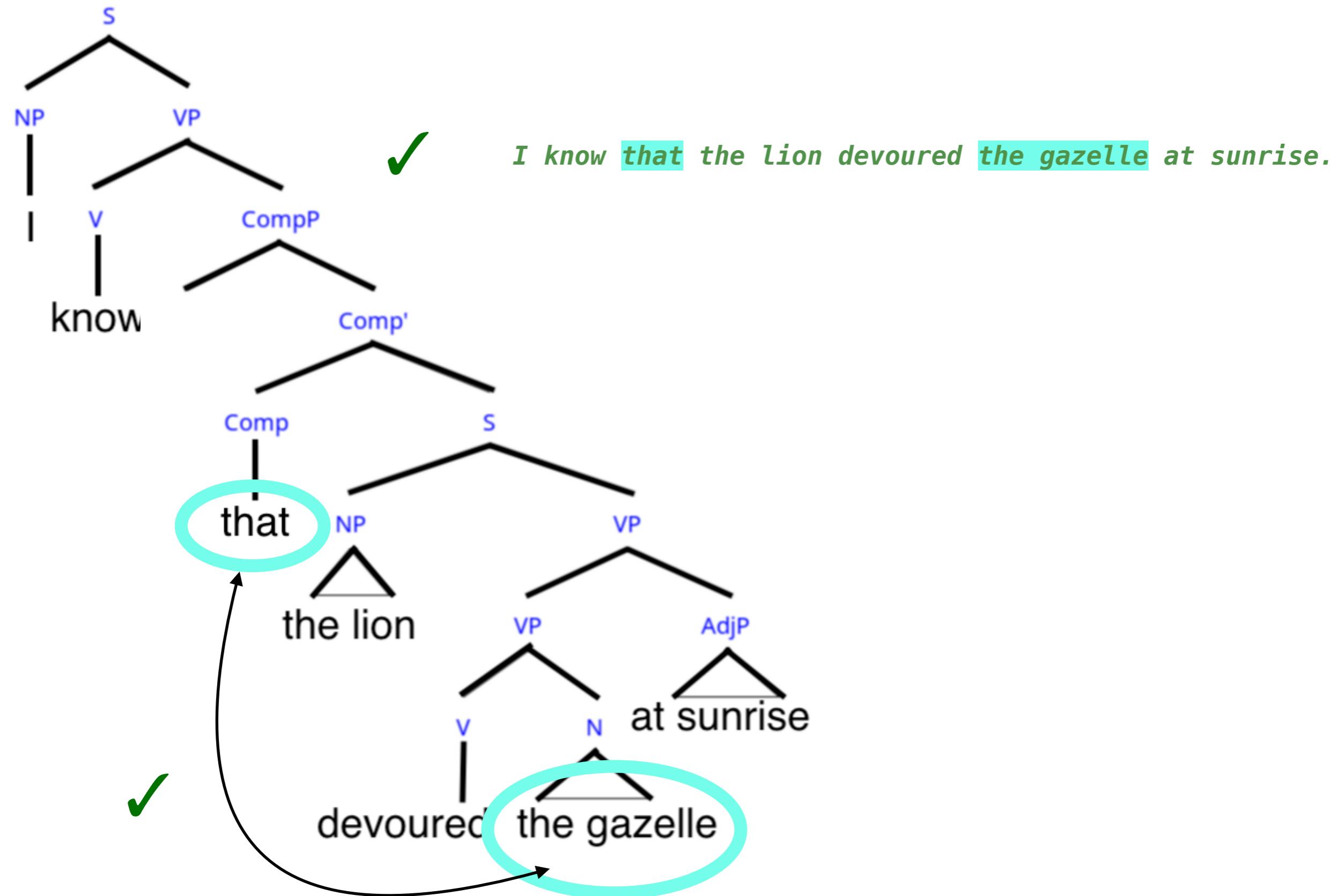


I know that the lion devoured the gazelle at sunrise.

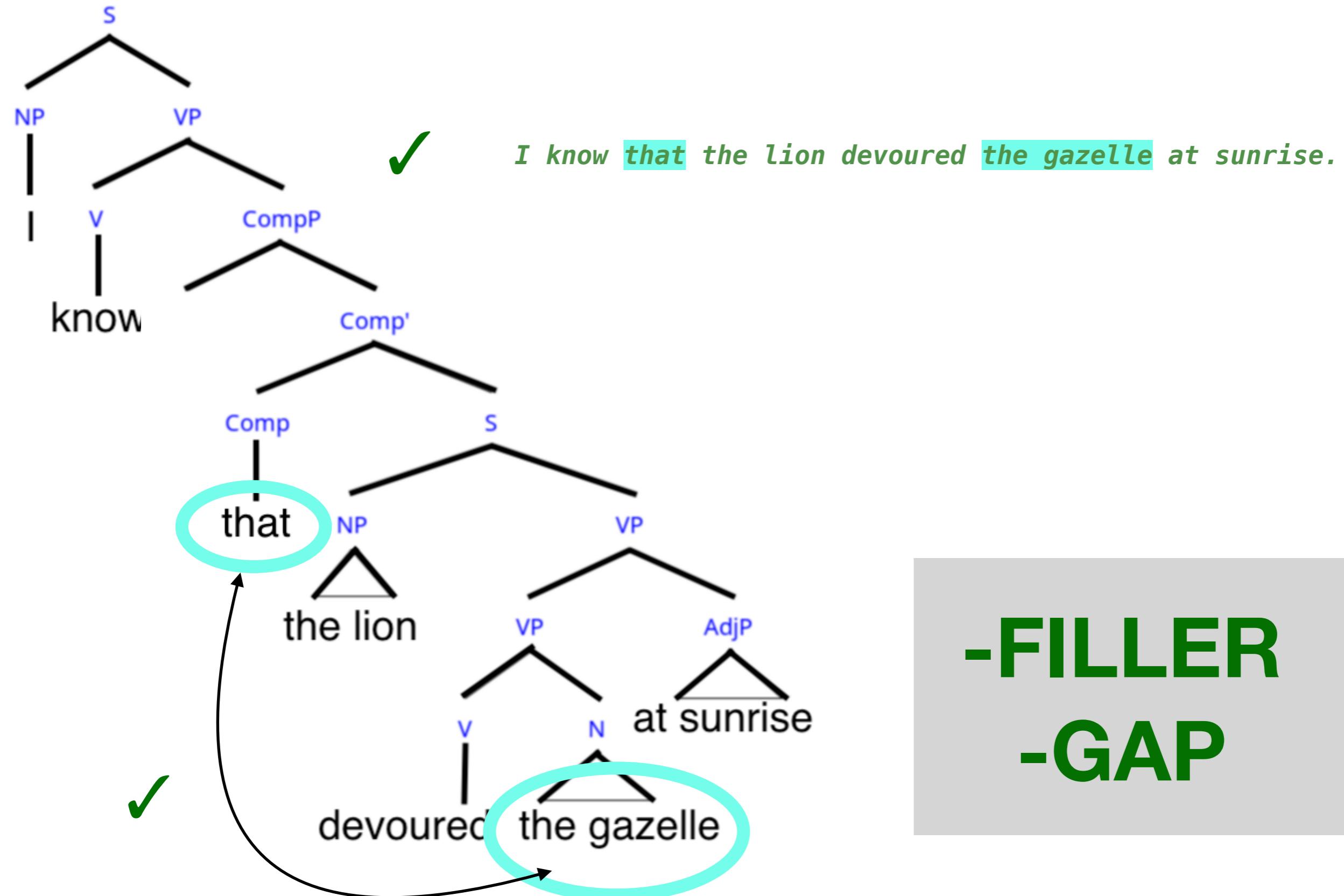
Filler–gap dependencies



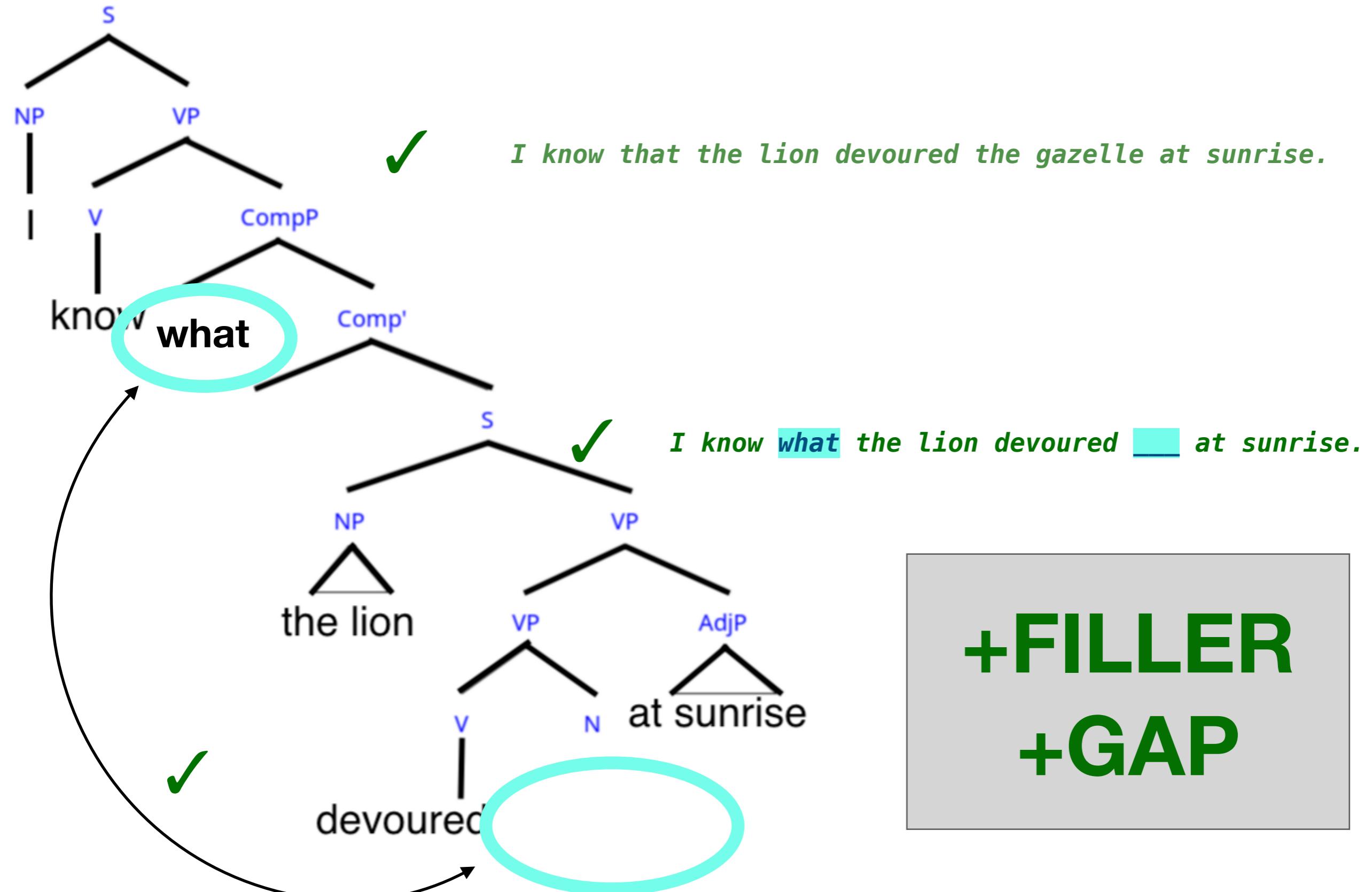
Filler–gap dependencies



Filler–gap dependencies

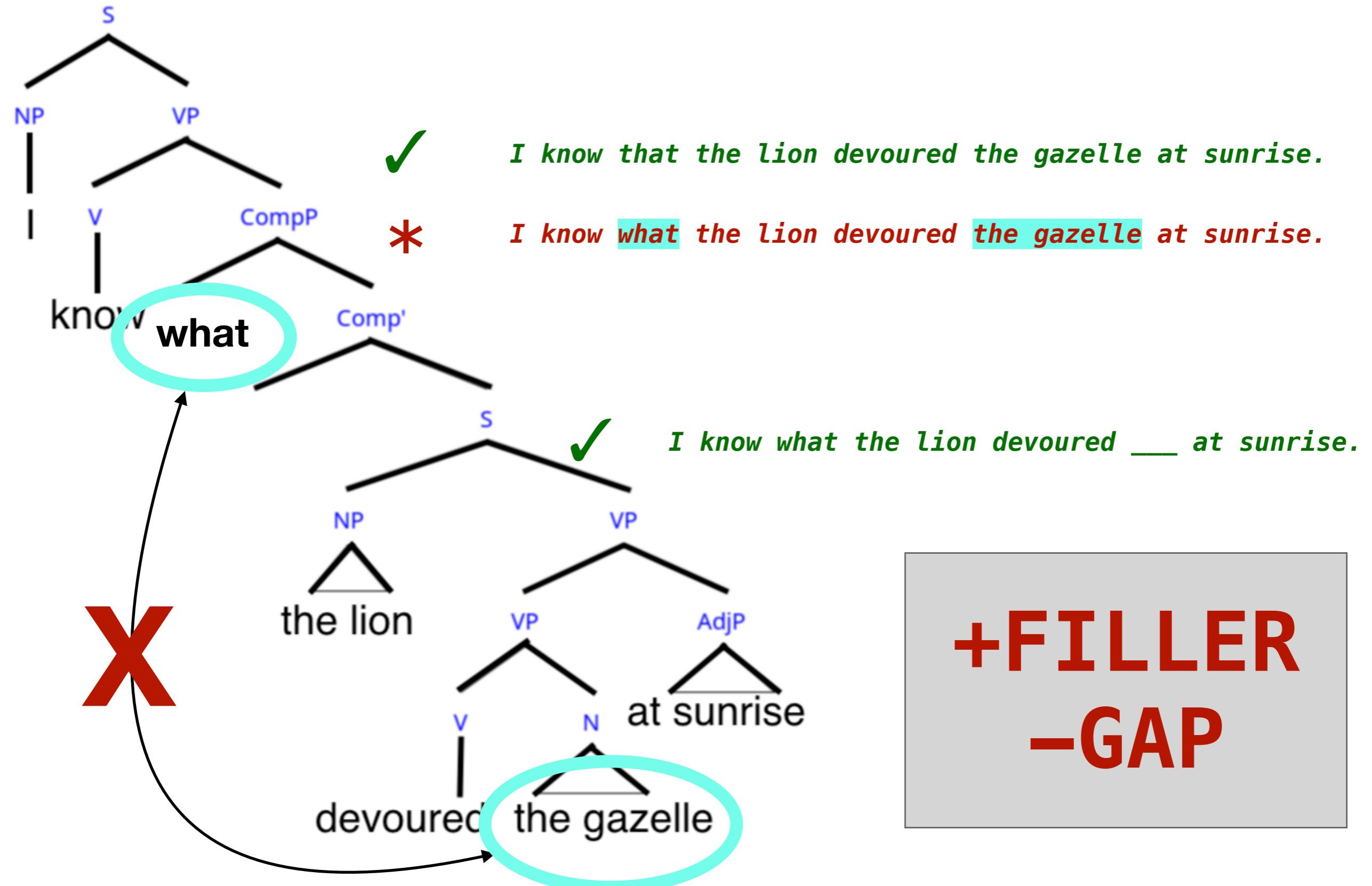


Filler–gap dependencies



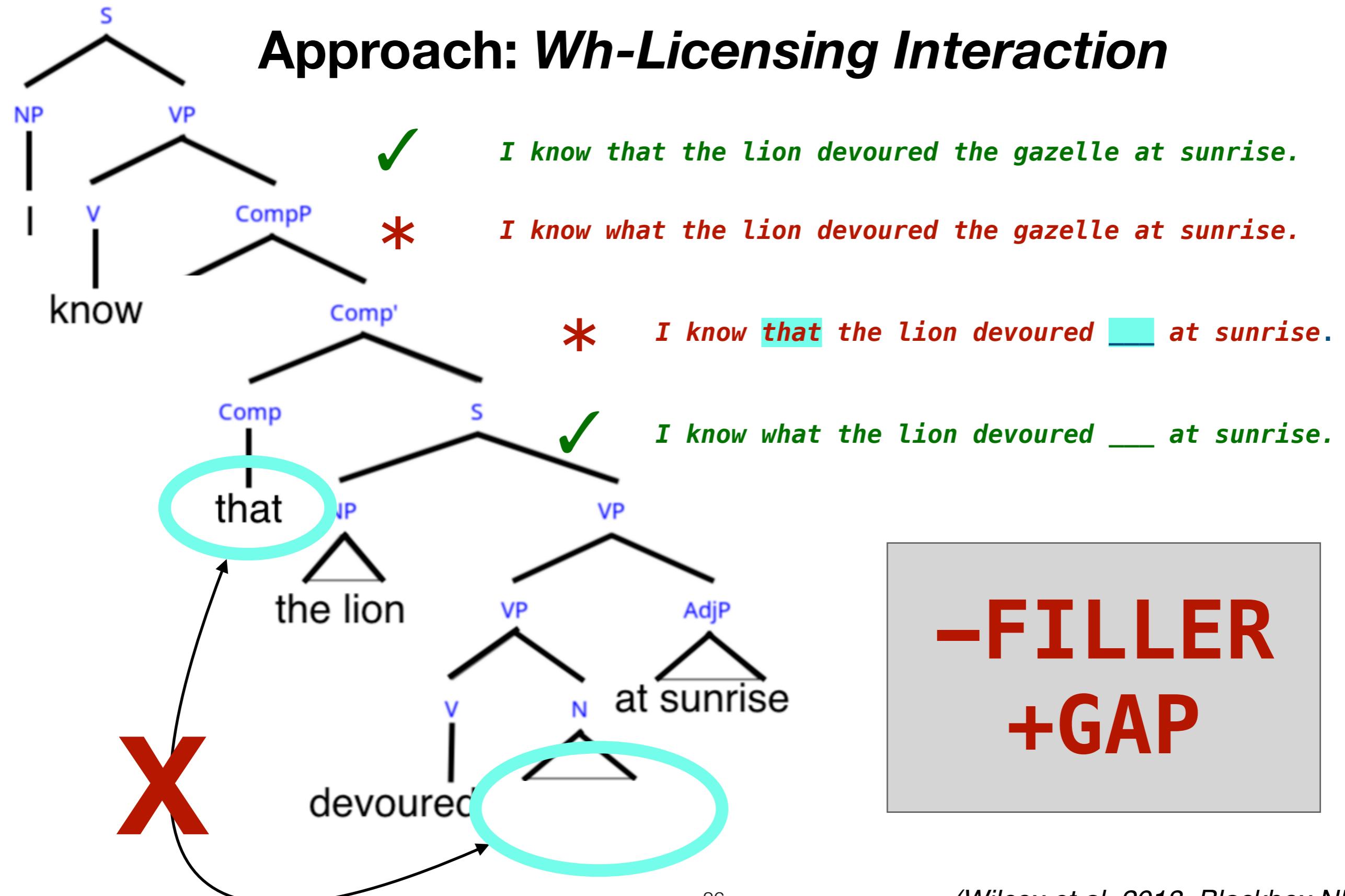
+FILLER
+GAP

Filler–gap dependencies



Filler–gap dependencies

Approach: *Wh-Licensing Interaction*



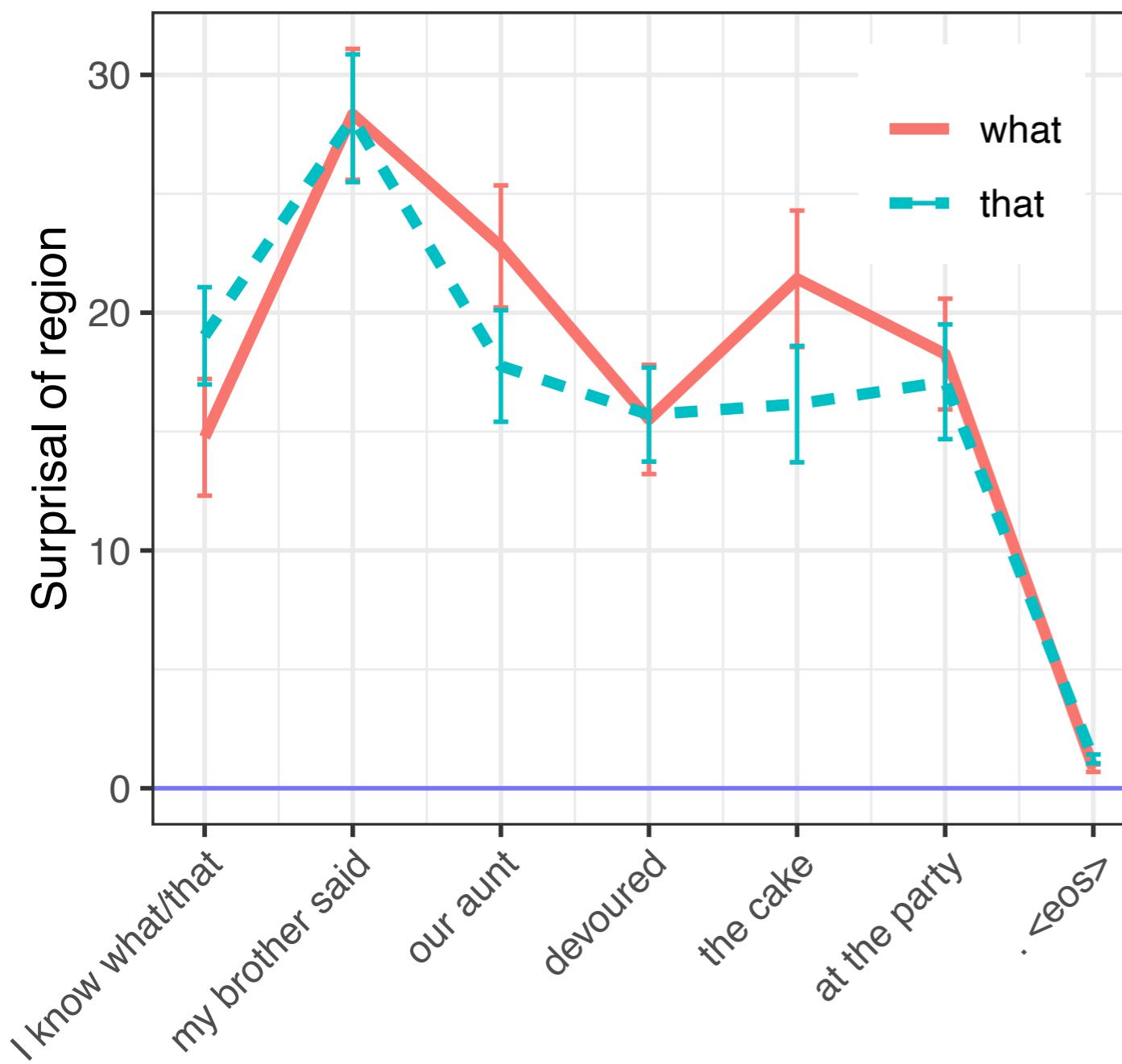
- ✓ *I know **that** my brother said our aunt devoured **the cake** at the party.*
- * *I know **what** my brother said our aunt devoured **the cake** at the party.*

✓

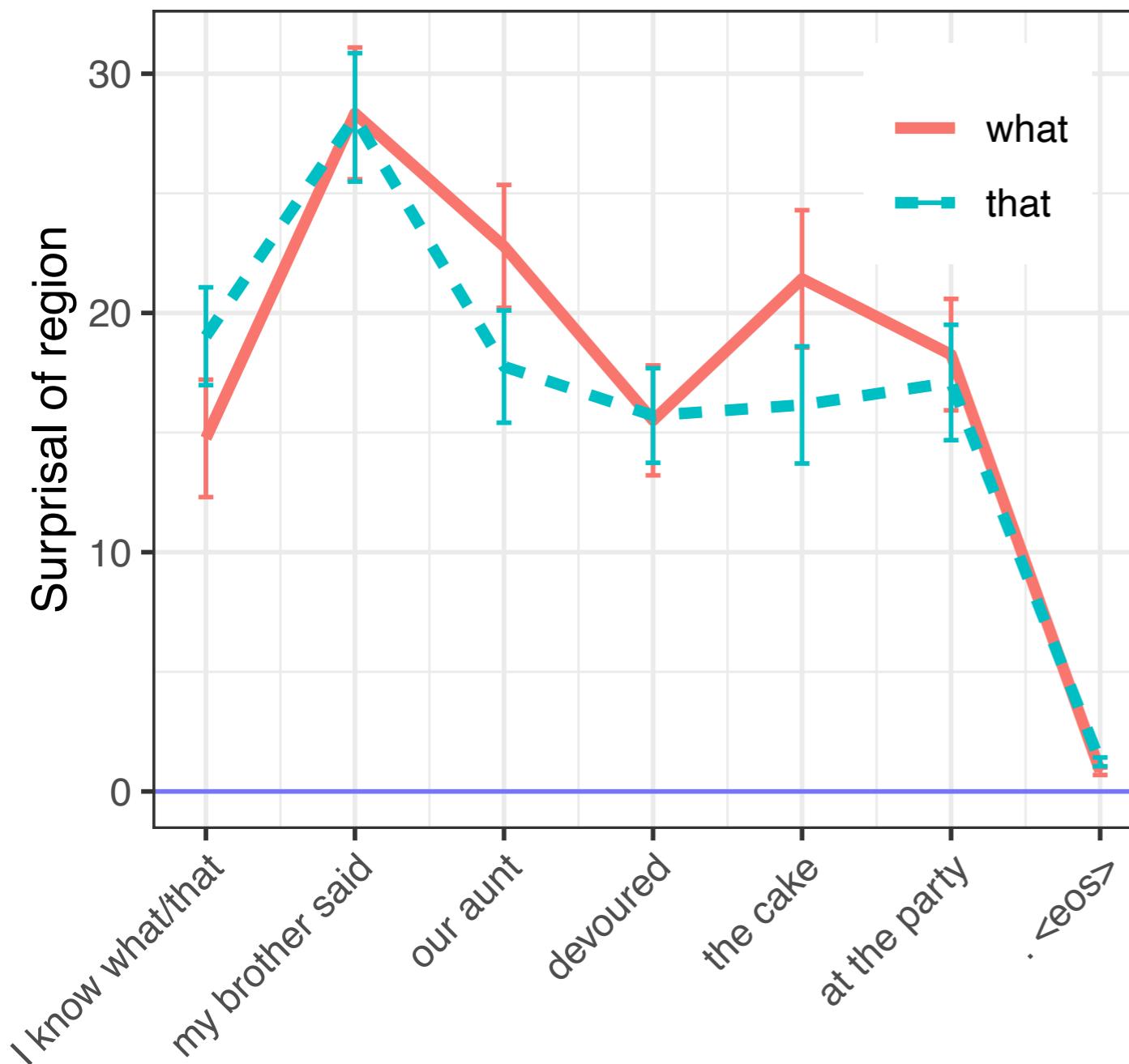
*I know **that** my brother said our aunt devoured **the cake** at the party.*

*

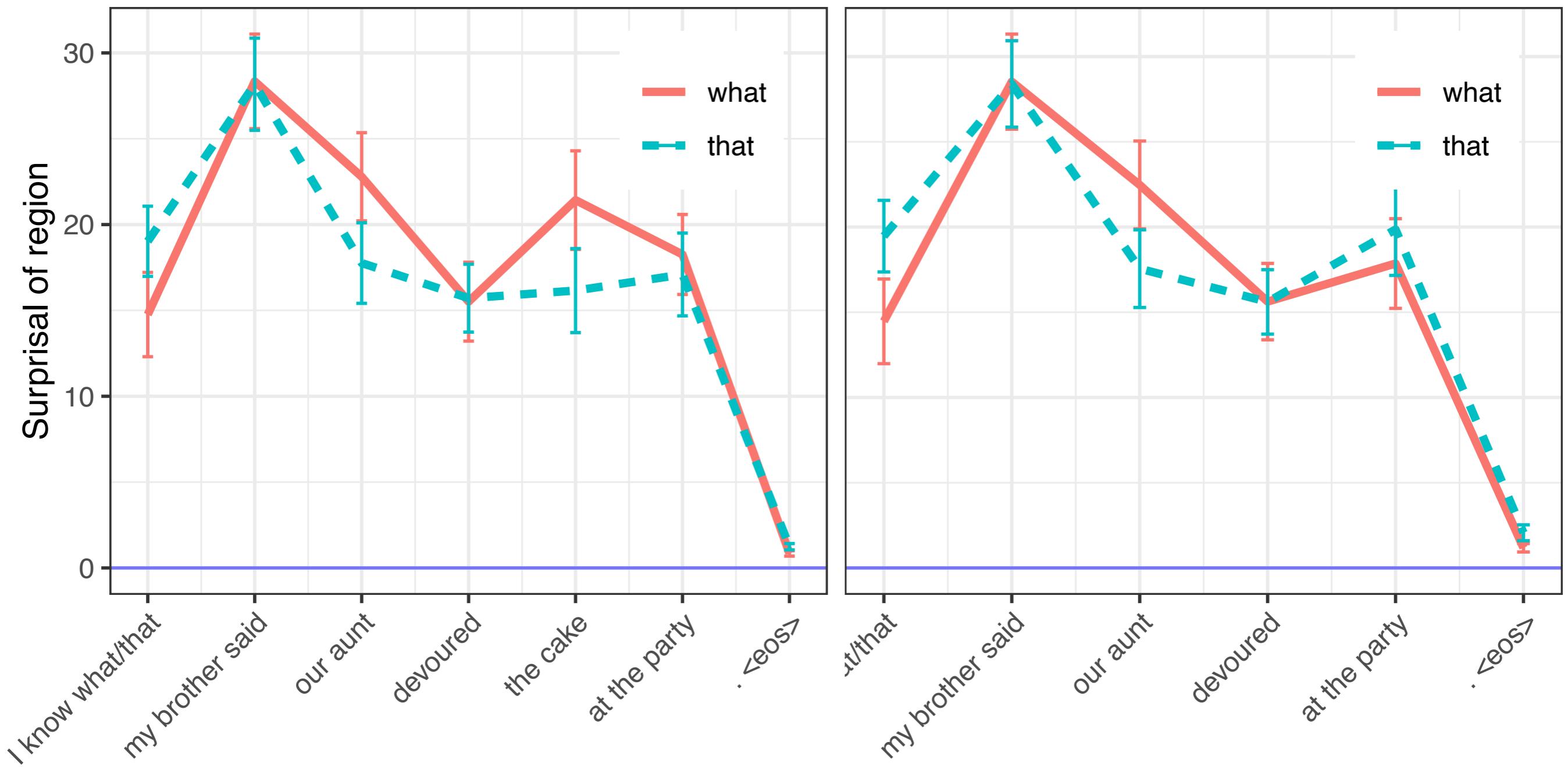
*I know **what** my brother said our aunt devoured **the cake** at the party.*



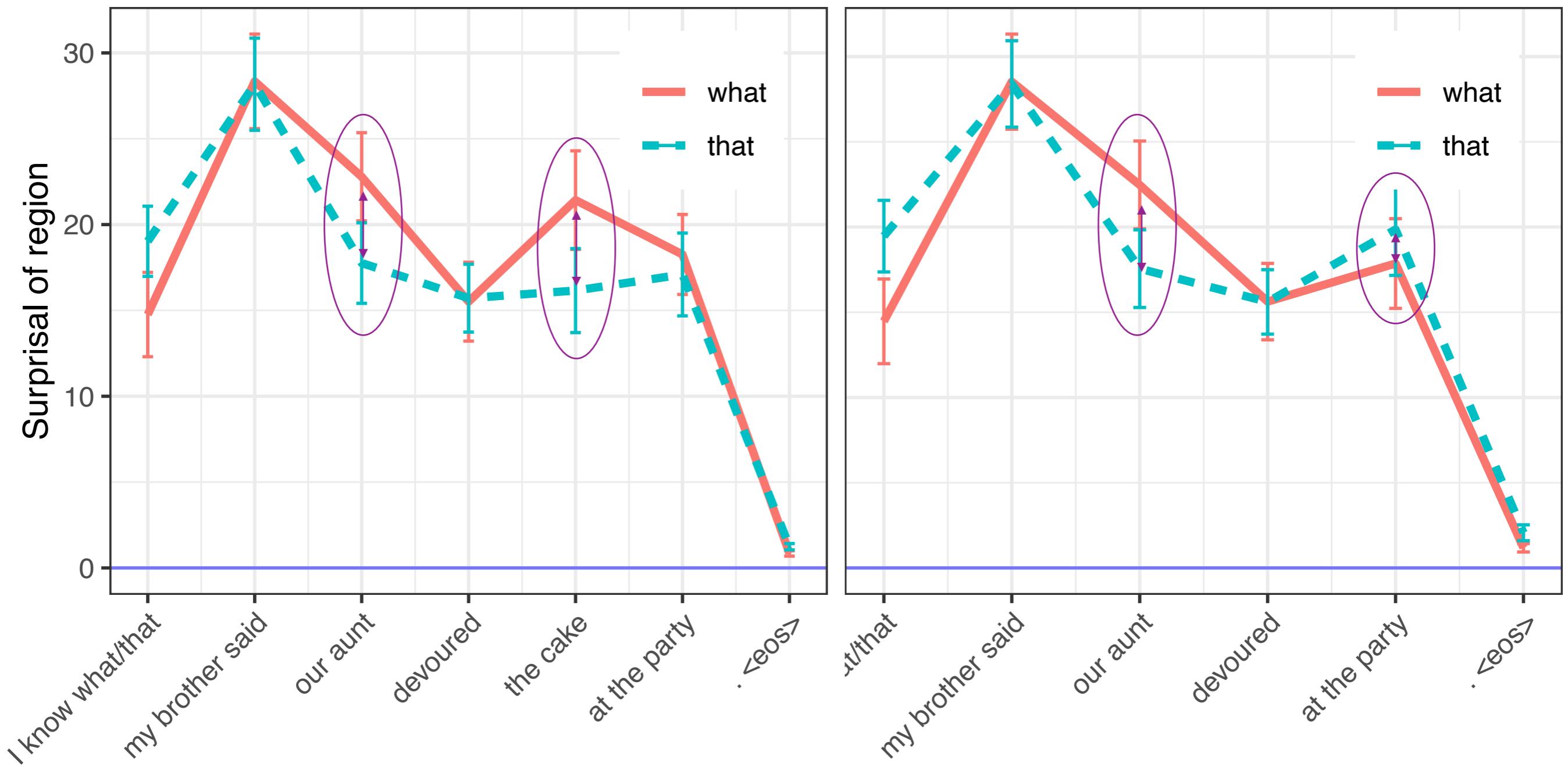
- ✓ *I know **that** my brother said our aunt devoured **the cake** at the party.*
- * *I know **what** my brother said our aunt devoured **the cake** at the party.*
- * *I know **that** my brother said our aunt devoured _____ at the party.*
- ✓ *I know **what** my brother said our aunt devoured _____ at the party.*



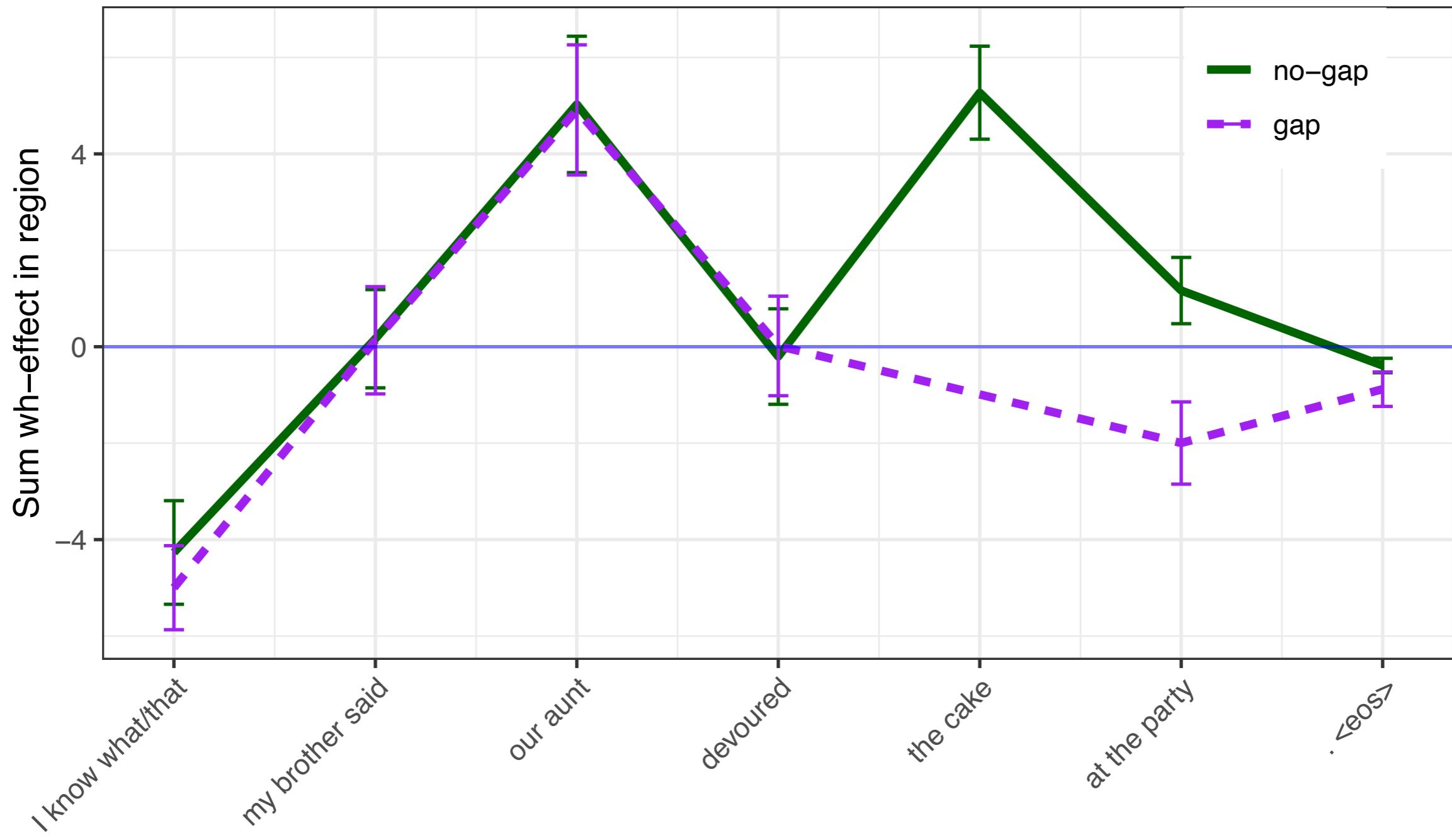
- ✓ *I know **that** my brother said our aunt devoured **the cake** at the party.*
- * *I know **what** my brother said our aunt devoured **the cake** at the party.*
- * *I know **that** my brother said our aunt devoured _____ at the party.*
- ✓ *I know **what** my brother said our aunt devoured _____ at the party.*



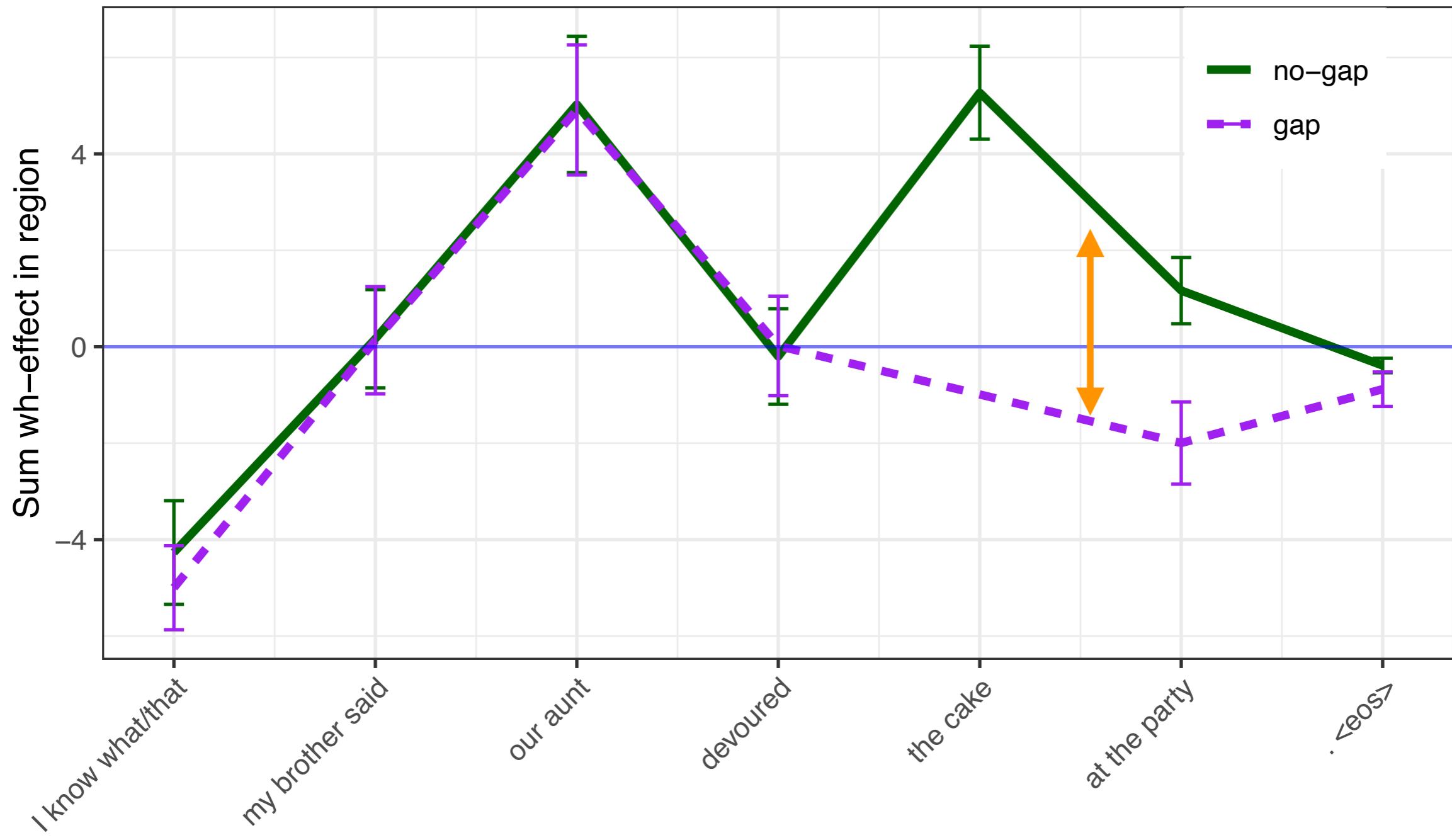
- ✓ *I know **that** my brother said our aunt devoured **the cake** at the party.*
- * *I know **what** my brother said our aunt devoured **the cake** at the party.*
- * *I know **that** my brother said our aunt devoured _____ at the party.*
- ✓ *I know **what** my brother said our aunt devoured _____ at the party.*



- ✓ *I know **that** my brother said our aunt devoured **the cake** at the party.*
- * *I know **what** my brother said our aunt devoured **the cake** at the party.*
- * *I know **that** my brother said our aunt devoured _____ at the party.*
- ✓ *I know **what** my brother said our aunt devoured _____ at the party.*



- ✓ *I know **that** my brother said our aunt devoured **the cake** at the party.*
- * *I know **what** my brother said our aunt devoured **the cake** at the party.*
- * *I know **that** my brother said our aunt devoured _____ at the party.*
- ✓ *I know **what** my brother said our aunt devoured _____ at the party.*



Unboundedness of *wh*-dependencies

I know what our mother gave __ to Mary last weekend.

Unboundedness of wh-dependencies

I know what our mother gave __ to Mary last weekend.

I know what our mother said that your friend gave __ to Mary last weekend.

Unboundedness of *wh*-dependencies

- 0 *I know what our mother gave __ to Mary last weekend.*
- 1 *I know what our mother said that your friend gave __ to Mary last weekend.*

Unboundedness of wh-dependencies

- 0 *I know what our mother gave __ to Mary last weekend.*
- 1 *I know what our mother said that your friend gave __ to Mary last weekend.*
- 2 *I know what our mother said that her friend remarked that your friend gave __ to Mary last weekend.*

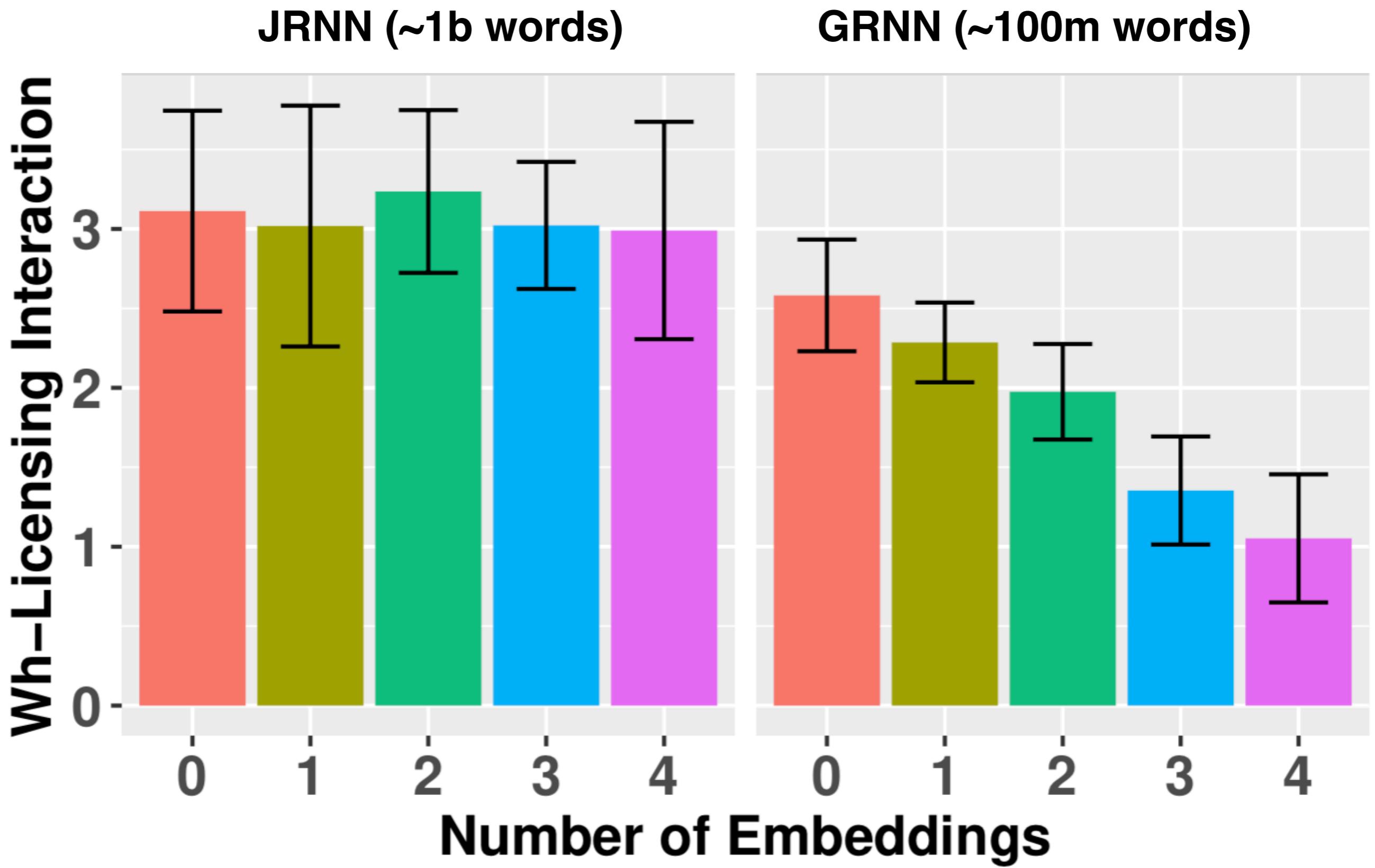
Unboundedness of wh-dependencies

- 0 *I know what our mother gave __ to Mary last weekend.*
- 1 *I know what our mother said that your friend gave __ to Mary last weekend.*
- 2 *I know what our mother said that her friend remarked that your friend gave __ to Mary last weekend.*
- 3 *I know what our mother said that her friend remarked that the park attendant wondered that your friend gave __ to Mary last weekend.*

Unboundedness of wh-dependencies

- 0 *I know what our mother gave to Mary last weekend.*
- 1 *I know what our mother said that your friend gave to Mary last weekend.*
- 2 *I know what our mother said that her friend remarked that your friend gave to Mary last weekend.*
- 3 *I know what our mother said that her friend remarked that the park attendant wondered that your friend gave to Mary last weekend.*
- 4 *I know what our mother said that her friend remarked that the park attendant wondered that the people stated that your friend gave to Mary last weekend.*

Unboundedness: Object Gap

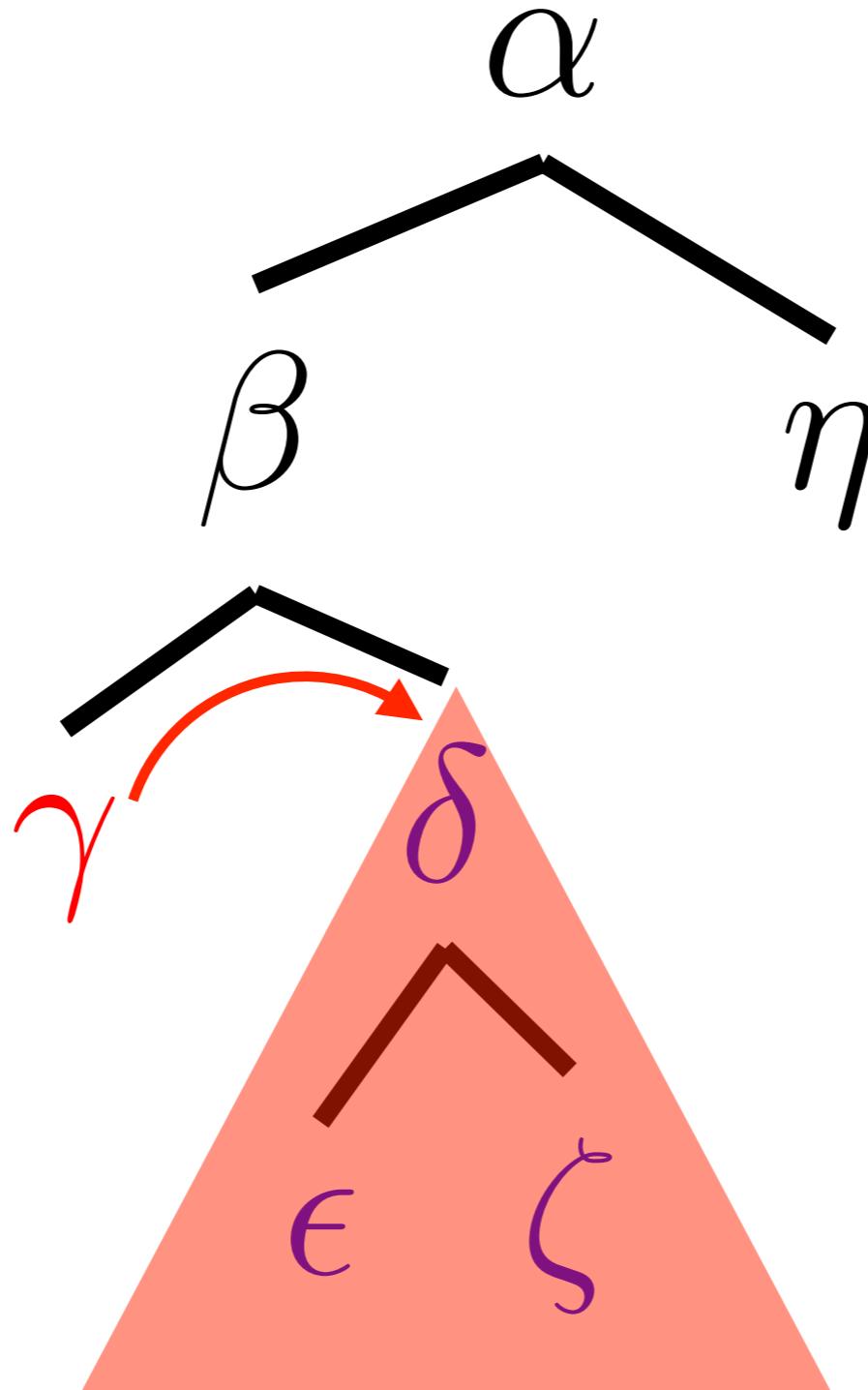


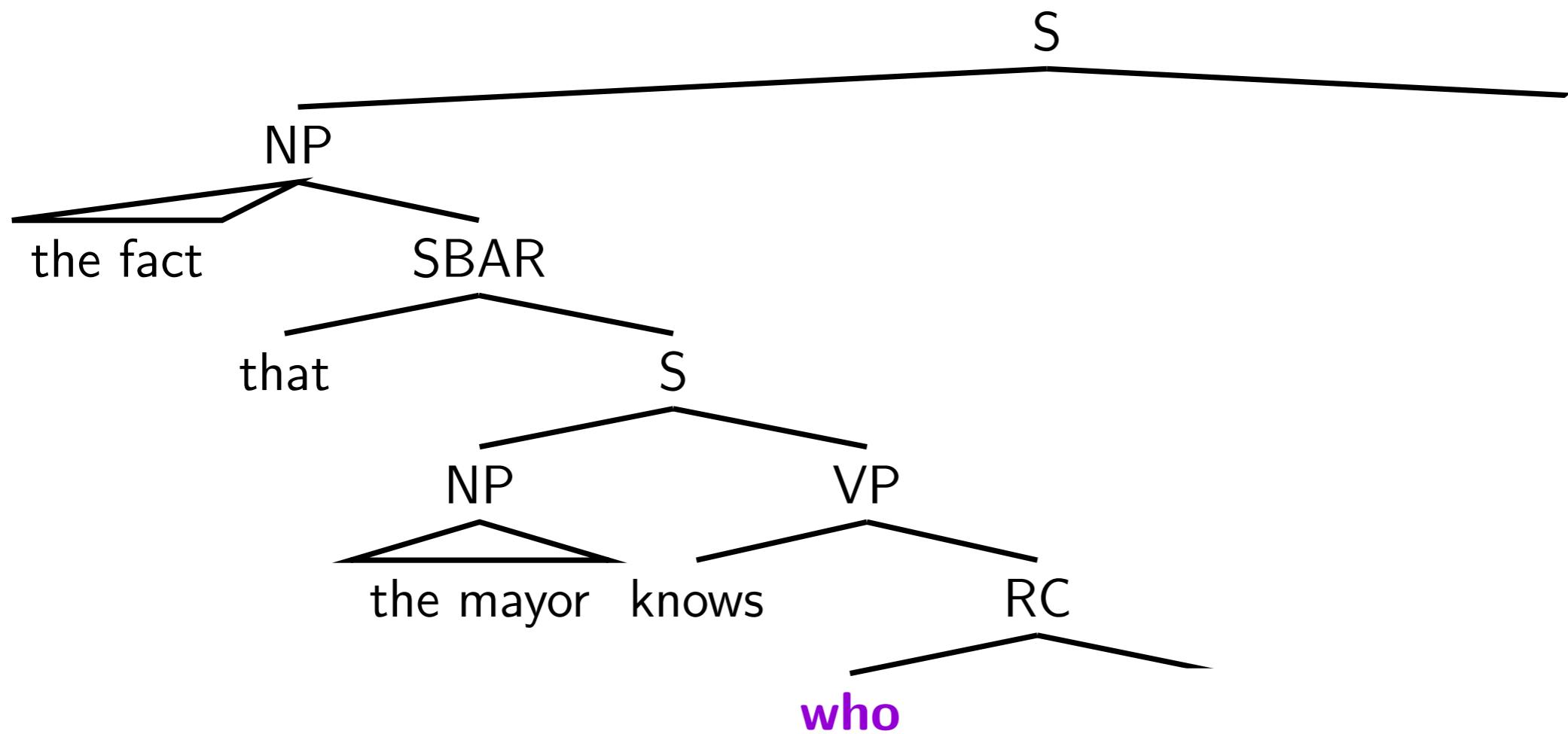
Potential concern #1

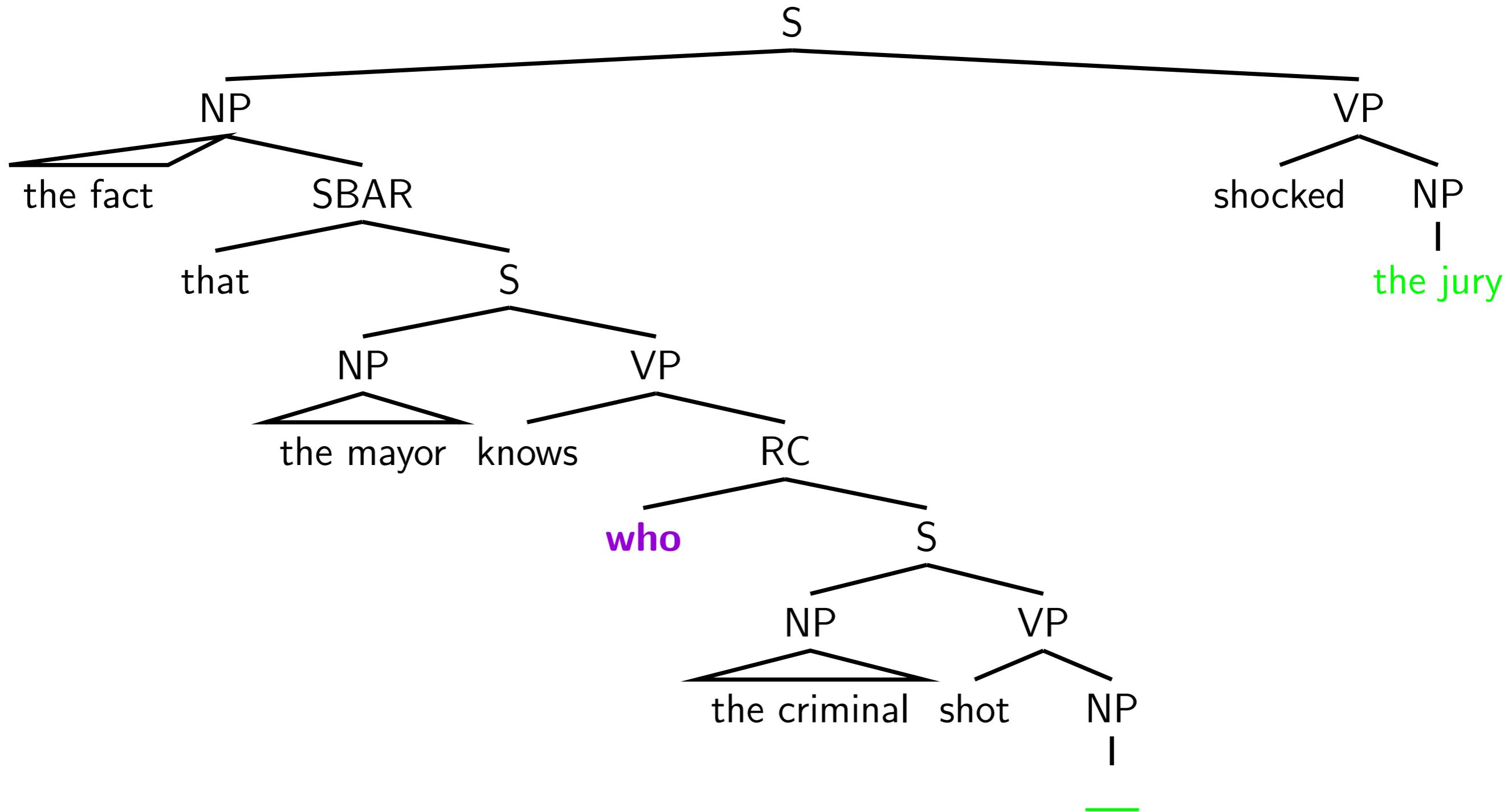
Couldn't the models be learning a *linear* dependency between
filler and gap, not a *hierarchical* dependency?

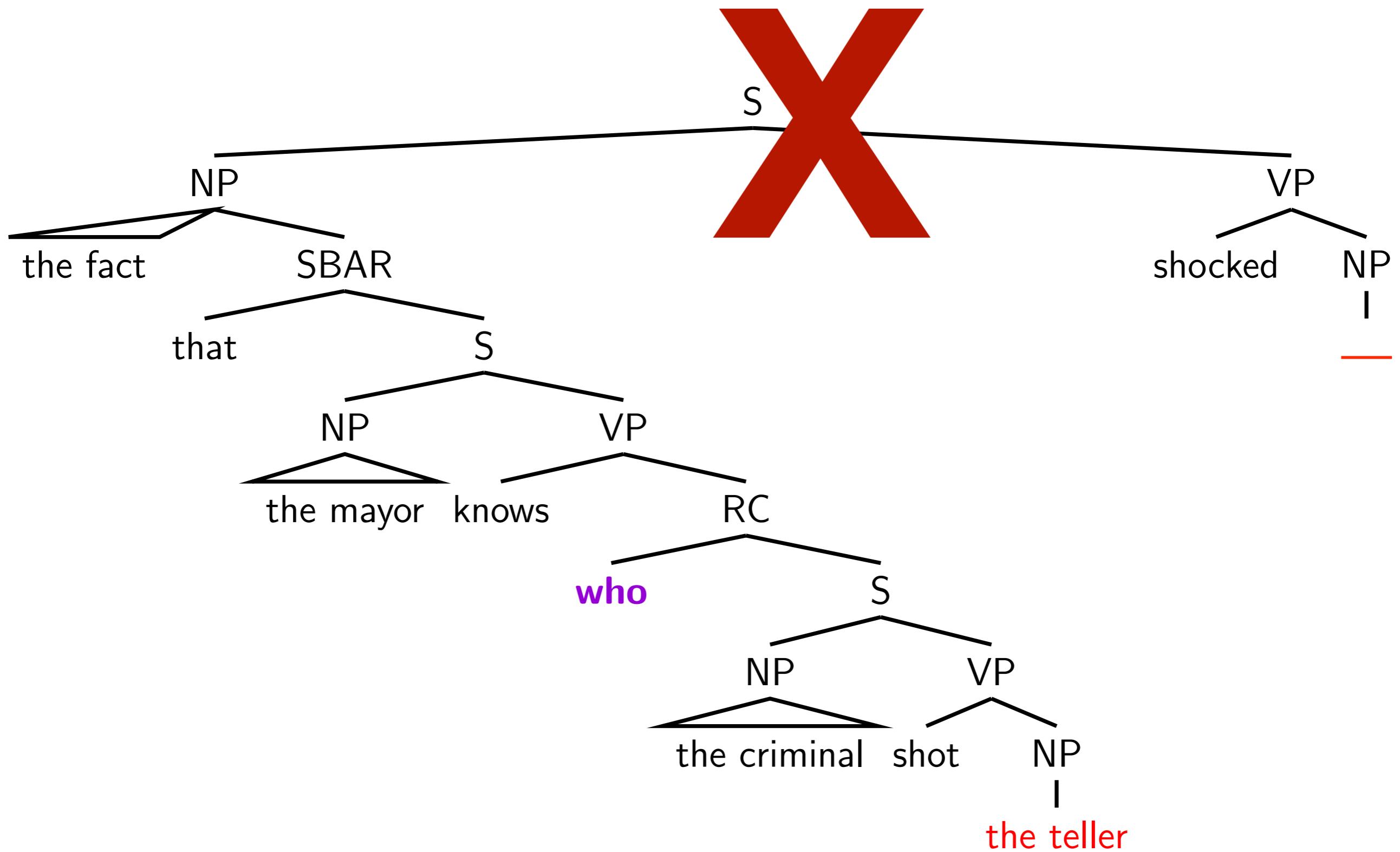
Syntactic Hierarchy

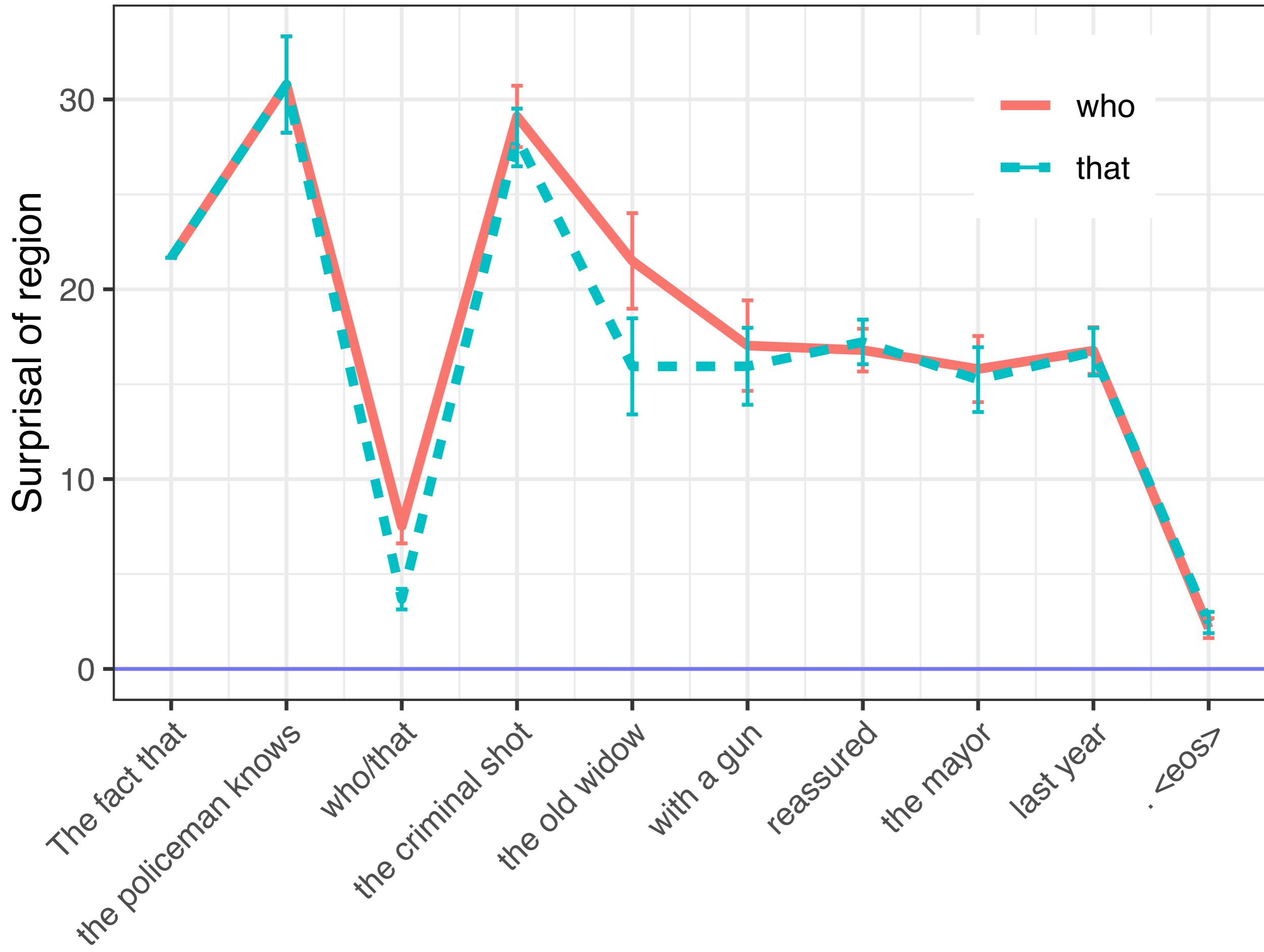
- A filler must be appropriately “above” its gap

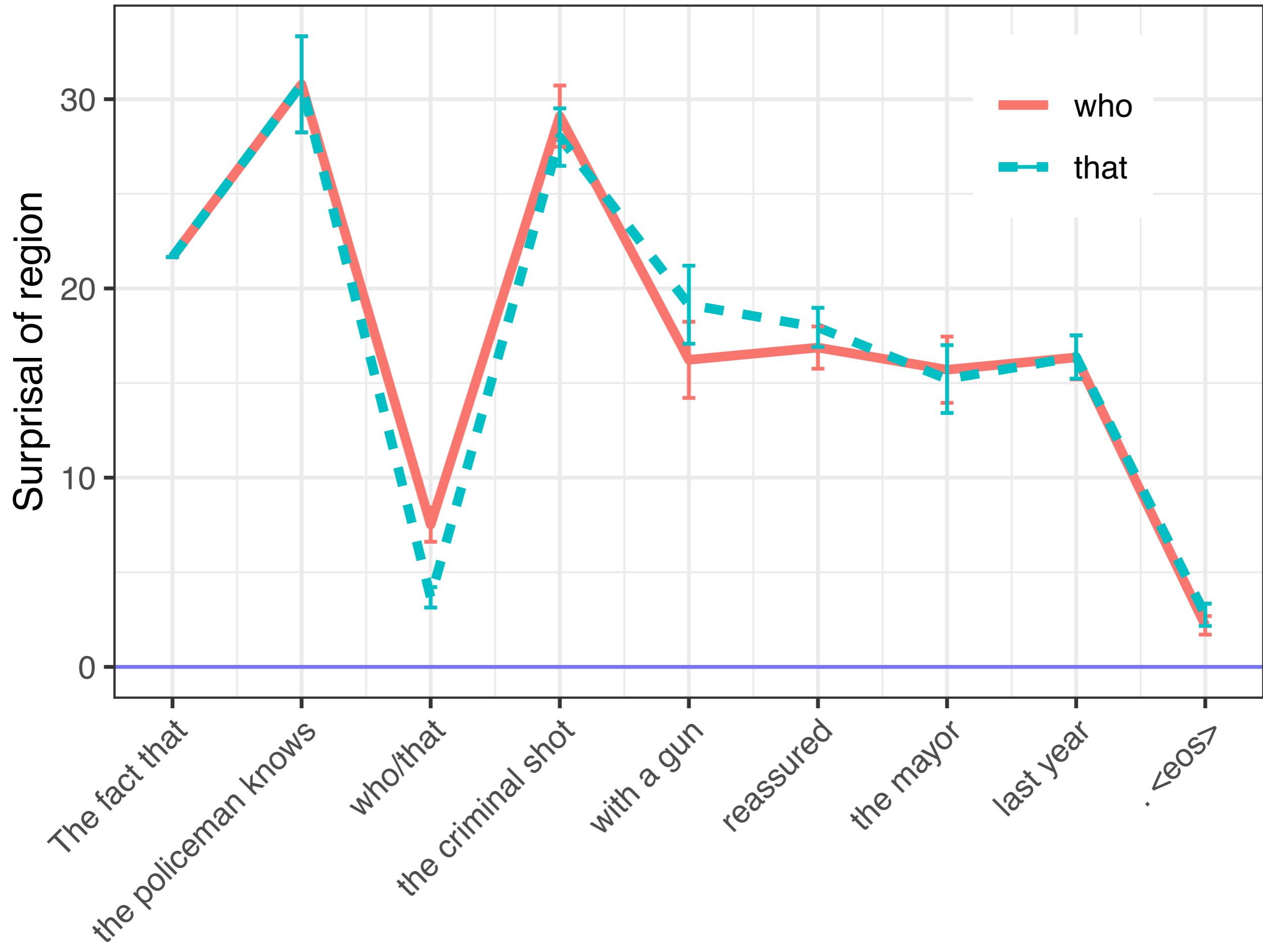


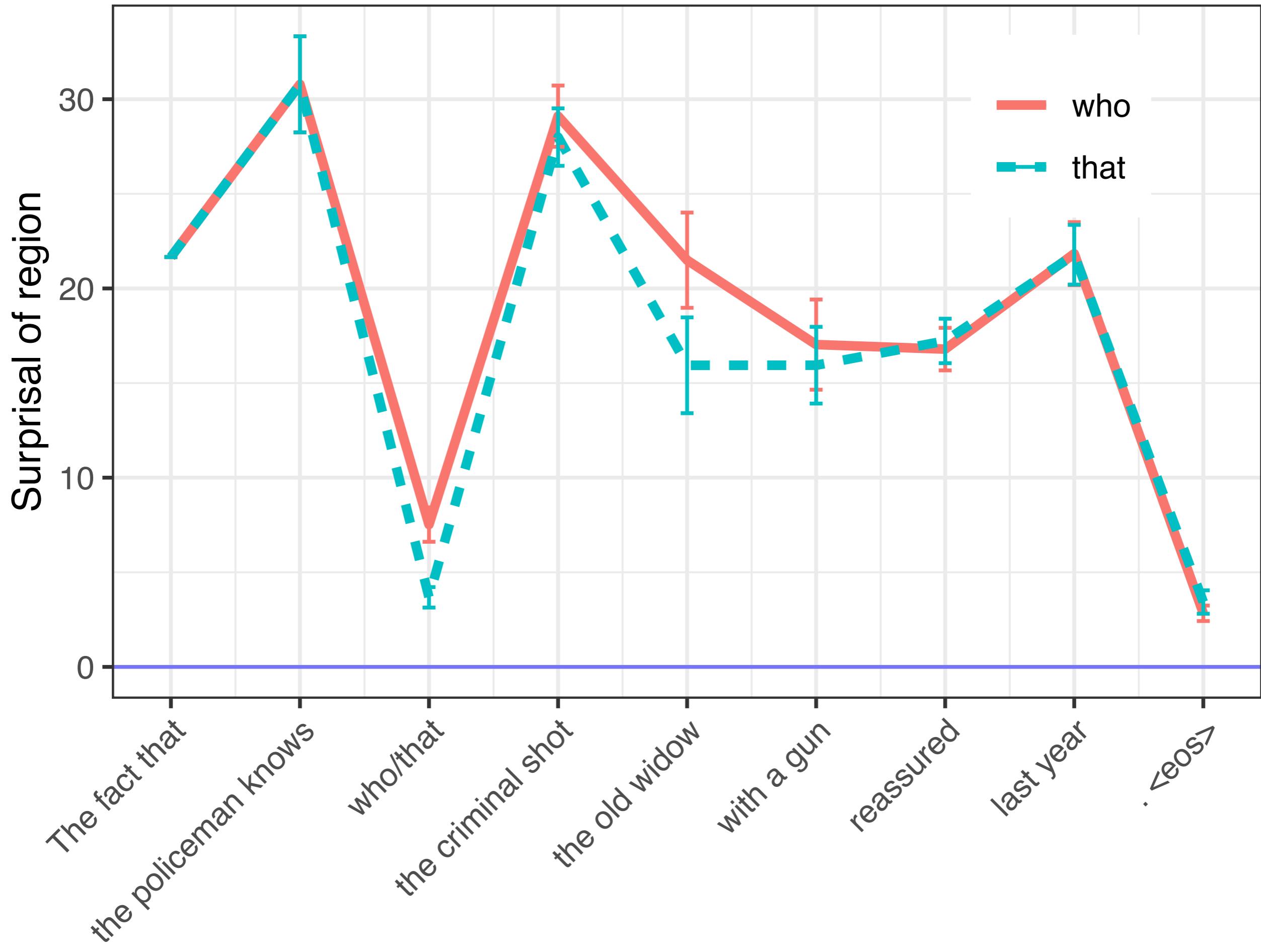


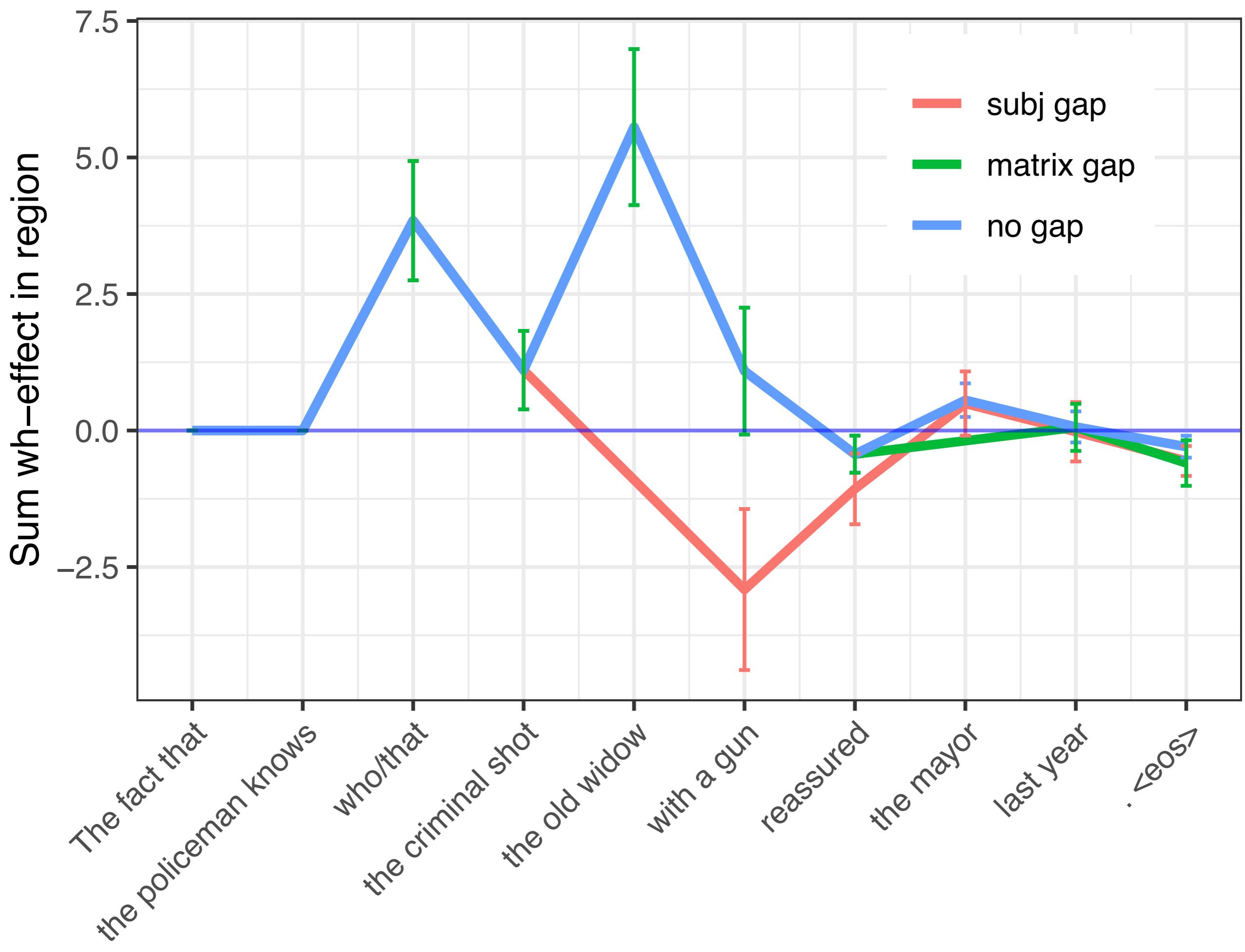












Potential concern #1

Couldn't the models be learning a *linear* dependency between filler and gap, not a *hierarchical* dependency?

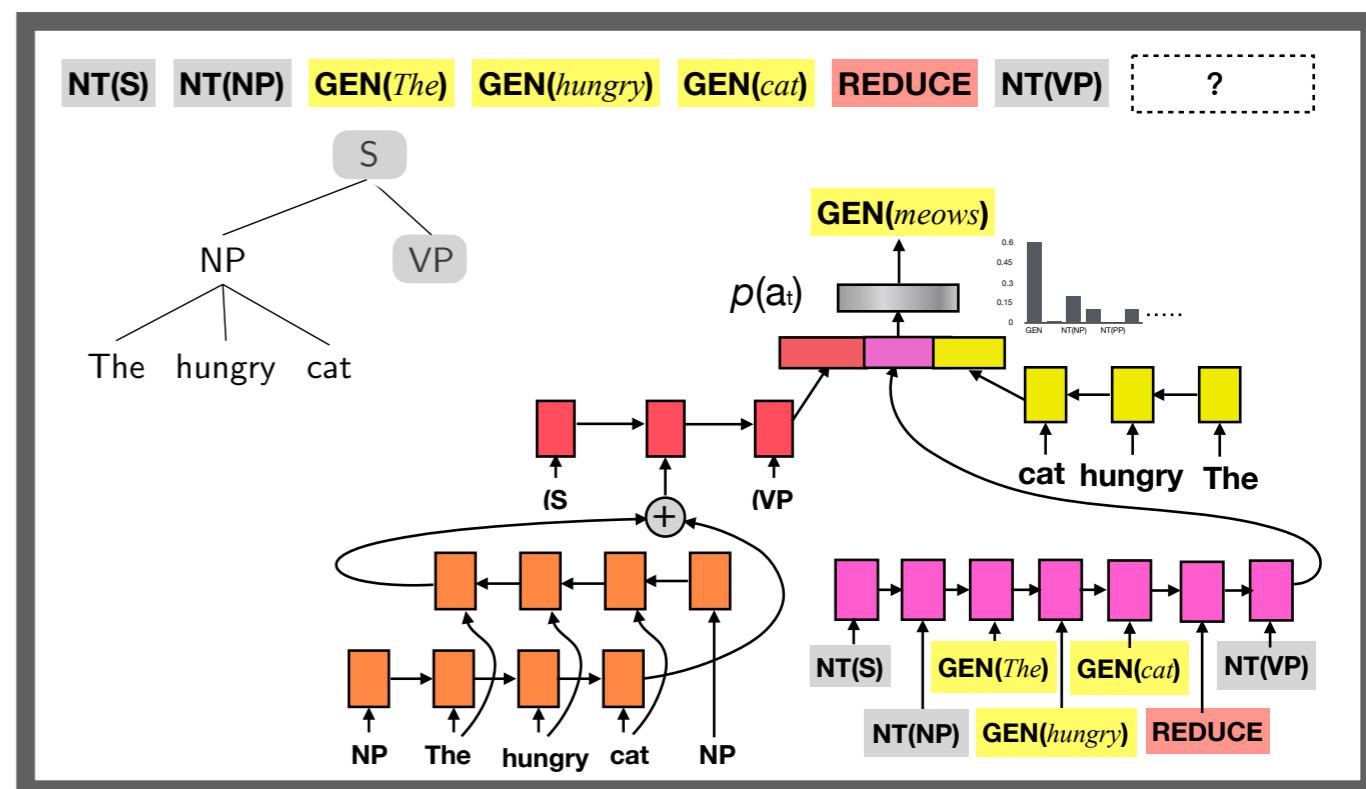
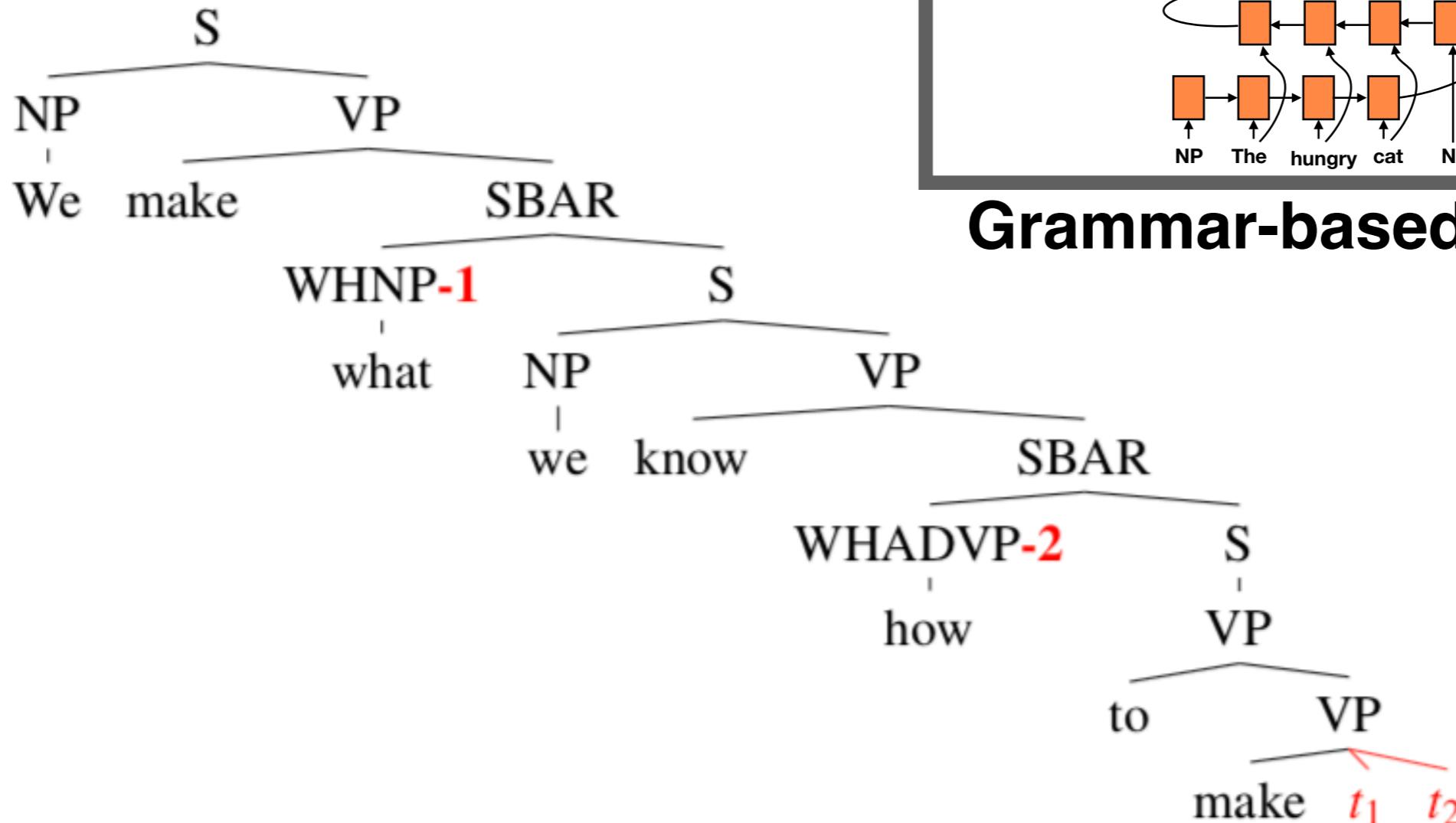
Potential concern #1 — *addressed*

Couldn't the models be learning a *linear* dependency between
filler and gap, not a *hierarchical* dependency?

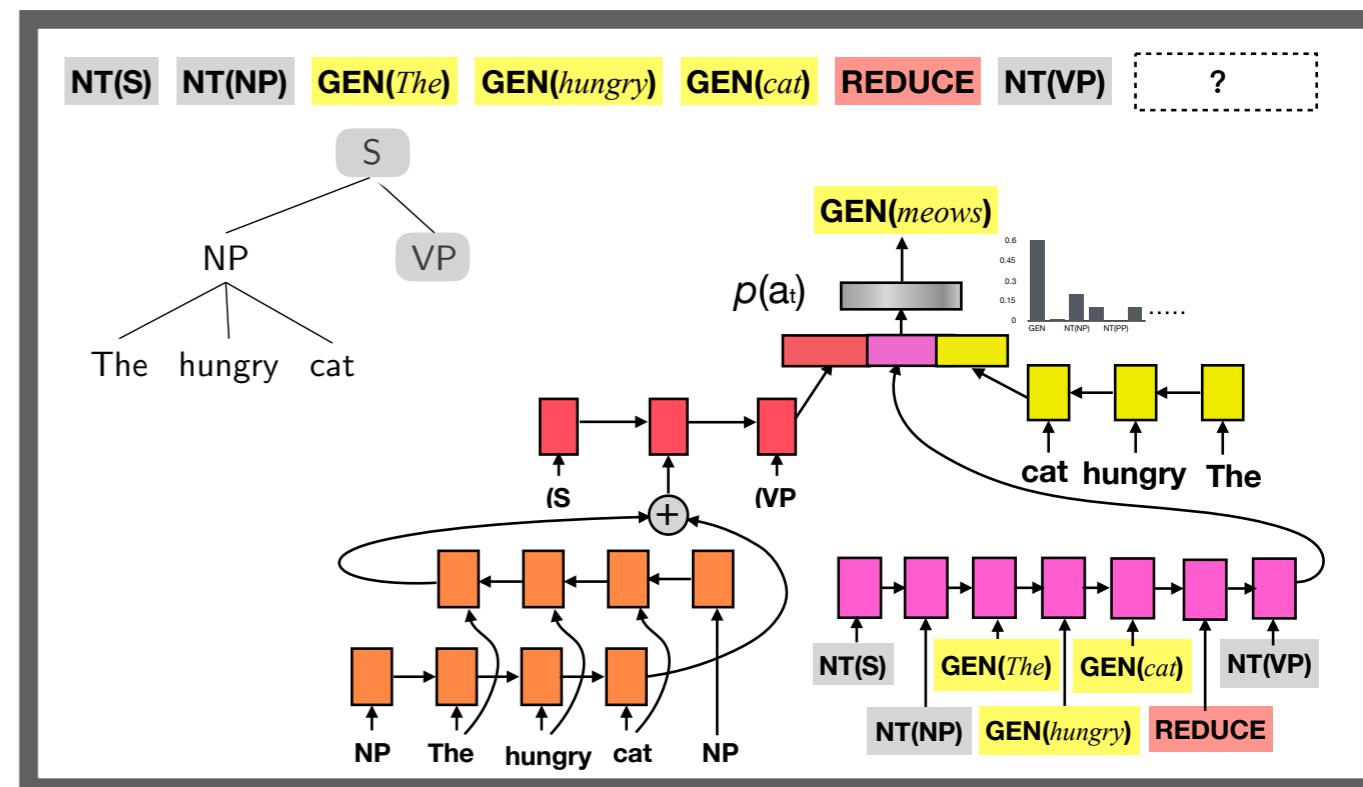
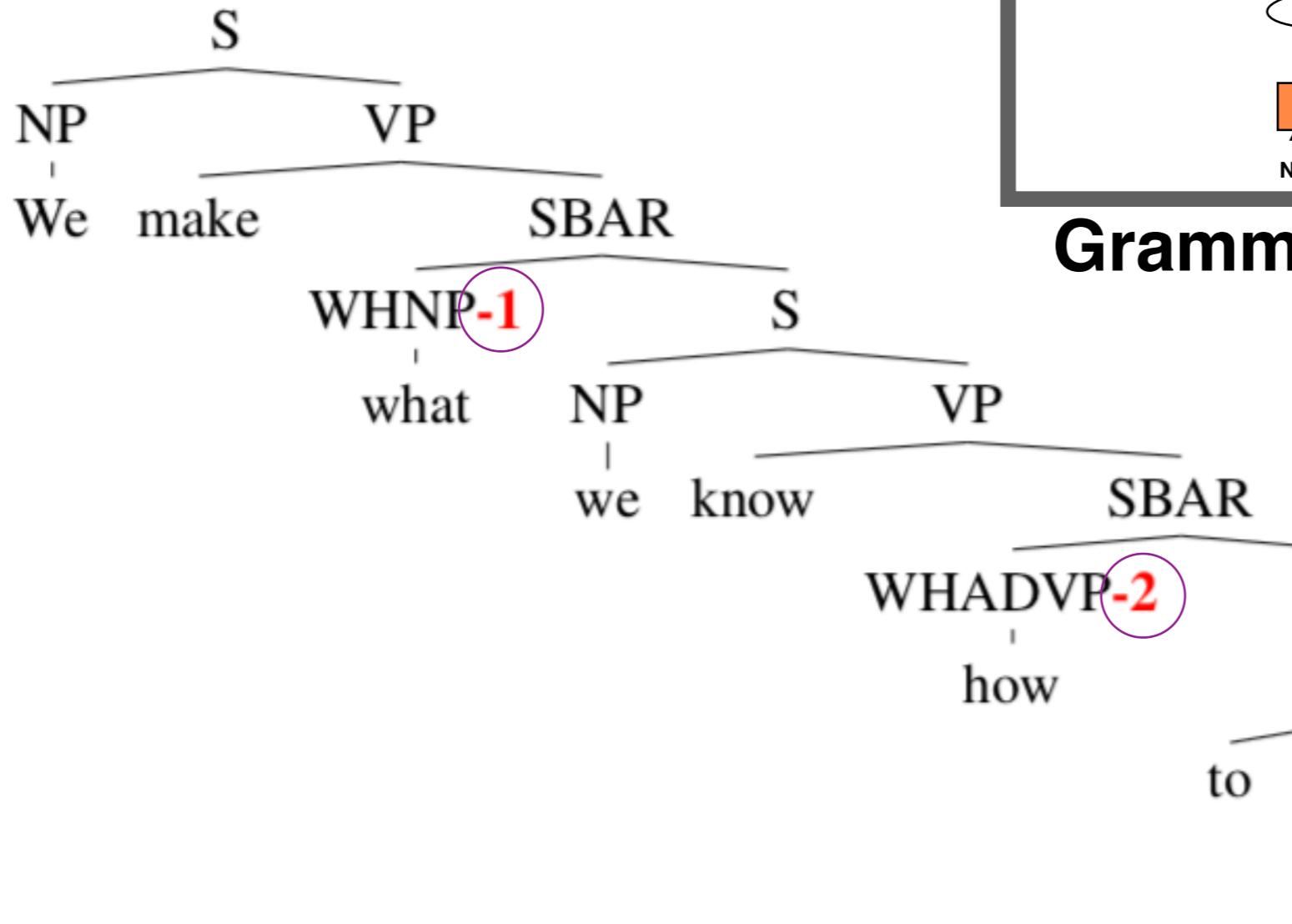


Our results suggest that RNN models trained on enough data
are sensitive to syntactic hierarchy for *wh*-dependency

Does syntactic supervision help?

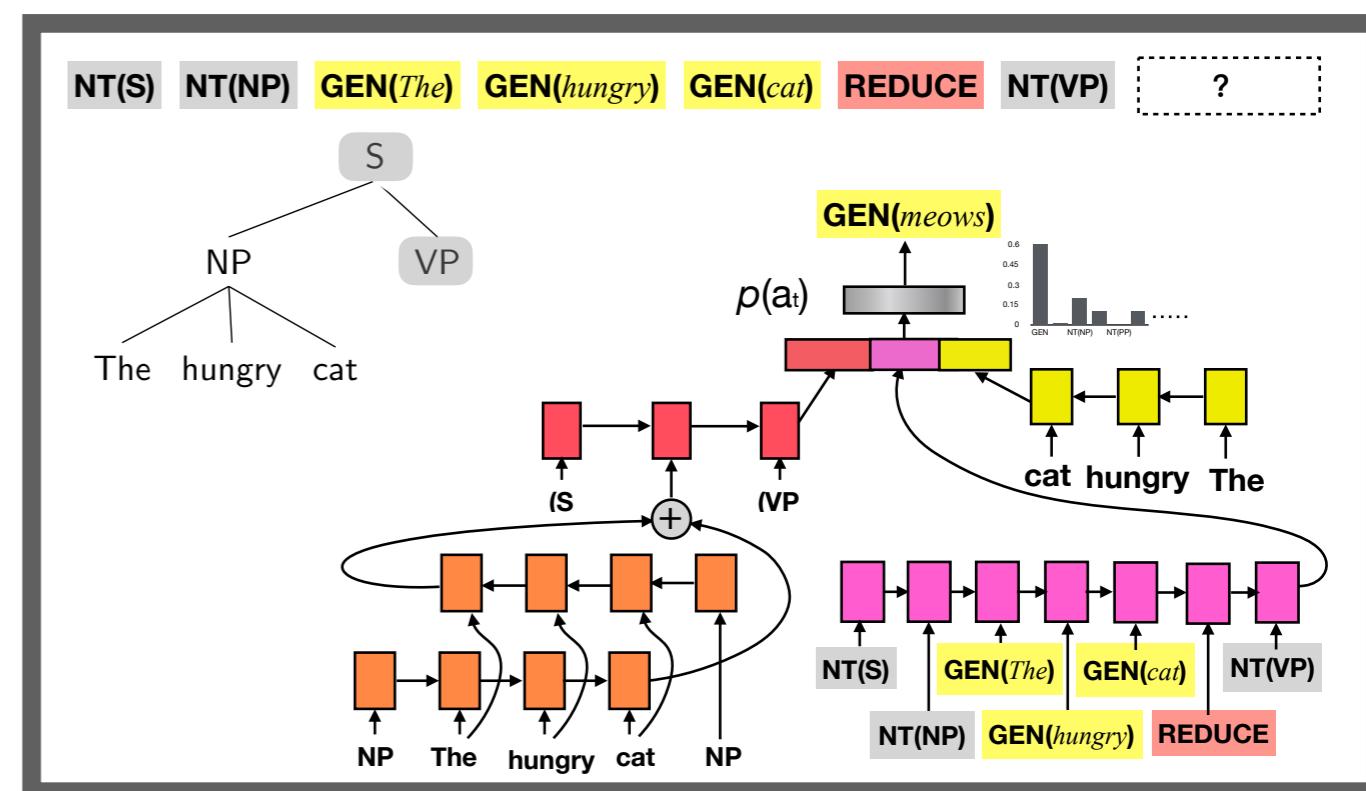
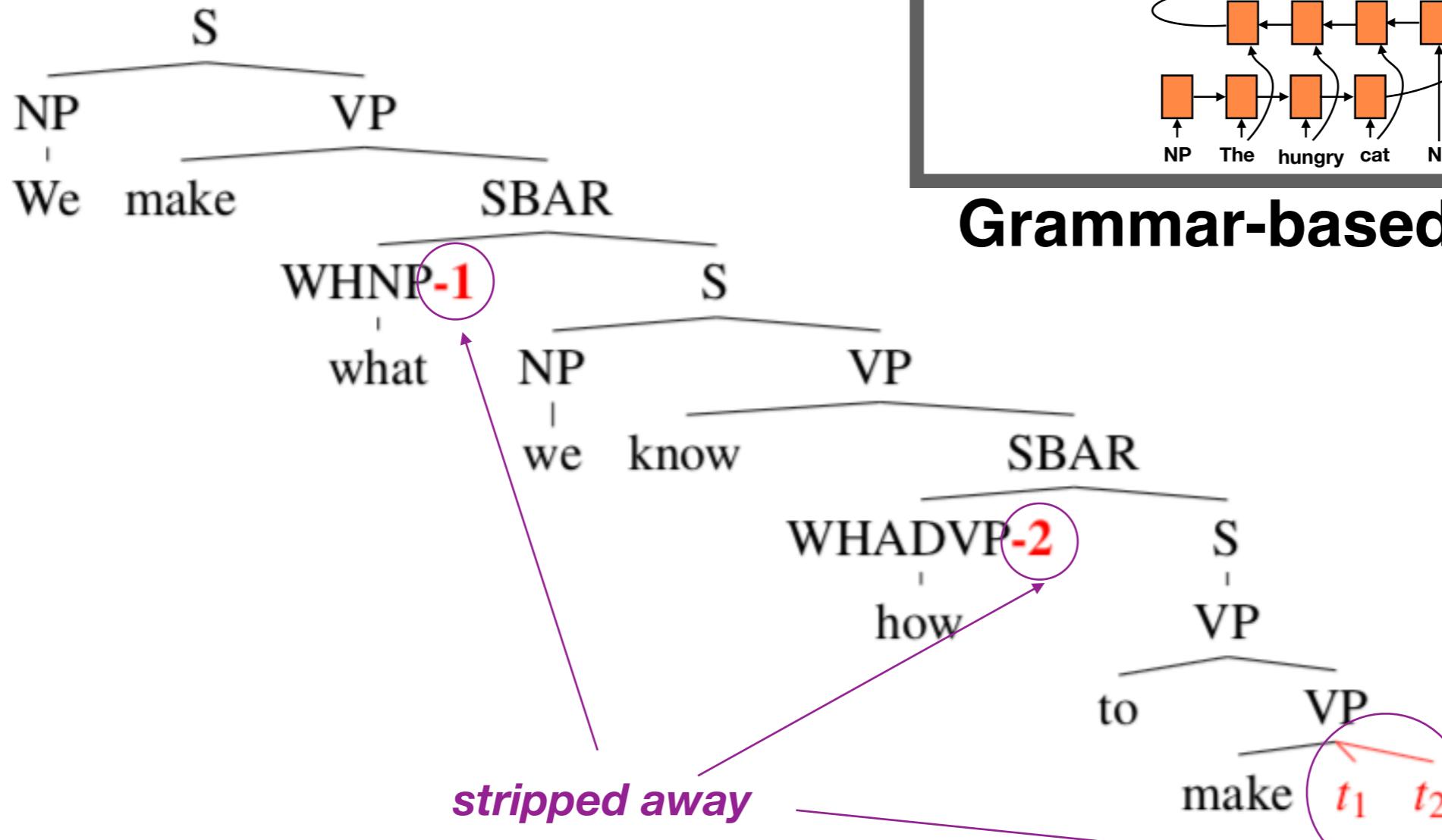


Does syntactic supervision help?



Grammar-based model (RNNG)

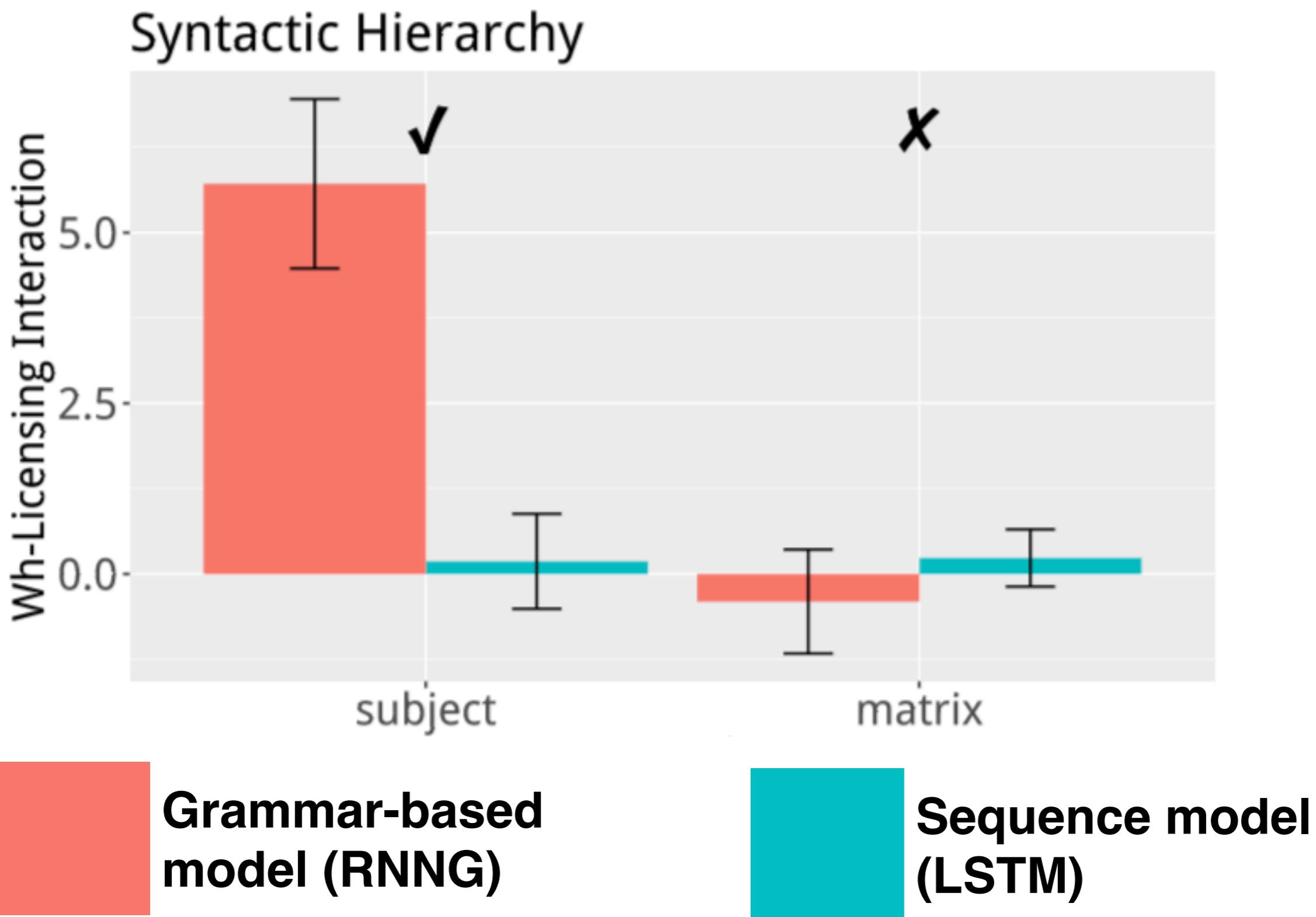
Does syntactic supervision help?



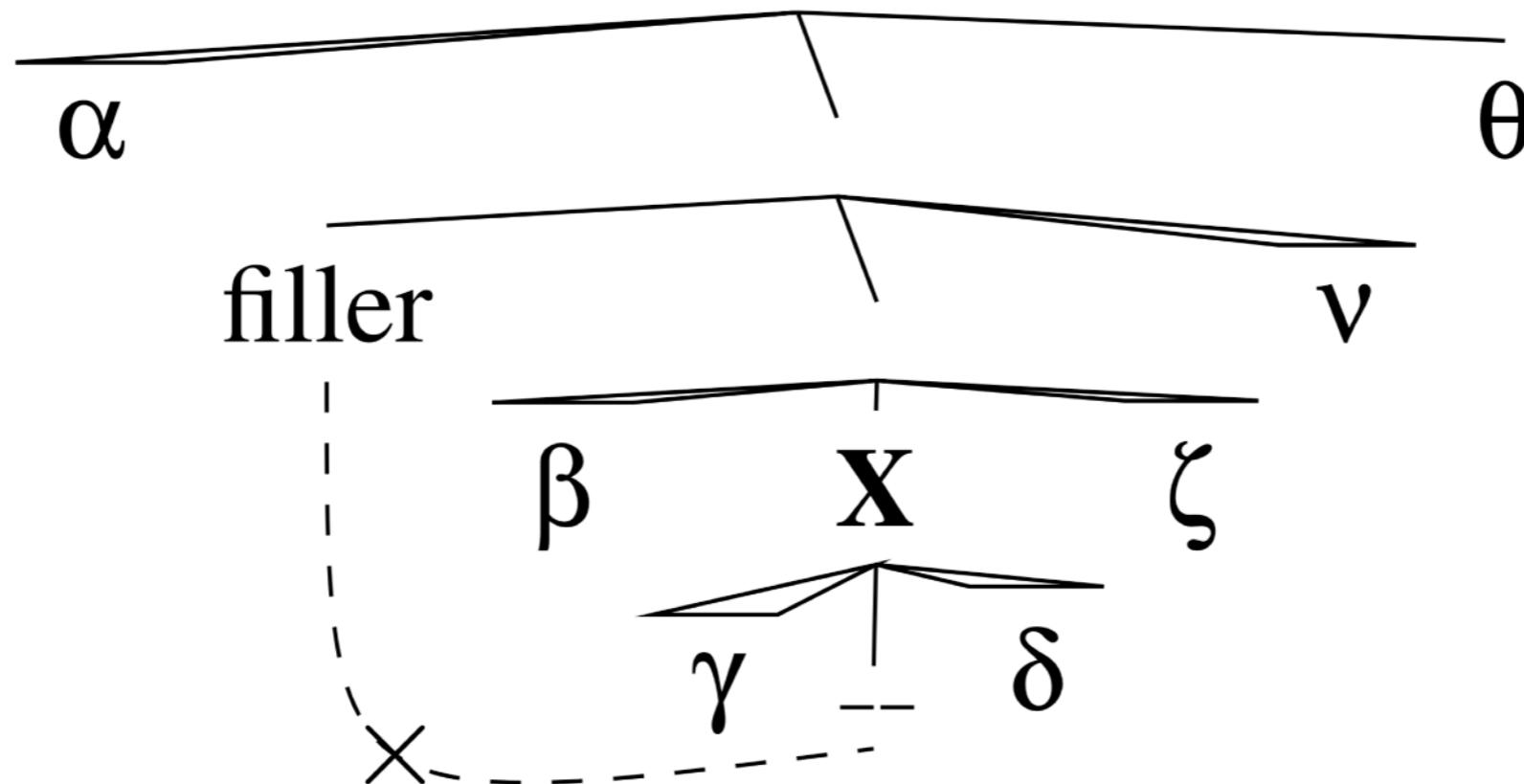
Grammar-based model (RNNG)

Syntactic supervision helps a lot!

- With small-dataset training (1m words):

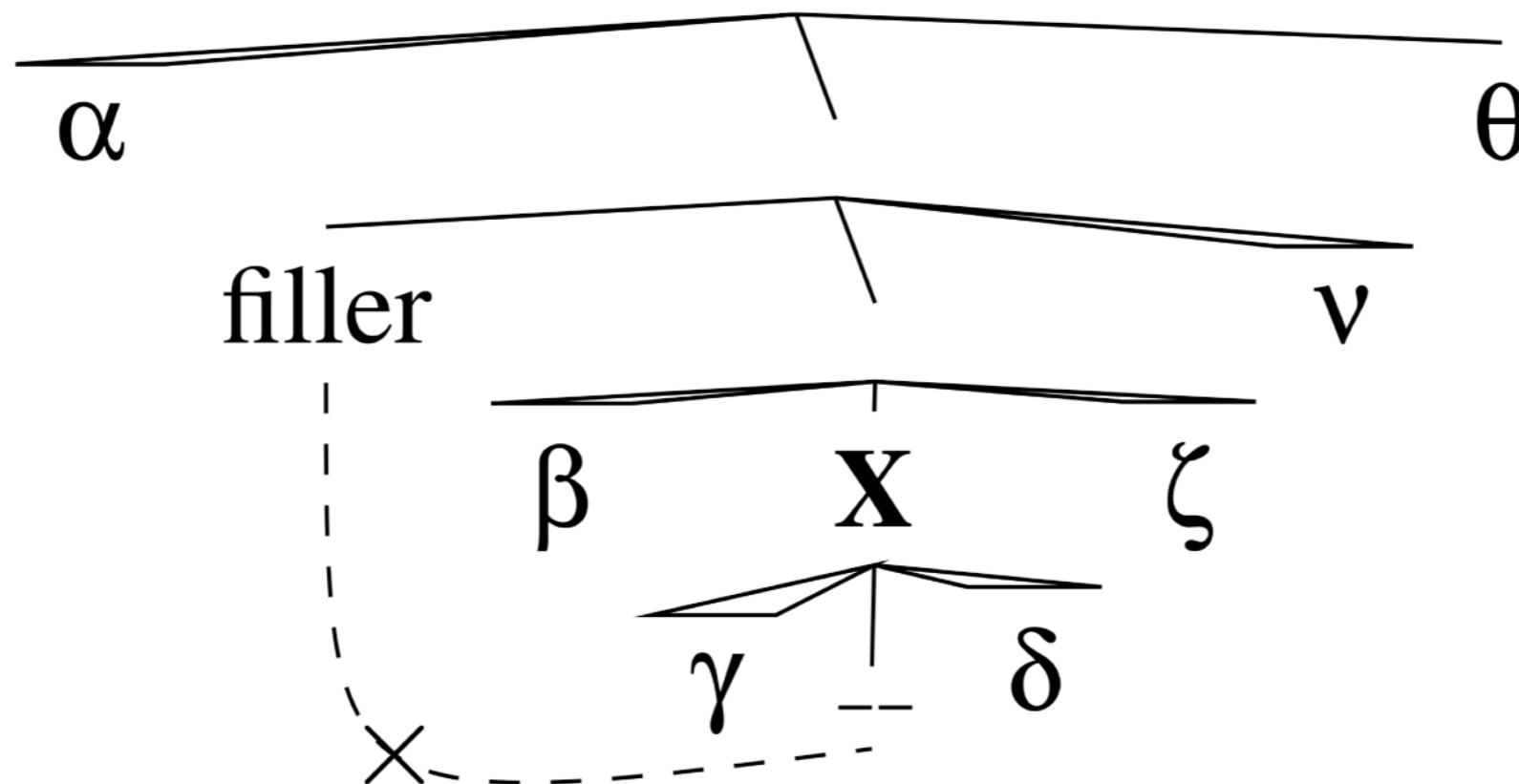


Syntactic island constraints



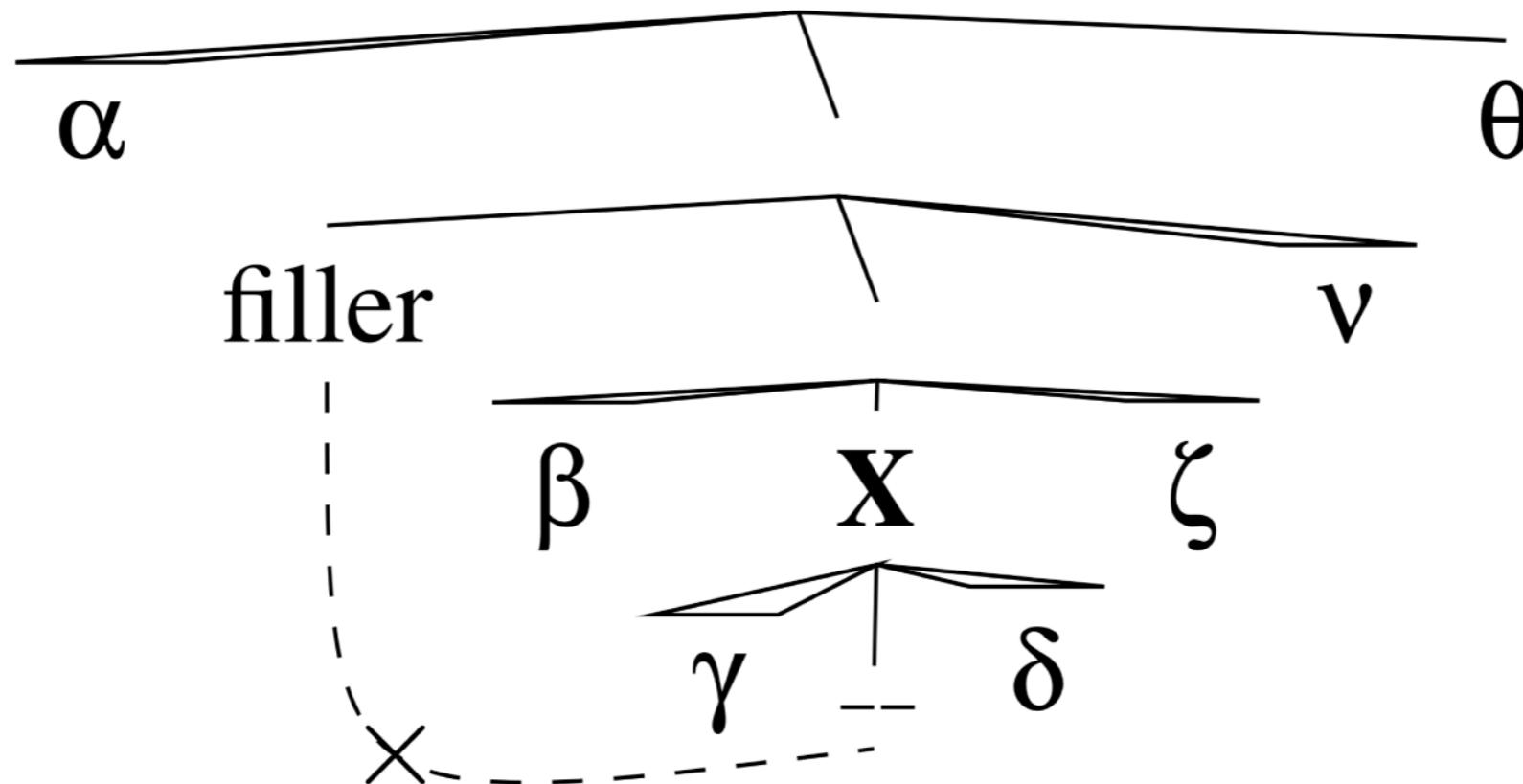
Syntactic island constraints

- Some types of phrases are *islands*: filler–gap dependencies cannot link from outside to inside of them



Syntactic island constraints

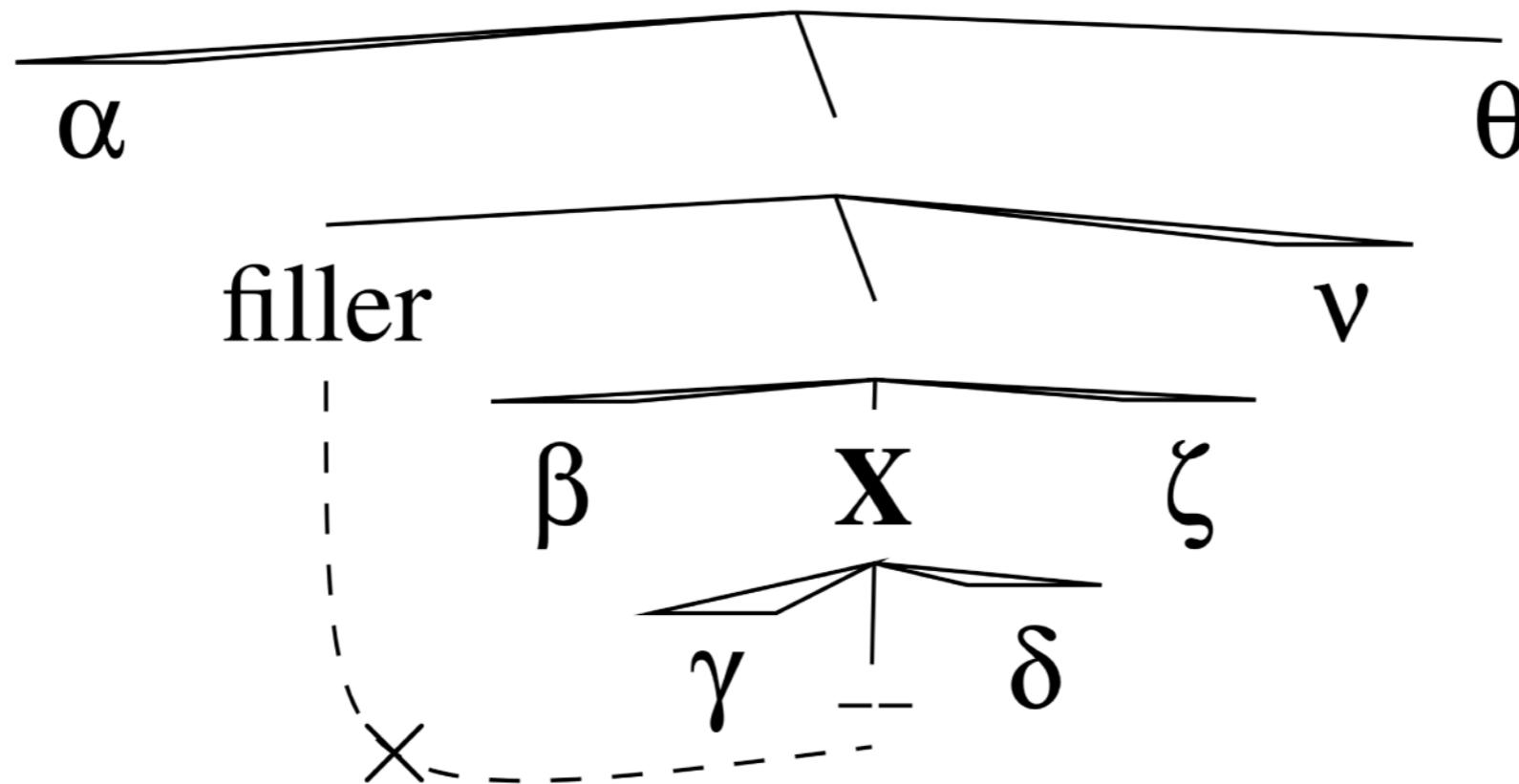
- Some types of phrases are *islands*: filler–gap dependencies cannot link from outside to inside of them



- Islands are prominent in learnability debates: they'd require learning from negative evidence, and are rare structures

Syntactic island constraints

- Some types of phrases are *islands*: filler–gap dependencies cannot link from outside to inside of them



- Islands are prominent in learnability debates: they'd require learning from negative evidence, and are rare structures
- We take a language model to have learned an island constraint if it *fails* to propagate filler-generated expectations for gaps into phrases that should be islands

Syntactic islands

Wh-complementizers block filler—gap dependencies:

I know what Alex said...

...your friend devoured █ at the party.
[null complementizer]

Syntactic islands

Wh-complementizers block filler—gap dependencies:

I know what Alex said...

✓ *...your friend devoured █ at the party.*
 [null complementizer]

Syntactic islands

Wh-complementizers block filler—gap dependencies:

I know what Alex said...

✓ *...your friend devoured █ at the party.*
 [null complementizer]

...that your friend devoured █ at the party.
 [that complementizer]

Syntactic islands

Wh-complementizers block filler—gap dependencies:

I know what Alex said...

- ✓ ...*your friend devoured* █ *at the party.*
[null complementizer]

- ✓ ...***that*** *your friend devoured* █ *at the party.*
[*that* complementizer]

Syntactic islands

Wh-complementizers block filler—gap dependencies:

I know what Alex said...

- ✓ *...your friend devoured █ at the party.*
[null complementizer]
- ✓ *...that your friend devoured █ at the party.*
[that complementizer]
- ...whether your friend devoured █ at the party.*
[wh-complementizer]

Syntactic islands

Wh-complementizers block filler—gap dependencies:

I know what Alex said...

- ✓ ...*your friend devoured* █ *at the party.*
[null complementizer]
- ✓ ...***that*** *your friend devoured* █ *at the party.*
[that complementizer]
- * ...***whether*** *your friend devoured* █ *at the party.*
[wh-complementizer]

Syntactic islands

Wh-complementizers block filler—gap dependencies:

I know what Alex said...

- ✓ ...*your friend devoured* █ *at the party.*
[null complementizer]
- ✓ ...***that*** *your friend devoured* █ *at the party.*
[that complementizer]
- * ...***whether*** *your friend devoured* █ *at the party.*
[wh-complementizer]

Do the RNNs learn this?

✓

I know that my brother said our aunt devoured the cake at the party.

*

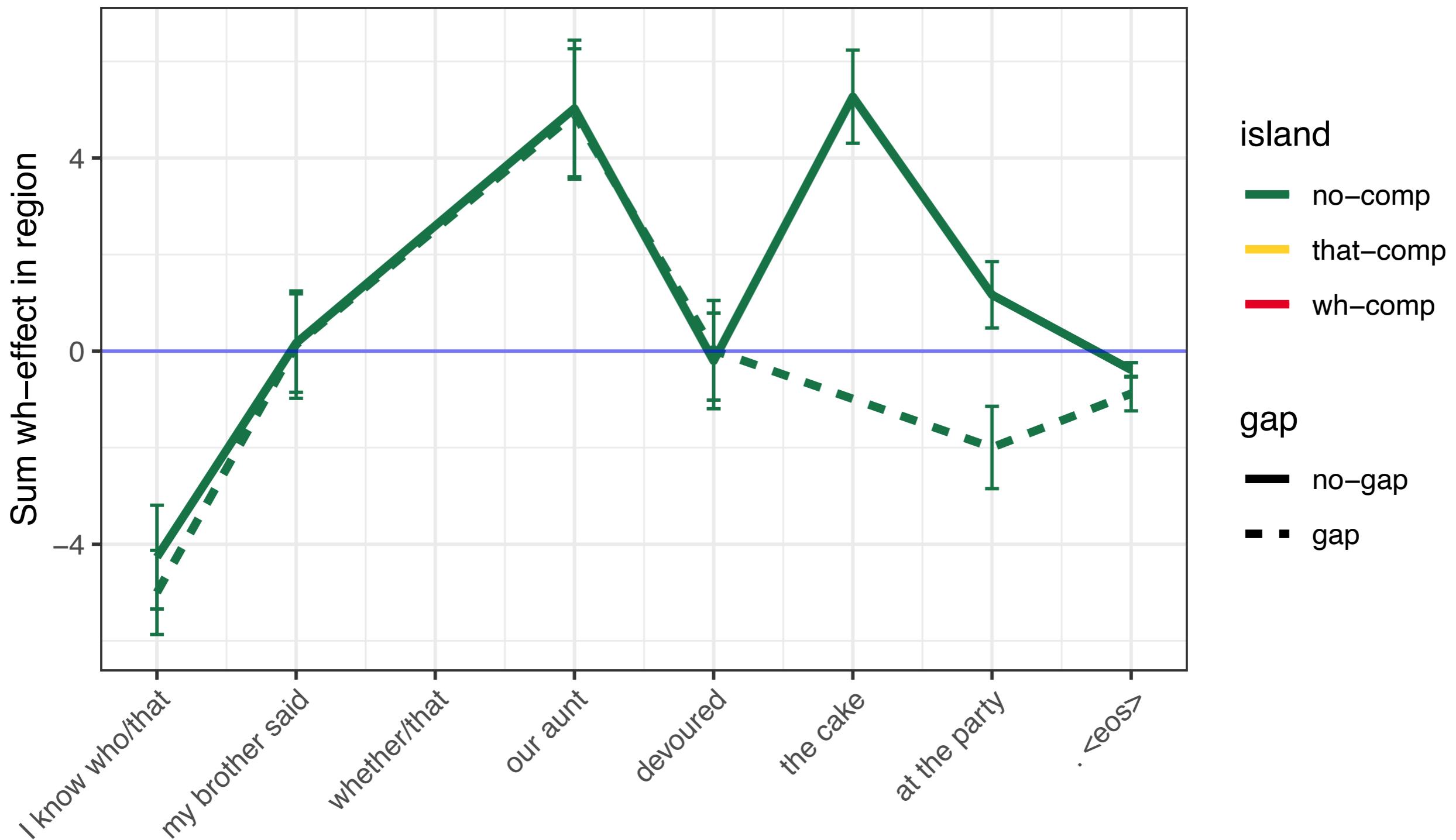
I know what my brother said our aunt devoured the cake at the party.

*

I know that my brother said our aunt devoured _____ at the party.

✓

I know what my brother said our aunt devoured _____ at the party.



✓

*I know that my brother said **that** our aunt devoured the cake at the party.*

*

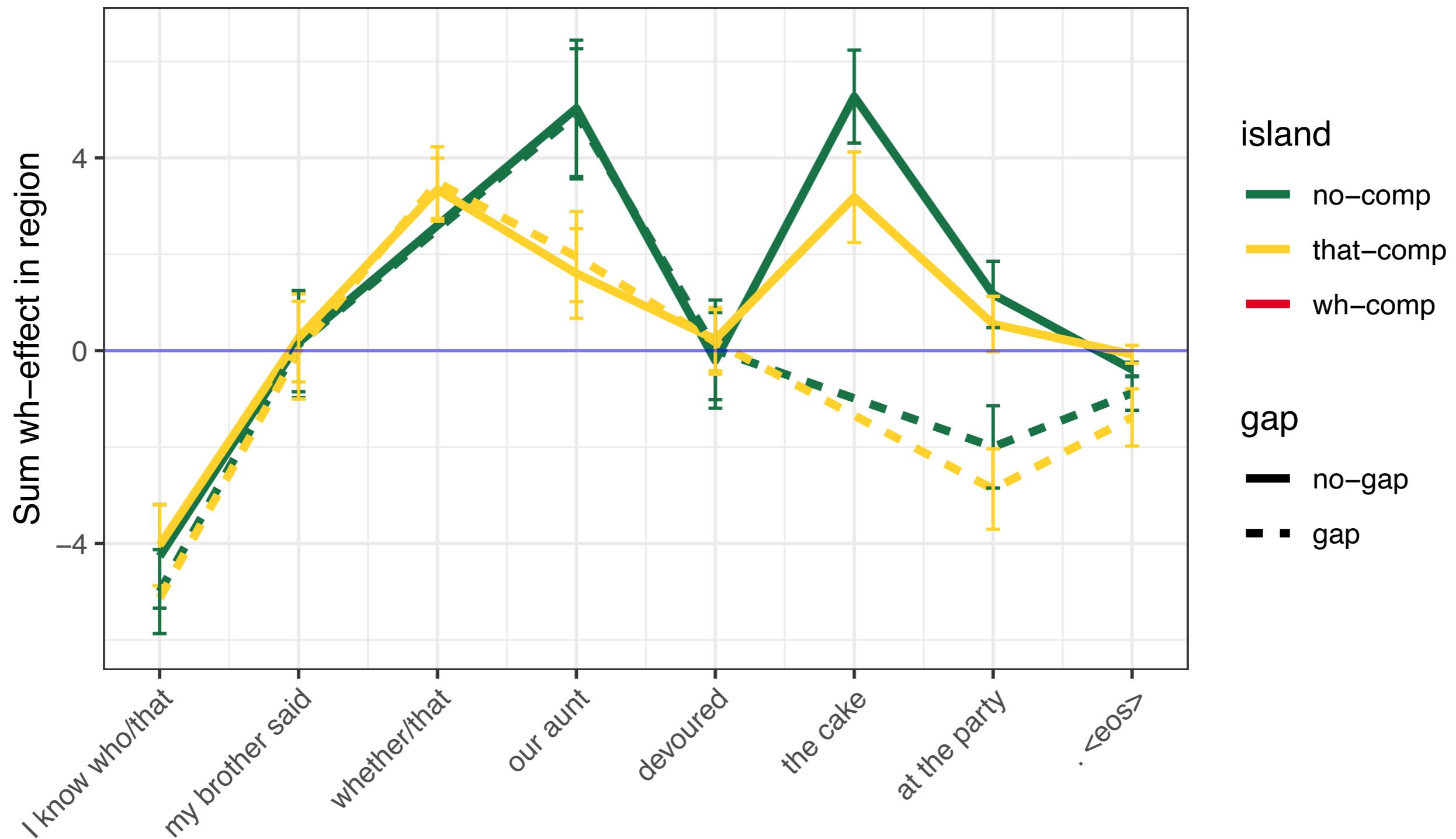
*I know **what** my brother said **that** our aunt devoured the cake at the party.*

*

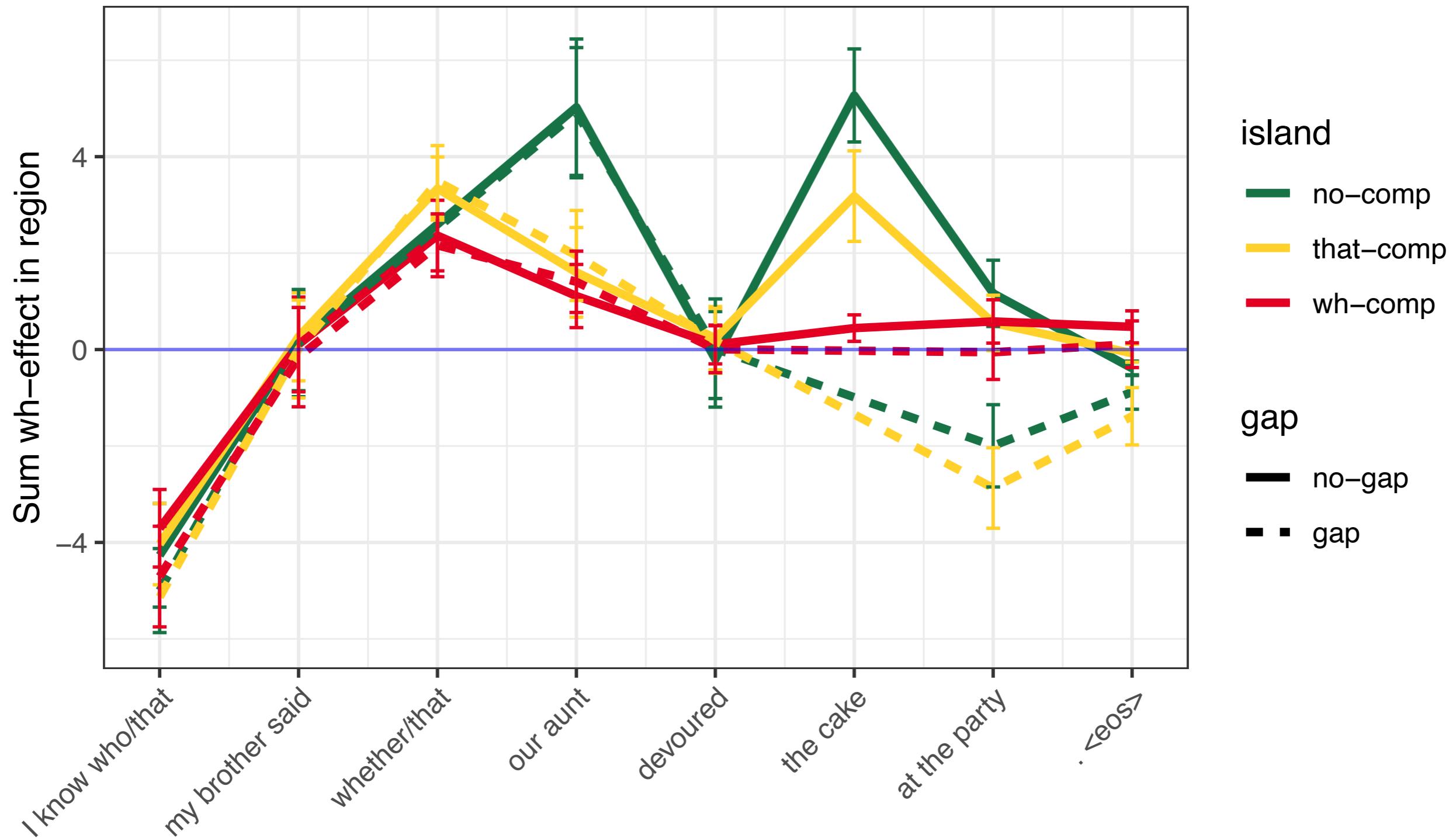
*I know that my brother said **that** our aunt devoured _____ at the party.*

✓

*I know **what** my brother said **that** our aunt devoured _____ at the party.*



- * I know that my brother said **whether** our aunt devoured the cake at the party.
- * I know **what** my brother said **whether** our aunt devoured the cake at the party.
- * I know that my brother said **whether** our aunt devoured _____ at the party.
- * I know **what** my brother said **whether** our aunt devoured _____ at the party.



Potential concern #2

Could RNNs have difficulty threading *any* type of expectation into a syntactic island?

Gendered-pronoun Expectation Control

Gendered-pronoun Expectation Control

- Worry: Can the models thread **any** expectation into islands?

Gendered-pronoun Expectation Control

- Worry: Can the models thread **any** expectation into islands?
- Test with expectation for **gendered pronouns** set up by **culturally or morphologically gendered subjects**.

Gendered-pronoun Expectation Control

- Worry: Can the models thread **any** expectation into islands?
- Test with expectation for **gendered pronouns** set up by **culturally or morphologically gendered subjects**.
 - ✓ *The actress said that they insulted her friends.*
[CONTROL, MATCH]

Gendered-pronoun Expectation Control

- Worry: Can the models thread **any** expectation into islands?
- Test with expectation for **gendered pronouns** set up by **culturally or morphologically gendered subjects**.
 - ✓ *The actress said that they insulted her friends.*
[CONTROL, MATCH]
 - # *The actress said that they insulted his friends.*
[CONTROL, MISMATCH]

Gendered-pronoun Expectation Control

- Worry: Can the models thread **any** expectation into islands?
- Test with expectation for **gendered pronouns** set up by **culturally or morphologically gendered subjects**.

Gender Expectation
Effect (#-✓ should be
positive)

- ✓ *The actress said that they insulted her friends.*
[CONTROL, MATCH]
- # *The actress said that they insulted his friends.*
[CONTROL, MISMATCH]

Gendered-pronoun Expectation Control

- Worry: Can the models thread **any** expectation into islands?
- Test with expectation for **gendered pronouns** set up by **culturally or morphologically gendered subjects**.

Gender Expectation
Effect (#-✓ should be
positive)

- ✓ *The actress said that they insulted her friends.*
[CONTROL, MATCH]
- # *The actress said that they insulted his friends.*
[CONTROL, MISMATCH]
- ✓ *The actress said whether they insulted her friends.*
[ISLAND, MATCH]

Gendered-pronoun Expectation Control

- Worry: Can the models thread **any** expectation into islands?
- Test with expectation for **gendered pronouns** set up by **culturally or morphologically gendered subjects**.

Gender Expectation
Effect (#-✓ should be
positive)

- ✓ *The actress said that they insulted her friends.*
[CONTROL, MATCH]
- # *The actress said that they insulted his friends.*
[CONTROL, MISMATCH]
- ✓ *The actress said whether they insulted her friends.*
[ISLAND, MATCH]
- # *The actress said whether they insulted his friends.*
[ISLAND, MISMATCH]

Gendered-pronoun Expectation Control

- Worry: Can the models thread **any** expectation into islands?
- Test with expectation for **gendered pronouns** set up by **culturally or morphologically gendered subjects**.

Gender Expectation
Effect (#-✓ should be
positive)

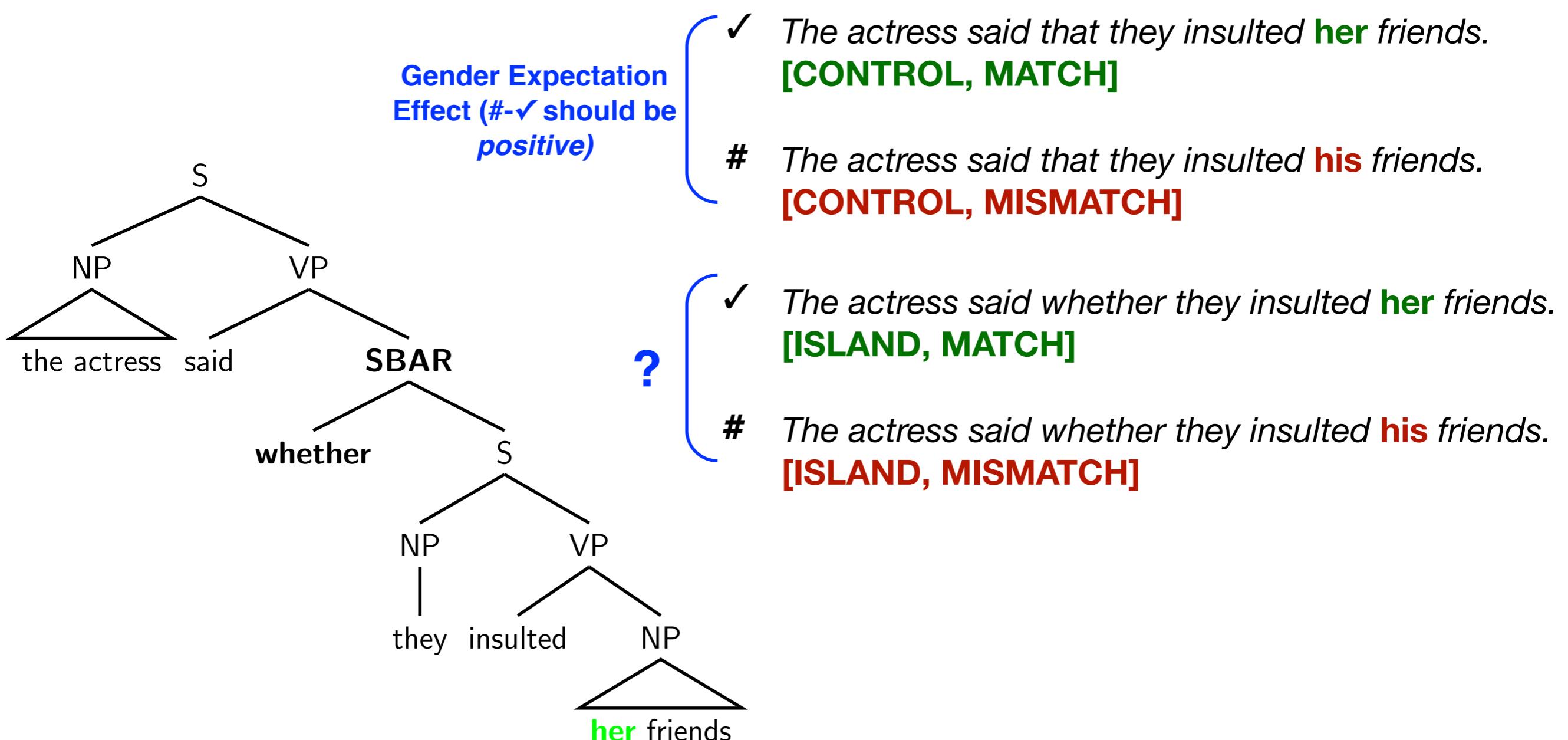
- ✓ *The actress said that they insulted her friends.*
[CONTROL, MATCH]
- # *The actress said that they insulted his friends.*
[CONTROL, MISMATCH]

?

- ✓ *The actress said whether they insulted her friends.*
[ISLAND, MATCH]
- # *The actress said whether they insulted his friends.*
[ISLAND, MISMATCH]

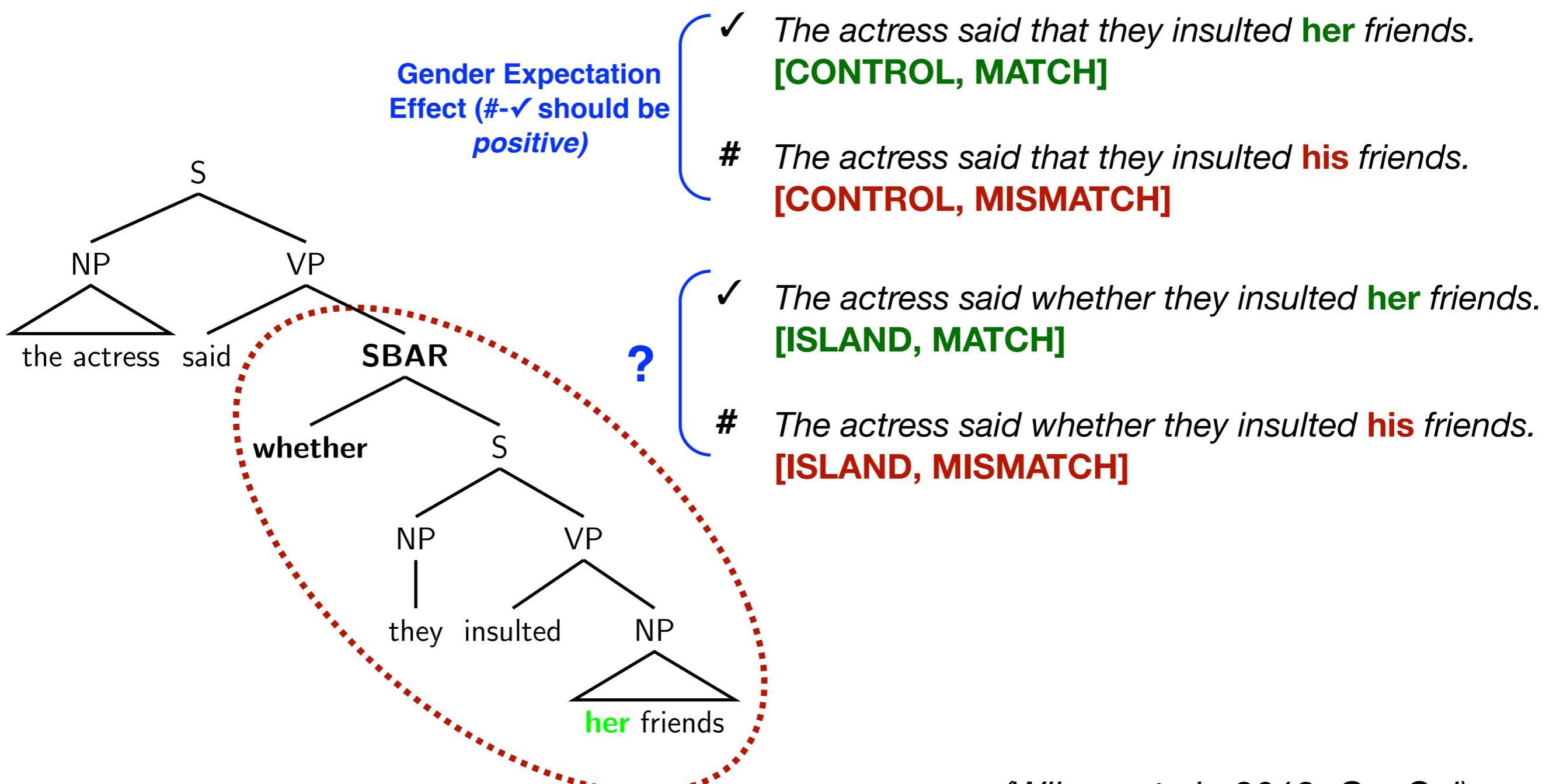
Gendered-pronoun Expectation Control

- Worry: Can the models thread **any** expectation into islands?
- Test with expectation for **gendered pronouns** set up by **culturally or morphologically gendered subjects**.



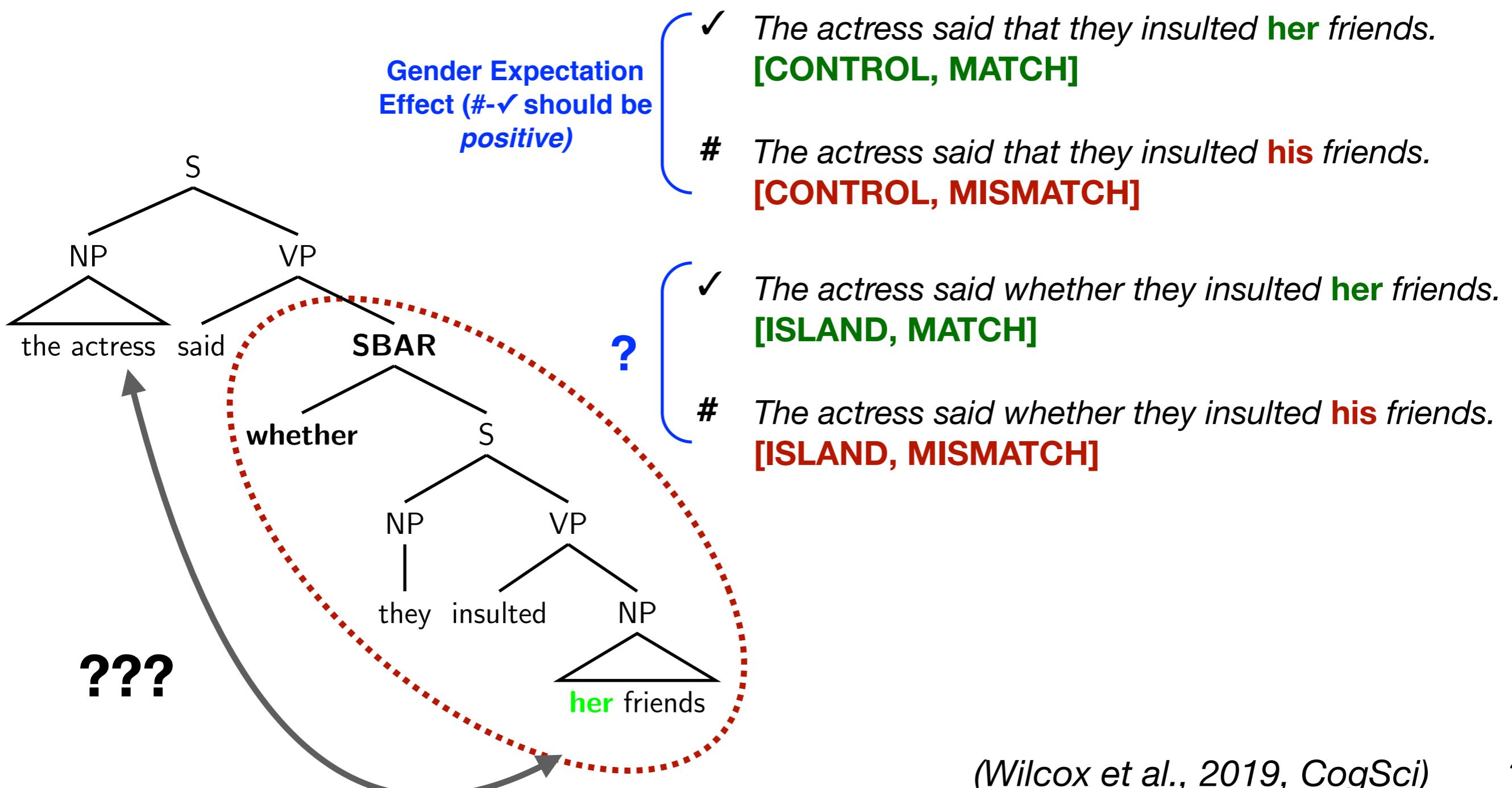
Gendered-pronoun Expectation Control

- Worry: Can the models thread **any** expectation into islands?
- Test with expectation for **gendered pronouns** set up by **culturally or morphologically gendered subjects**.



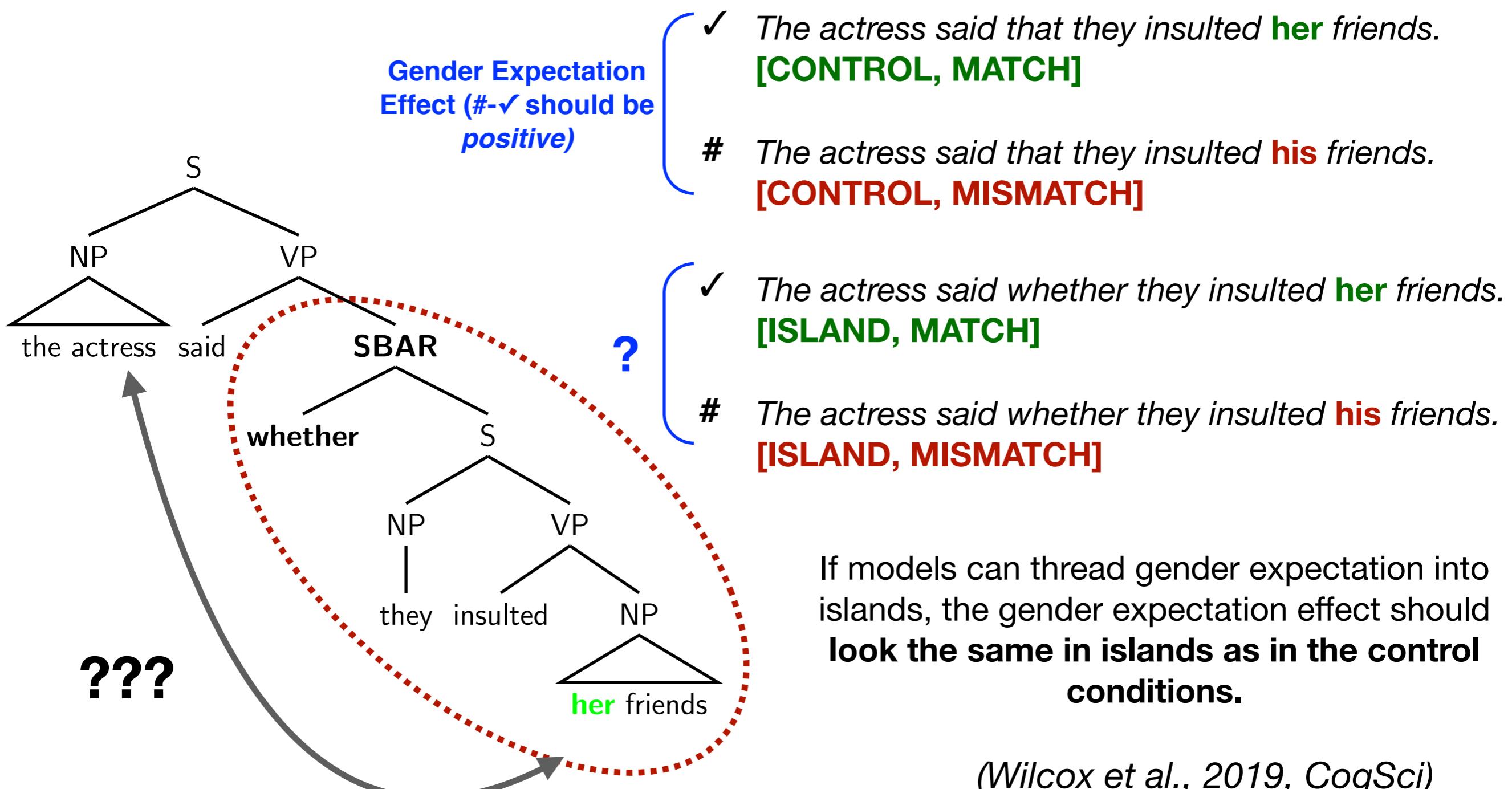
Gendered-pronoun Expectation Control

- Worry: Can the models thread **any** expectation into islands?
- Test with expectation for **gendered pronouns** set up by **culturally or morphologically gendered subjects**.



Gendered-pronoun Expectation Control

- Worry: Can the models thread **any** expectation into islands?
- Test with expectation for **gendered pronouns** set up by **culturally or morphologically gendered subjects**.

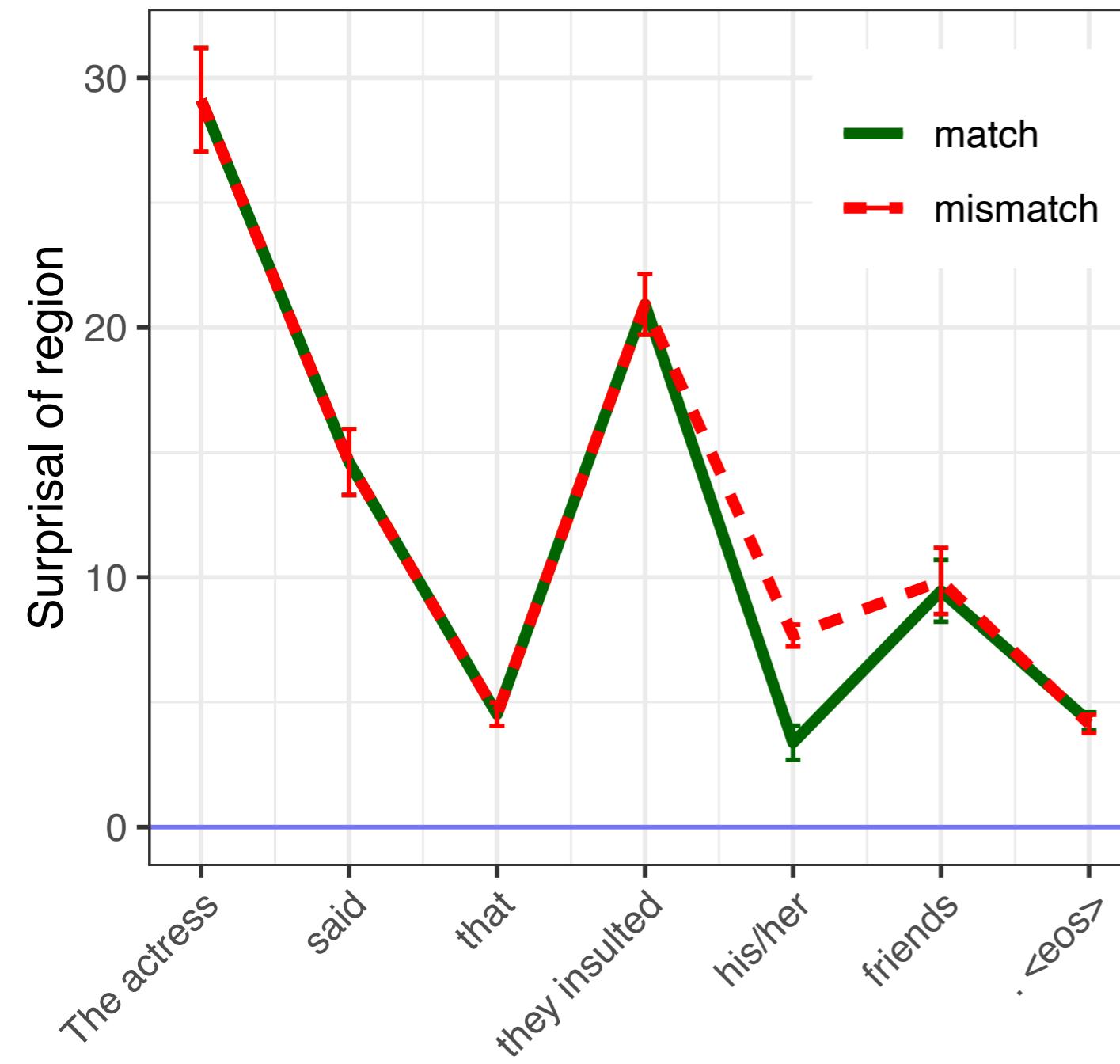


The actress said that they insulted her friends.

The actress said that they insulted his friends.

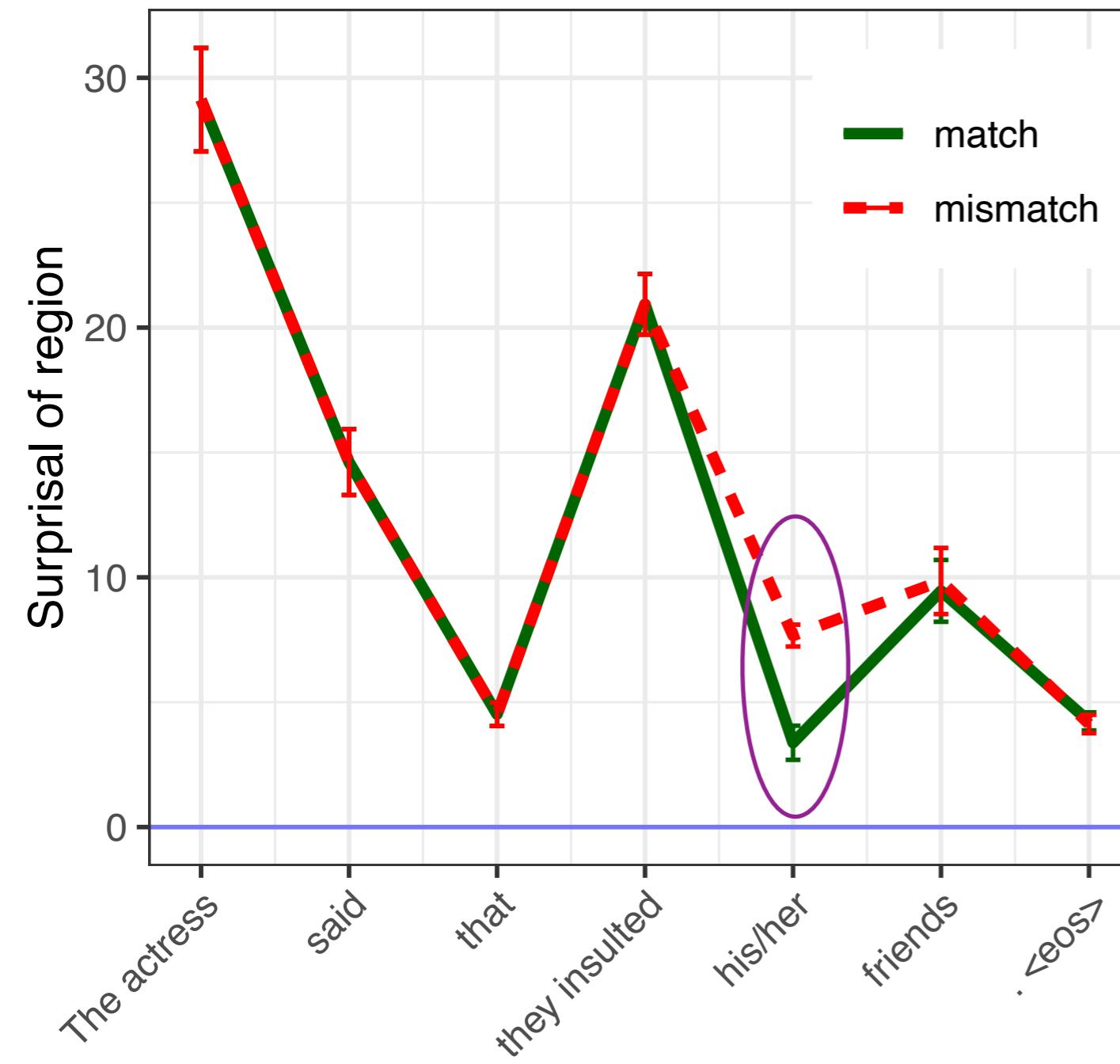
The actress said that they insulted her friends.

The actress said that they insulted his friends.



The actress said that they insulted her friends.

The actress said that they insulted his friends.

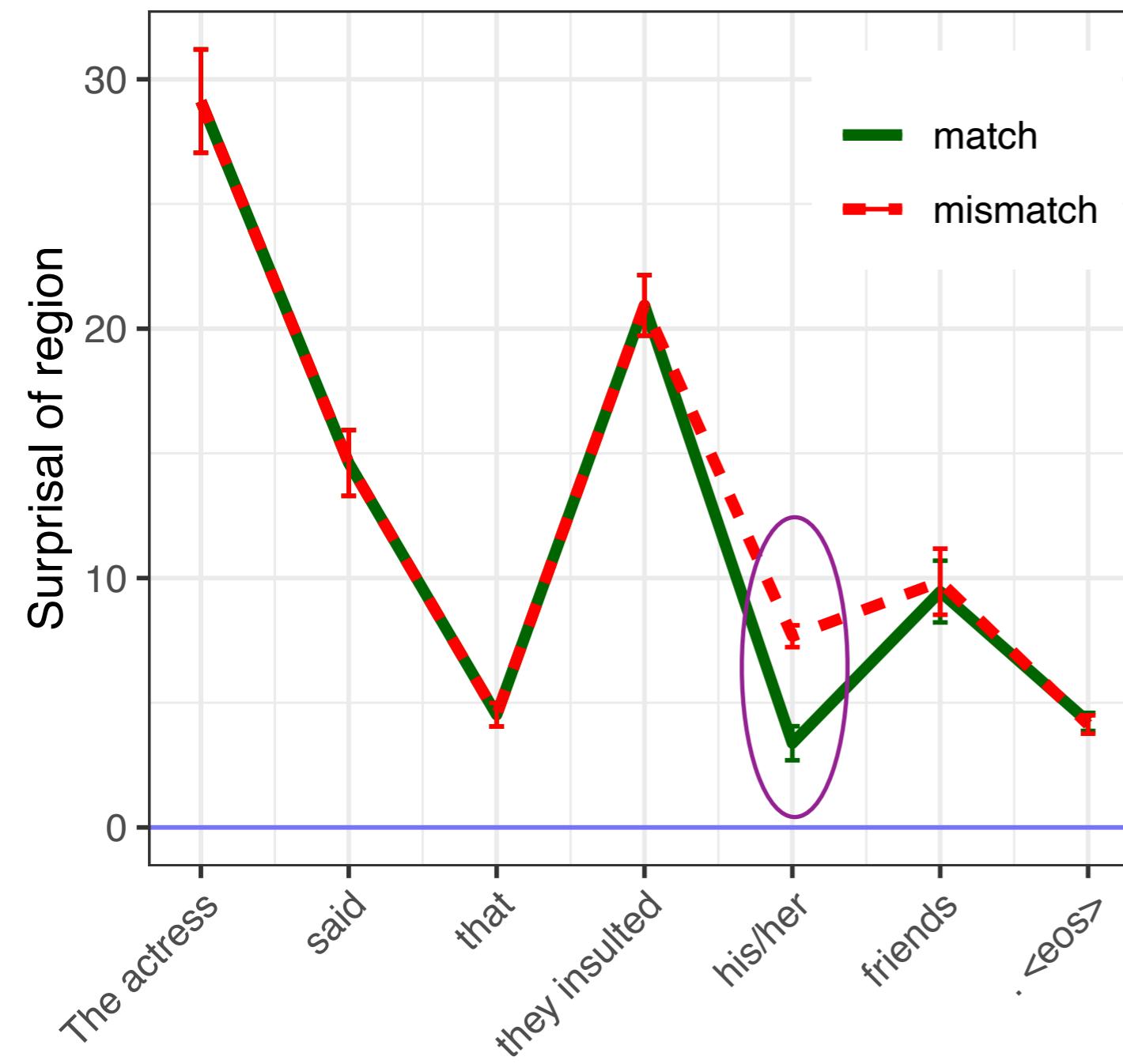


*The actress said that they insulted **her** friends.*

*The actress said that they insulted **his** friends.*

*The actress said whether they insulted **her** friends.*

*The actress said whether they insulted **his** friends.*

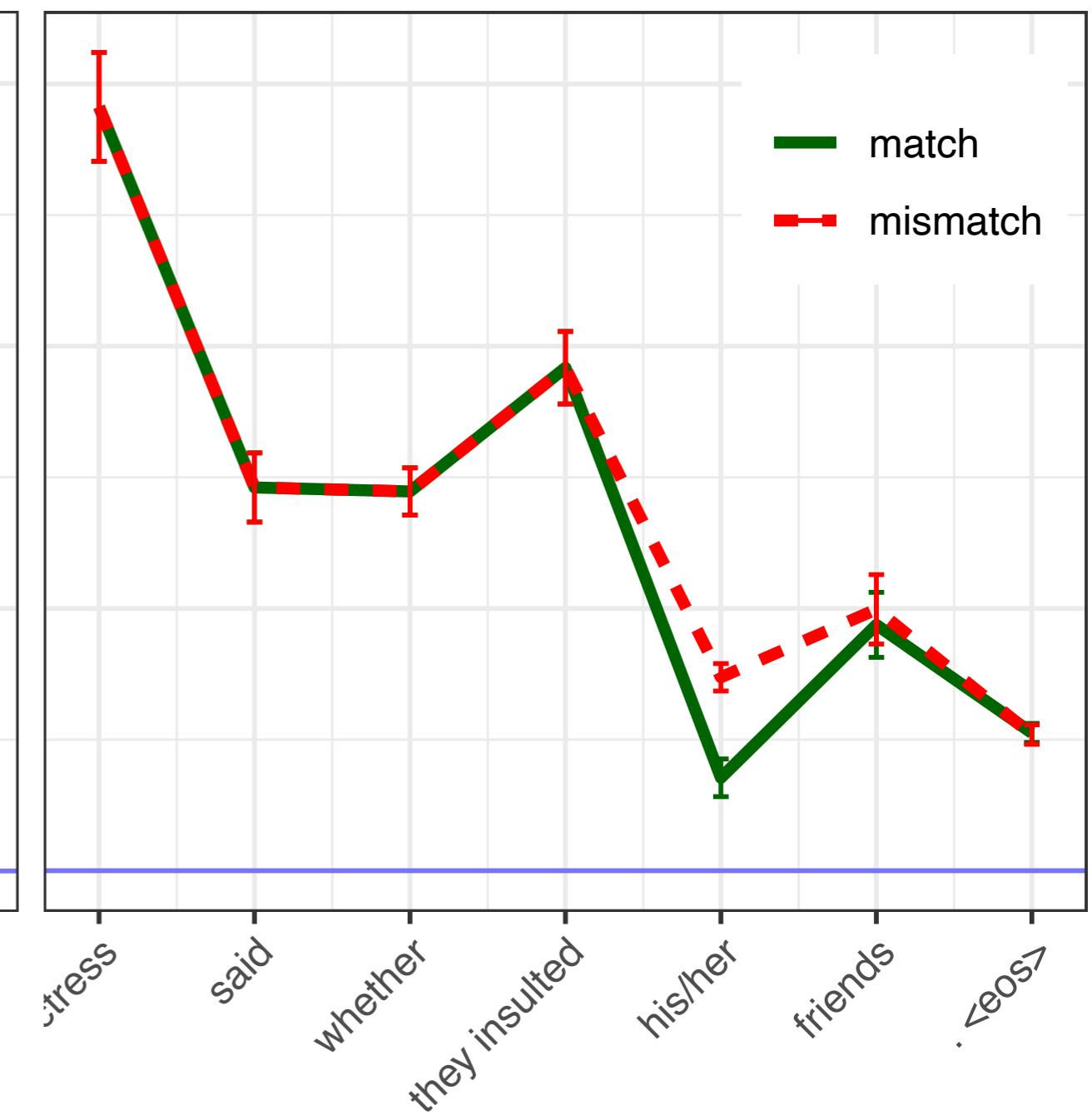
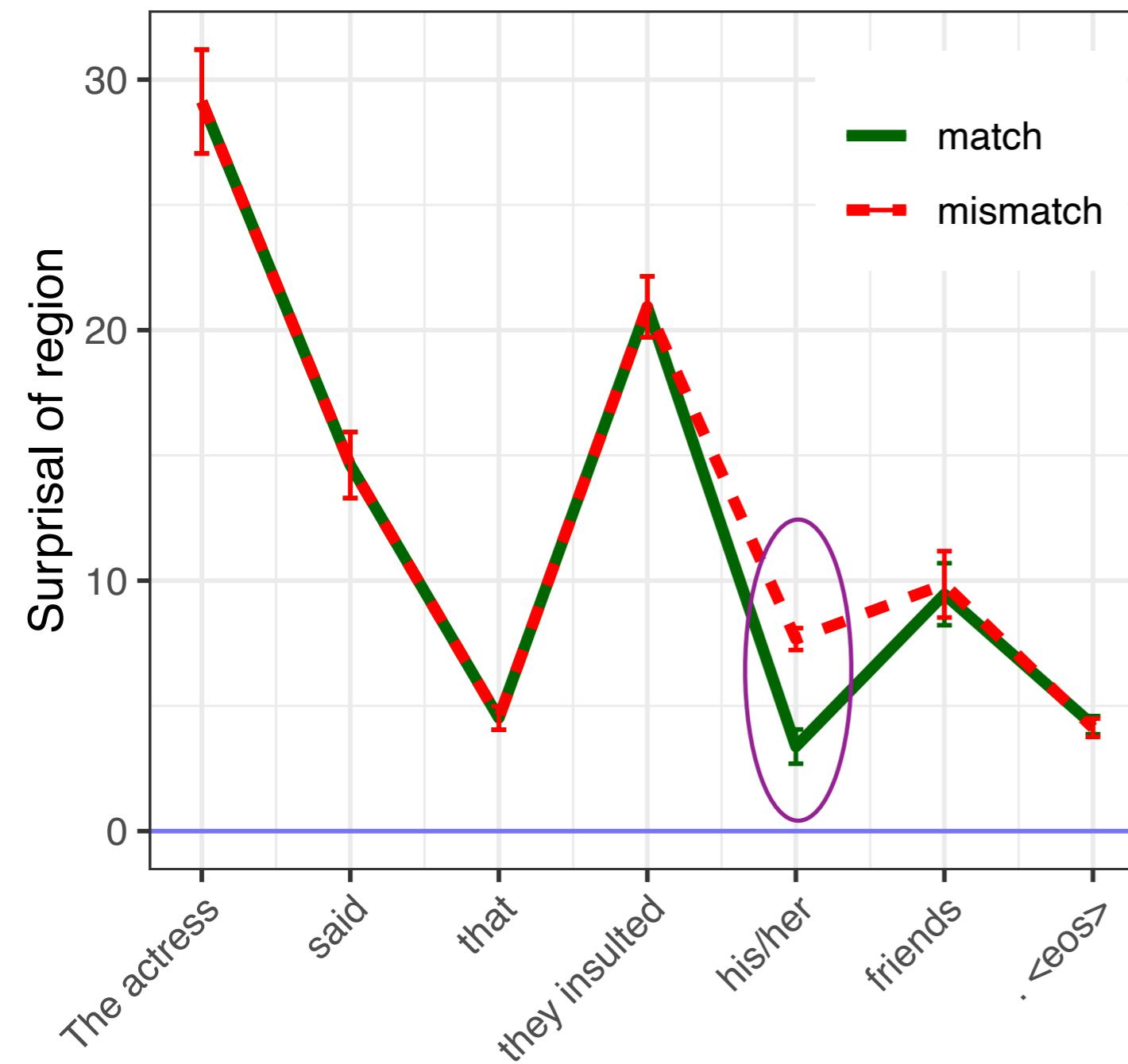


*The actress said that they insulted **her** friends.*

*The actress said that they insulted **his** friends.*

*The actress said whether they insulted **her** friends.*

*The actress said whether they insulted **his** friends.*

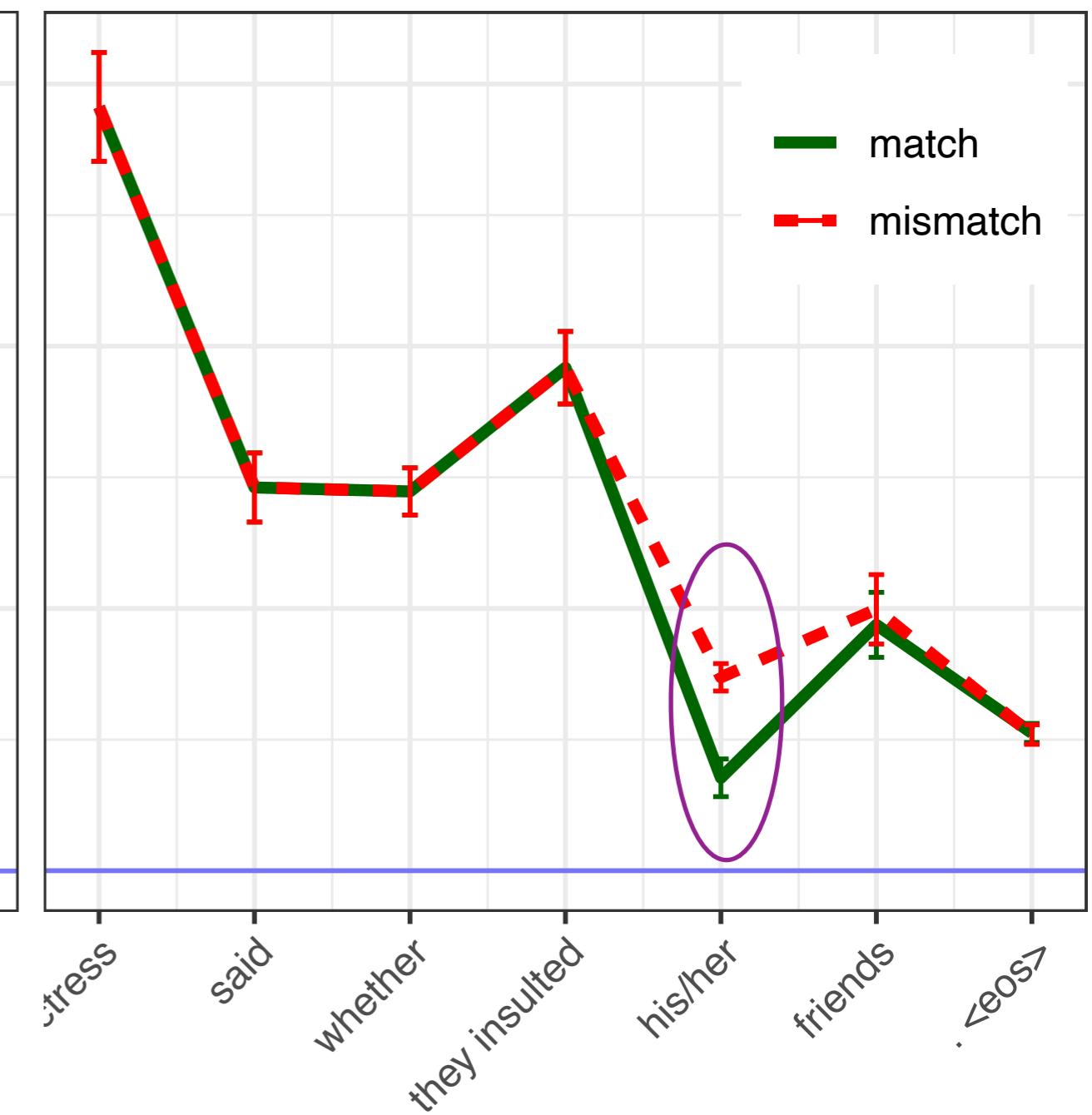
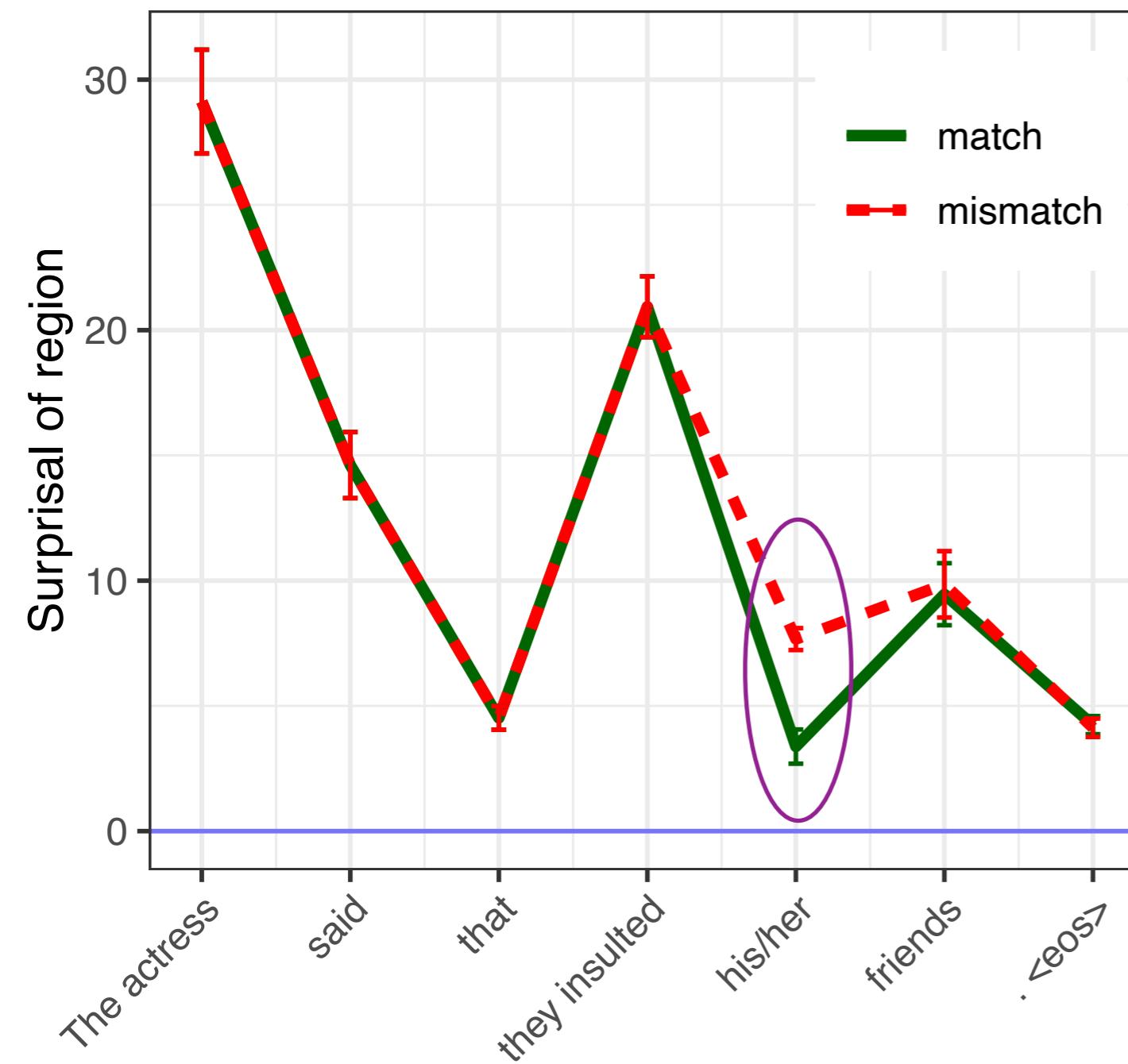


*The actress said that they insulted **her** friends.*

*The actress said that they insulted **his** friends.*

*The actress said whether they insulted **her** friends.*

*The actress said whether they insulted **his** friends.*



Potential concern #2

Could RNNs have difficulty threading *any* type of expectation into a syntactic island?

Potential concern #2 — *addressed*

Could RNNs have difficulty spreading *any* type of expectation into a syntactic island?



RNN models that learn island constraints still propagate pronoun gender expectations into islands

References

- Ba, J. L., Kiros, J. R., & Hinton, G. E. (2016). Layer normalization. arXiv preprint arXiv:1607.06450.
- Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. CoRR, abs/1409.0473. arXiv: 1409.0473
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of the 17th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.
- Gauthier, J., Hu, J., Wilcox, E., Qian, P., & Levy, R. P. (2020). SyntaxGym: An online platform for targeted evaluation of language models. In Proceedings of the 58th annual meeting of the Association for Computational Linguistics.
- Hu, J., Gauthier, J., Qian, P., Wilcox, E., & Levy, R. P. (2020). A systematic assessment of syntactic generalization in neural language models. In Proceedings of the 58th annual meeting of the Association for Computational Linguistics.
- Liu, P. J., Saleh, M., Pot, E., Goodrich, B., Sepassi, R., Kaiser, L., & Shazeer, N. (2018). Generating wikipedia by summarizing long sequences. In Proceedings of ICLR.
- Luong, T., Pham, H., & Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. In Proceedings of the 2015 conference on empirical methods in natural language processing (pp. 1412–1421).
- Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. In Proceedings of Neural Information Processing Systems (pp. 5998–6008).
- Wilcox, E., Levy, R. P., & Futrell, R. (2019). What syntactic structures block dependencies in RNN language models? In Proceedings of the 41st annual meeting of the Cognitive Science Society (pp. 1199–1205).
- 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.
- Wilcox, E., Qian, P., Futrell, R., Ballesteros, M., & Levy, R. (2019). Structural supervision improves learning of non-local grammatical dependencies. In Proceedings of the 18th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 3302–3312).