The Role of Context Gating in Predictive Sentence Processing

Yasemin Gokcen

How do we process sentences?

During language comprehension, humans routinely use information from preceding context to implicitly predict upcoming input.

This is thought to be a core computation in language comprehension.

Ryskin & Nieuwland (2023), Federmeier et al. (2015), Dell & Chang (2014)

Surprisal Theory

I'm going to return my book to the…

Library → expected, low surprisal, high probability with context 

Dog → not expected, high surprisal, lower probability with context 

Surprisal

- Measures of how unexpected a word is, given context
- Surprisal = -logP(word | context)

Levy (2008), Frank et al. (2015), Shain et al. (2024), Smith & Levy (2013)

Behavioral Measure of Surprisal

- Surprisal estimates from language models associated with reading times
  - Smith & Levy (2008), Frank (2009), Shain et al. (2024)
  - Higher surprisal positively correlated with higher reading times/eye fixations

We have established the importance of prediction in sentence processing…

What mechanisms *enable us to predict?*

Context Memory

- Need some memory for previous words to predict the next word
- Memory representation of the context is lossy
- Not every word contributes to next-word prediction equally

I'm going to return my book to the…

I'm going to return my book to the…

?? ?? ?? return ?? book ?? the…

Hahn et al. (2022), van Schijndel & Linzen (2021)

Memory for Context

- Working memory
  - Adaptive
  - Selecting information to track
    - Based on current situation
    - Ex: driving at a busy intersection
  - Help us keep track of *context*
    - From “I’m going to return my book to the…”
    - Hold onto return and book

Noelle (2012), Miyake & Shah (1999)

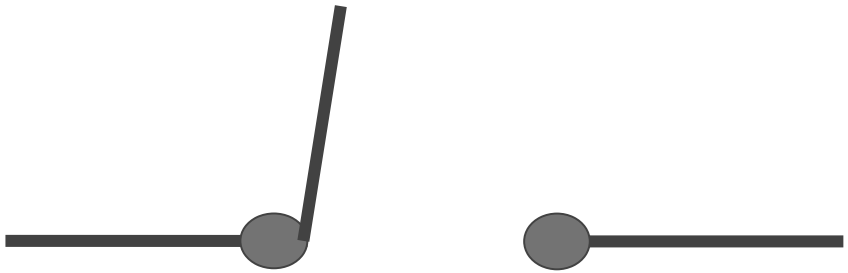
Working Memory and Computational Models

- Computational models of working memory (WM) mechanisms associated with prefrontal cortex (PFC) use *gating*
  - Capture ability to actively maintain and rapidly update contextual/task-relevant information in memory.

- Anatomy and physiology of PFC is thought to support *gating*

Hazy et al. (2013), Noelle (2012)

Similar gating mechanisms may be involved in context memory for sentence processing



Return Book To The…

Research Question

To what degree are the gating mechanisms of prefrontal cortex used to maintain context for the language network?

Methodology Outline

Comparing gating vs non-gating models

- Ability to predict the next word
- Architecture difference
- Compare to WM demand measures

Language Learning Models

- Compare performance across language models
  - Trained to predict next word
  - Different architectures
  - Recurrent Neural Networks (RNN) and Long Short-Term Memory (LSTM)

***LSTM has PFC-like gating mechanisms*, RNN does not**

  - Trained on the Wikitext 2 dataset and tested on Natural Stories (Futrell et al., 2018)
  - Natural Stories corpus has sentences of various linguistic properties; used in neuroimaging studies (Shain et al., 2022; Whebe et al., 2021)

Model Architectures

If you were to journey to the North of England, you would come to a valley that is surrounded by moors as high as mountains.

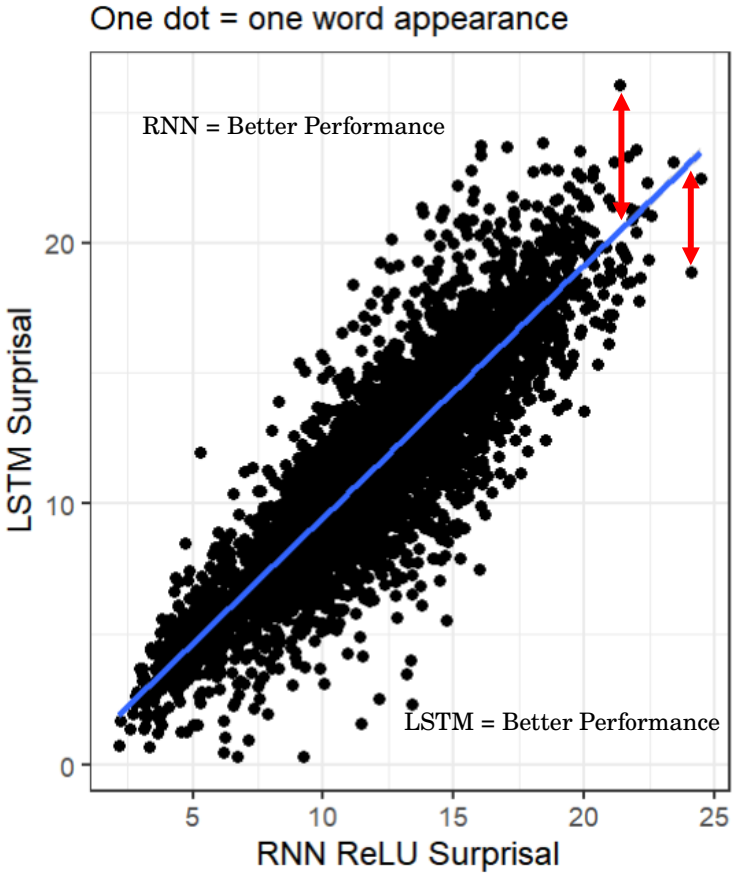
Word	LSTM	RNN(Tanh)	RNN(ReLU)
If	10.3523	8.3332	9.0970
You	4.5668	6.1161	6.7887
Were	5.1052	5.3894	4.8747
To	5.5413	5.2790	6.4274
Journey	17.3029	15.7554	14.2847

- We trained and tested 3 different model types
  - LSTM
  - RNN
    - 2 different activation functions: Tanh and ReLU
- Each model produces surprisal values for each word

Van Schijndel & Linzen (2018), Hochreiter & Schmidhuber (1997)

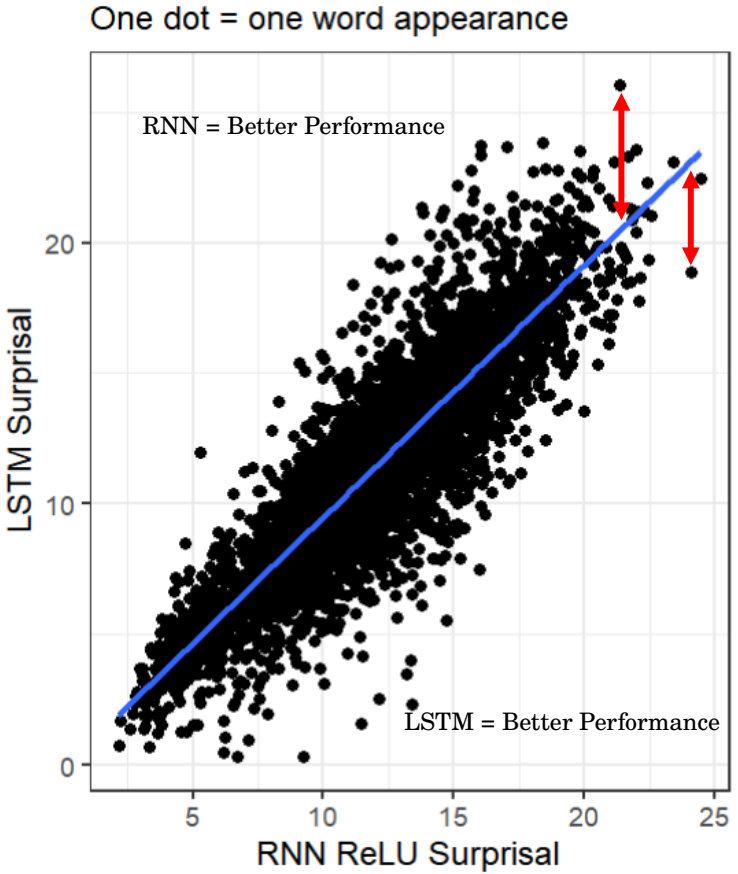
How do we compare model performance?

- Compare all of the surprisal values for each word
- Compute the *residual*
  - How well does the RNN predict the LSTM value?



How do we compare model performance?

- For each model and word we have:
  - Residual values that represent performance difference
- Negative → LSTM is performing better
  - Gating may be playing an important role



Working Memory Demand Measures

- 3 theoretical frameworks to access WM demand (Shain et al., 2022)
  - Dependency Locality Theory
    - How far words are away from each other
  - ACT-R model
    - Temporal costs
    - Memory decay
  - Left-Corner Parsing
    - e.g., How many phrases are between words?

Calculated and implemented for each word independent of the LSTM/RNN models

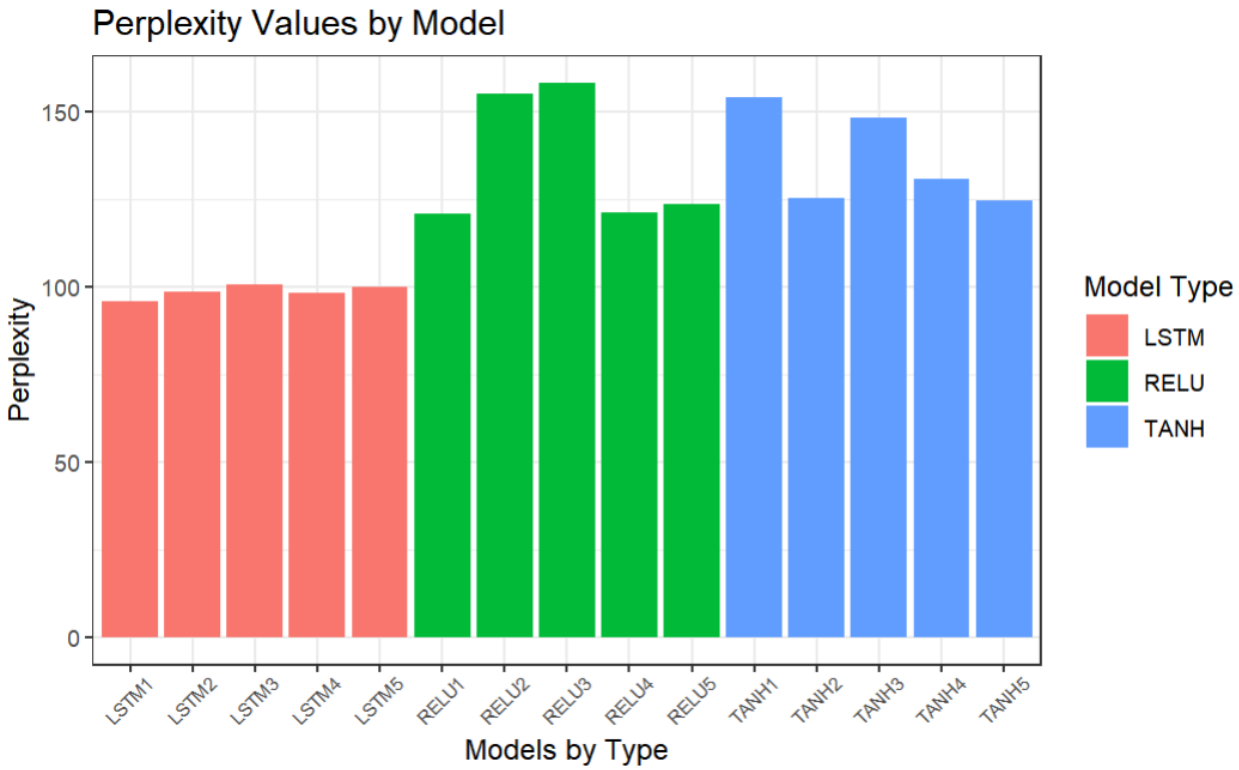
WM Measures & Residuals

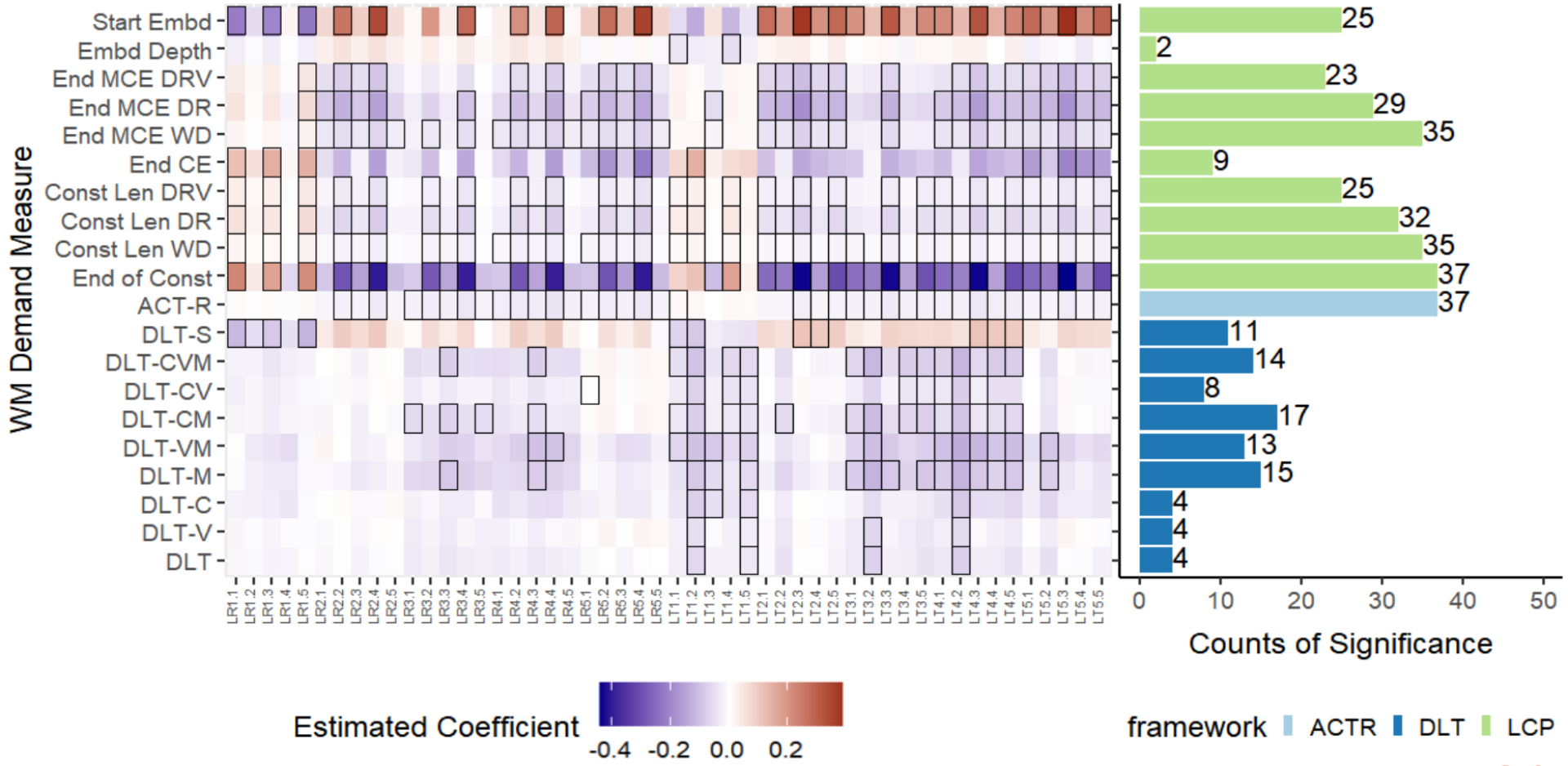
- We computed 5 iterations of each model (all with different seeds)
  - 650 hidden units, 128 batch size
  - Total of 15 models

- Computed for each combination (LSTM ~ Tanh and LSTM ~ ReLU)
- $LSTM_1$ vs $Tanh_1$, $LSTM_2$ vs $Tanh_1$, $LSTM_3$ vs $Tanh_1$ ….
  - 25 residuals for LSTM ~ Tanh + 25 residuals for LSTM ~ ReLU **Is gating related to WM demands?**
  - Correlate Residuals with WM demand measures for each

Comparing Models by Surprisal

- All three models are performing similarly
- LSTMs generally performing better
- Slight differences across iterations




WM Demand Measure
Start Embd
Embd Depth
End MCE DRV
End MCE DR
End MCE WD
End CE
Const Len DRV
Const Len DR
Const Len WD
End of Const
ACT-R
DLT-S
DLT-CVM
DLT-CV
DLT-CM
DLT-VM
DLT-M
DLT-C
DLT-V
DLT
LR1.1 LR1.2 LR1.3 LR1.4 LR1.5 LR2.1 LR2.2 LR2.3 LR2.4 LR2.5 LR3.1 LR3.2 LR3.3 LR3.4 LR3.5 LR4.1 LR4.2 LR4.3 LR4.4 LR4.5 LR5.1 LR5.2 LR5.3 LR5.4 LR5.5 LT1.1 LT1.2 LT1.3 LT1.4 LT1.5 LT2.1 LT2.2 LT2.3 LT2.4 LT2.5 LT3.1 LT3.2 LT3.3 LT3.4 LT3.5 LT4.1 LT4.2 LT4.3 LT4.4 LT4.5 LT5.1 LT5.2 LT5.3 LT5.4 LT5.5
25
2
23
29
35
9
25
32
35
37
37
11
14
8
17
13
15
4
4
4
0 10 20 30 40 50
Counts of Significance
Estimated Coefficient
-0.4 -0.2 0.0 0.2
framework ACTR DLT LCP

Summary

- We can use model surprisal values as estimates for human processing difficulty with processing words
- Gating appears to be important for contexts with higher WM demand
- DLT measures were rarely significantly correlated with residuals
- Constituent and many embedding measures more often significantly correlated with residuals
- LSTM does relatively better (compared to RNN) as these measures of WM demand increase
  - Regardless of activation function

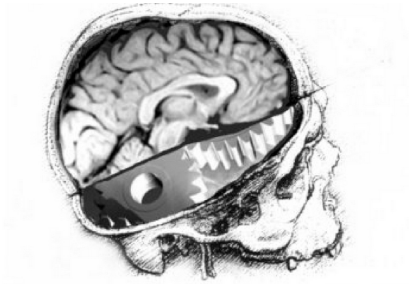
Back to our research question

- To what degree are the gating mechanisms of prefrontal cortex used to maintain context for the language network?
  - There are types of sentence structures in which gating is helpful in lowering surprisal.
  - The language network in the brain could make use of PFC-like gating when WM demand is high.

Future Directions

- Analyze changes in hidden layer activation from word-to-word
- Examine gating units for each cell in the LSTM
  - Can we identify where/when active maintenance and rapid updating occurs in a sentence?
- **EEG data collection in progress! → correlate with model predictions**

Thanks!



Computational Cognitive Neuroscience Laboratory
University of California, Merced

Special thanks to Dr. David Noelle and Dr. Rachel Ryskin for mentorship and guidance throughout this project!

Further thanks to the CCN lab and LInC lab for extra feedback and camaraderie ☺