# Effects of Recency Bias on Transformers' Predictions of Reading Times

Christian Clark Byung-Doh Oh William Schuler

38th Annual Conference on Human Sentence Processing



# Surprisal-Based Models of Sentence Processing

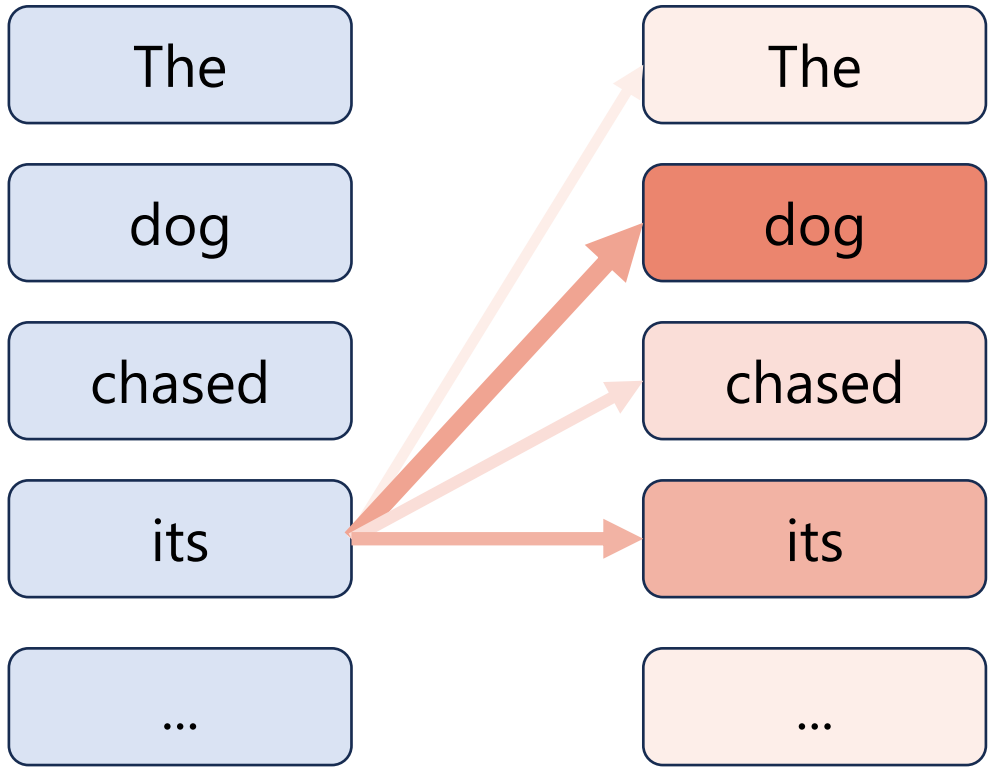
- Surprisal theory (Hale, 2001; Levy, 2008) connects a word's processing difficulty to its predictability in context:

$$\text{difficulty}(w_i) \propto -\log P(w_i | w_1, \ldots, w_{i-1})$$

- Creates a link between psycholinguistic modeling and language modeling
- Studies have compared surprisal estimates from a range of language models (e.g., Wilcox et al., 2020; Merkx and Frank, 2021)
  - *n*-gram models
  - Recurrent neural networks
  - Transformers
- Transformer surprisal shows a strong fit to human reading times

# Transformer Attention vs. Human Memory

- Transformers' attention mechanism resembles models of cue-based retrieval
  (Ryu and Lewis, 2021; Oh and Schuler, 2022; Timkey and Linzen, 2023; Yoshida et al., 2025)



# Transformer Attention vs. Human Memory

- Transformers' attention mechanism resembles models of cue-based retrieval (Ryu and Lewis, 2021; Oh and Schuler, 2022; Timkey and Linzen, 2023; Yoshida et al., 2025)
- But Transformers make other unrealistic assumptions about memory
  - Lossless representations
  - Long context window (text history)
  - Capable of "needle-in-haystack" tasks over long contexts (Gemini Team, 2024)
- This runs contrary to theories positing tight bounds on human memory (Chomsky and Miller, 1963; Miller and Isard, 1964; Gibson, 2000; Lewis and Vasishth, 2005)

# Recency Bias in Transformers

- We therefore try adding a **recency bias** to Transformers
  - Shifts some attention to more recent words
  - Resembles notions of decay or lossy context found in cognitive models (Lewis and Vasishth, 2005; Futrell et al., 2020)
- Surprisal estimates are taken from Transformers with and without recency bias
- Reading times are predicted from broad-coverage, naturalistic corpora
- Improved predictions from Transformers with ALiBi recency bias (Press et al., 2022)

# Outline

Background

Regression Experiments

Discussion

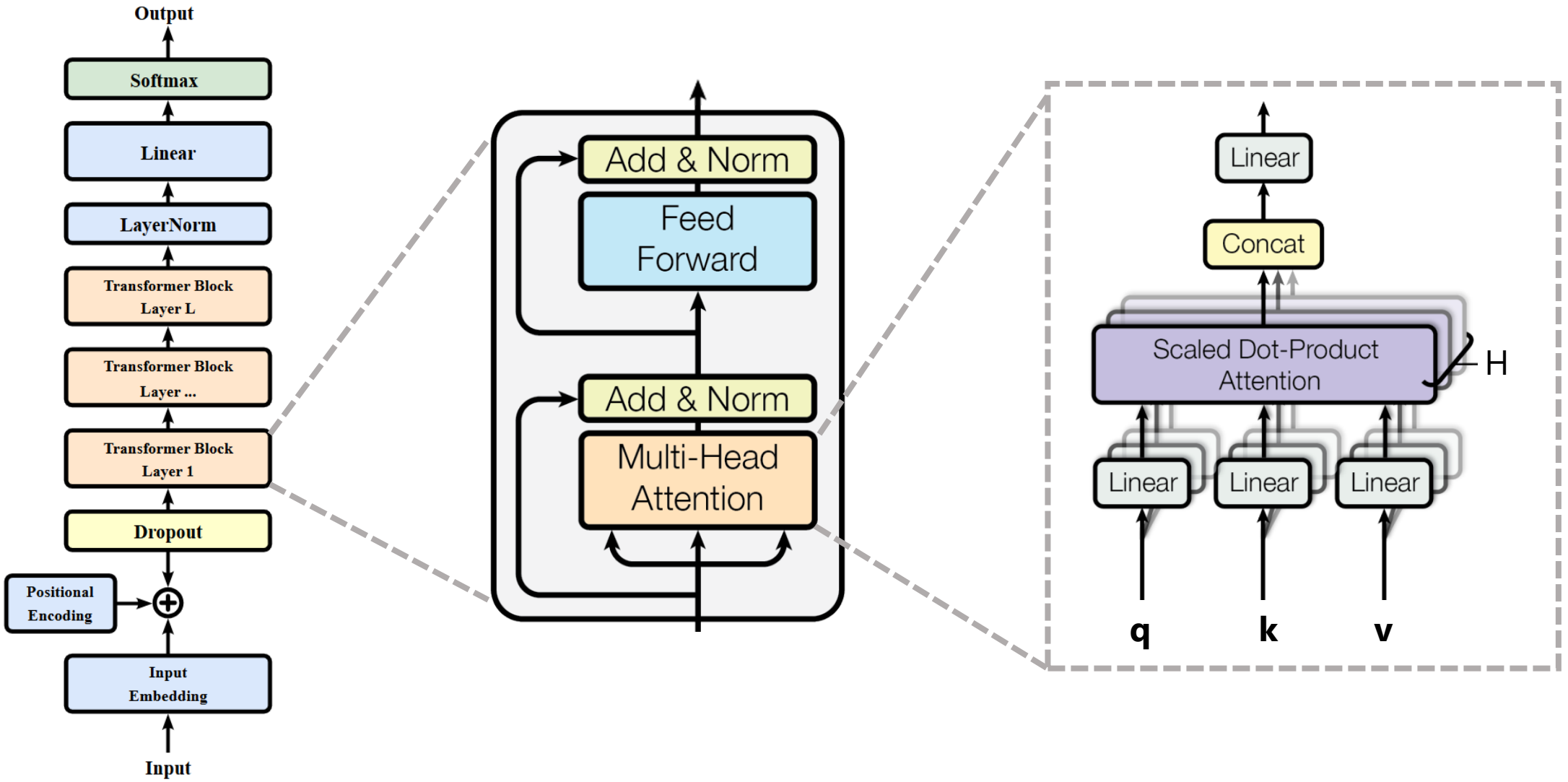
# Outline

Background

Regression Experiments

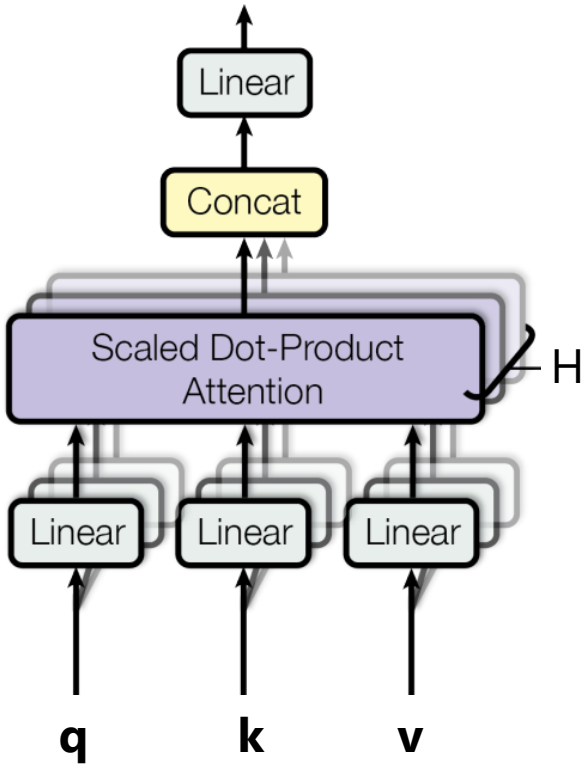
Discussion

# Transformer Architecture



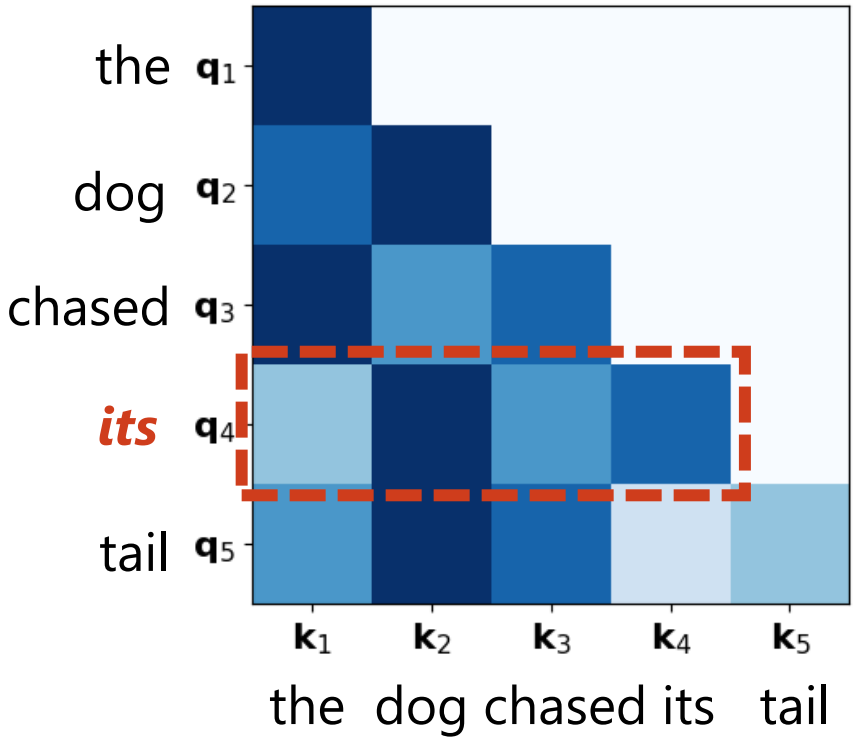
# Scaled Dot-Product Attention

- Uses query ($\mathbf{q}$), key ($\mathbf{k}$), value ($\mathbf{v}$) word vectors
- Steps when processing word $i$:
  1. Calculate similarity $z_{i,j}$ between $\mathbf{q}_i$ and each $\mathbf{k}_j$, $j \in [1..i]$
     - Scaled dot product: $z_{i,j} = \frac{1}{\sqrt{d}} \mathbf{q}_i^{\mathrm{T}} \mathbf{k}_j$
     - $d$ is the vector dimension
  2. Exponentiate and renormalize to get **attention scores**
     - $\mathrm{Attention}(i,j) = \sigma(\begin{bmatrix} z_{i,1} \\ \cdots \\ z_{i,i} \end{bmatrix})_j = \frac{e^{z_{i,j}}}{\sum_k e^{z_{i,k}}}$
  3. Combine value vectors based on attention scores
     - Used for probabilistic next-word predictions
- Multiple **attention heads** compute scores in parallel



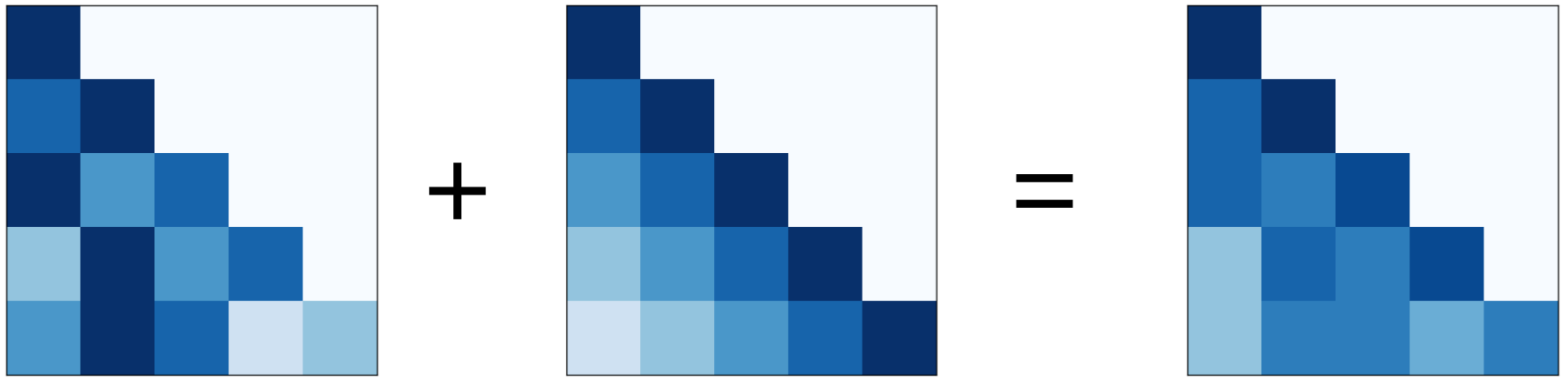
# Scaled Dot-Product Attention

- Uses query ($\mathbf{q}$), key ($\mathbf{k}$), value ($\mathbf{v}$) word vectors
- Steps when processing word $i$:
  1. Calculate similarity $z_{i,j}$ between $\mathbf{q}_i$ and each $\mathbf{k}_j$, $j \in [1..i]$
     - Scaled dot product: $z_{i,j} = \frac{1}{\sqrt{d}} \mathbf{q}_i^{\mathrm{T}} \mathbf{k}_j$
     - $d$ is the vector dimension
  2. Exponentiate and renormalize to get **attention scores**
     - $\text{Attention}(i,j) = \sigma(\begin{bmatrix} z_{i,1} \\ \cdots \\ z_{i,i} \end{bmatrix})_j = \frac{e^{z_{i,j}}}{\sum_k e^{z_{i,k}}}$
  3. Combine value vectors based on attention scores
     - Used for probabilistic next-word predictions
- Multiple **attention heads** compute scores in parallel



# Recency Bias

- A recency bias can be added to attention scores to upweight recent words




- We test two implementations of recency bias from previous work (de Varda and Marelli, 2024; Press et al., 2022)

# Recency Bias: de Varda and Marelli (2024)

- Add a recency bias term $\mathbf{b}_i \in \mathbb{R}^i$ to the similarity scores prior to renormalization
- Terms in the bias vector follow an exponential pattern: $\mathbf{b}_i[j] = e^{-\lambda(i-j)}$
- Hyperparameter $\lambda$ determines the rate of decay
- Modified attention formulation:

Query-key similarity scores

Recency bias

$$\text{Attention}_{\text{dVM}}(i,j) = \sigma(\alpha \mathbf{b}_i + (1-\alpha)\begin{bmatrix} z_{i,1} \\ \cdots \\ z_{i,i} \end{bmatrix})$$

- Hyperparameter $\alpha$ weights the bias and raw similarity scores
- Following the original study, we set $\lambda = 82.86$ and $\alpha = 0.37$

# Recency Bias: ALiBi (Press et al., 2022)

- Attention with Linear Biases
- Originally developed as a method for input length extrapolation
- Uses a linear recency bias term: $\mathbf{b}'_i \in \mathbb{R}^i$, where $\mathbf{b}'_i[j] = m \cdot (j - i)$
- Modified attention formulation:

Recency bias

Query-key similarity scores

$$\text{Attention}_{\text{ALiBi}}(i, j) = \sigma(\mathbf{b}'_i + \begin{bmatrix} z_{i,1} \\ \cdots \\ z_{i,i} \end{bmatrix})$$

- The hyperparameter $m$ determines the rate of decay
- Press et al. use a different decay rate for each attention head in a layer

# Outline

Background

Regression Experiments

Discussion

# Outline

Background

Regression Experiments

Discussion

# Experiments: Reading Time Corpora

- Reading times come from a set of broad-coverage psycholinguistic corpora

**Self-paced reading corpora:**

- Brown (Smith and Levy, 2013)
- Natural Stories (Futrell et al., 2021)
- UCL (Frank et al., 2013)

**Eye-tracking corpora:**

- UCL (Frank et al., 2013)
- GECO (Cop et al., 2017)
- Dundee (Kennedy et al., 2003)
- Provo (Luke and Christianson, 2018)

- Eye-tracking corpora included scan path, first-pass, and go-past durations
- Corpora were partitioned into fit, exploratory, and held-out partitions
  - Fit partition (50% of data points) used to fit regression models
  - Exploratory partition (25%) used for evaluation of regression models
  - Held-out partition (25%) used for statistical significance testing

# Experiments: Language Models

- The Transformers used in the experiments were based on Pythia models (Biderman et al., 2023)
  - Context window of 2048 tokens (word pieces)
  - Trained from scratch on the first 1000 batches (~2B tokens) of the Pile (Gao et al., 2020)
- 2 Transformer layers with 4 attention heads each (~27M parameters)
- Model settings were optimal values from a previous reading time study (Oh and Schuler, 2023)

# Experiments: Regression Modeling

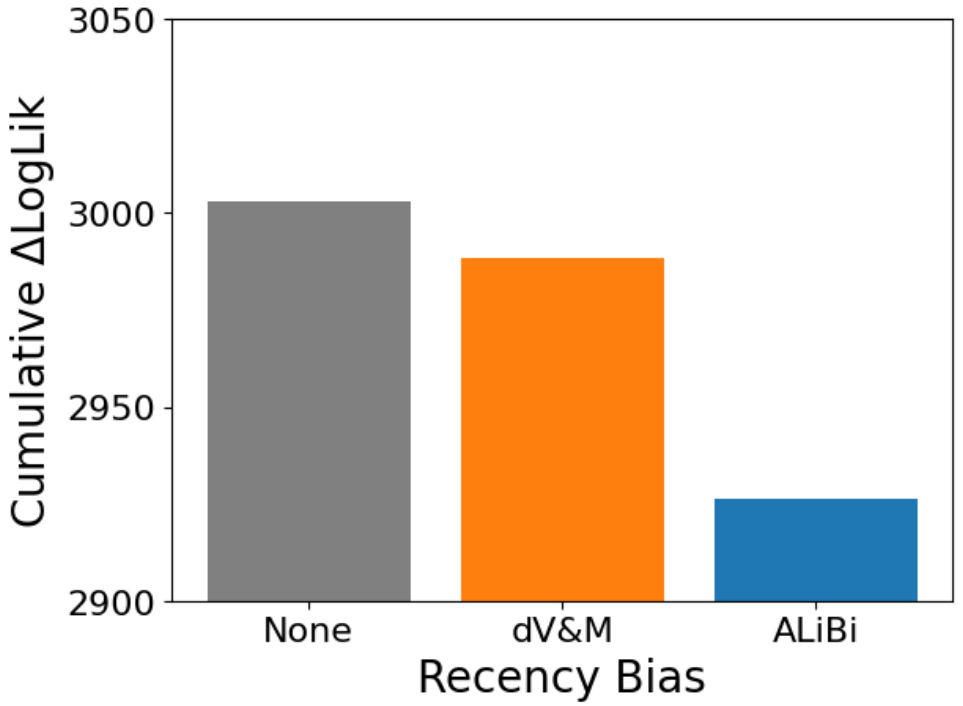
- Reading times predicted with linear mixed-effects regression models (Bates et al., 2015)
- Regression models included baseline predictors as well as Transformer surprisal
- Spillover predictors from the previous word $i-1$ were included for surprisal

$$\begin{aligned}\mathrm{RT}_{\text{Self-Paced Reading}} \sim\ & \mathrm{surp}_i + \mathrm{surp}_{i-1} + \mathrm{wordLen} + \mathrm{wordIndex} + \mathrm{unigramSurp} \\ & + (\mathrm{surp}_i + \mathrm{surp}_{i-1} + \mathrm{wordLen} + \mathrm{wordIndex} + 1 \mid \mathrm{subject}) \\ \mathrm{RT}_{\text{Eye Tracking}} \sim\ & \mathrm{surp}_i + \mathrm{surp}_{i-1} + \mathrm{wordLen} + \mathrm{wordIndex} + \mathrm{unigramSurp} \\ & + \mathrm{prevWordFixated} + (\mathrm{surp}_i + \mathrm{wordIndex} + 1 \mid \mathrm{subject})\end{aligned}$$

- Evaluation metric: increase in log likelihood (ΔLogLik)
  - Compares regression models with and without surprisal predictors
  - Aggregated over all corpora

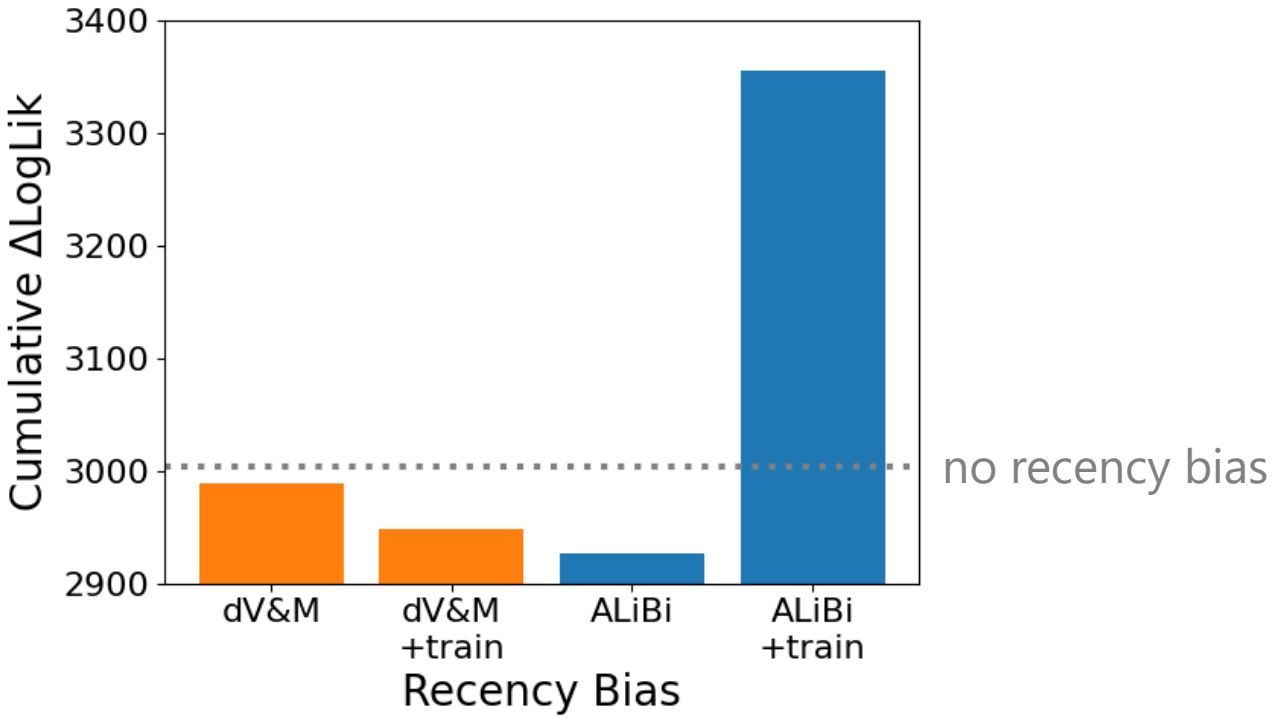
# Experiment 1: De Varda and Marelli (2024) Replication

- We do not replicate the improvement reported by dV&M
- We predict per-subject reading times; dV&M predict averages
- Like dV&M, we see variable results across individual corpora



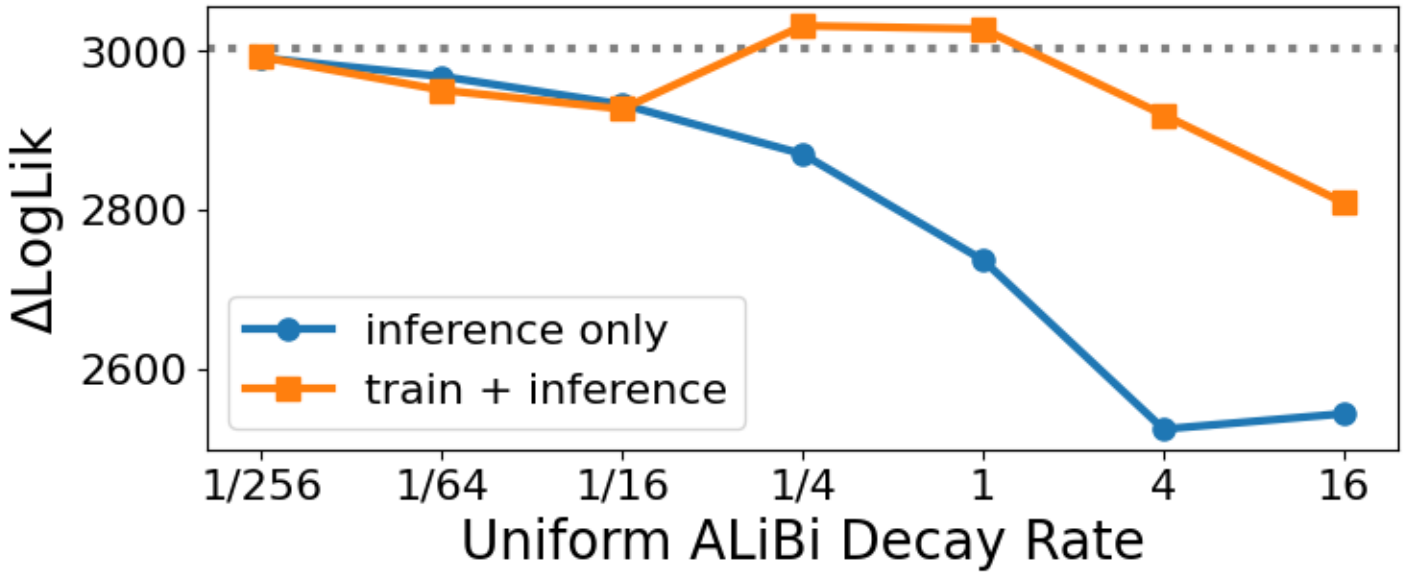
# Experiment 2: Recency Bias at Training and Inference

- Following Press et al. (2022), recency bias used throughout
- Still no ΔLogLik improvement from the de Varda and Marelli bias
- Significant improvement from ALiBi ($p < 0.001$)



# Experiment 3: Uniform ALiBi Decay Rate

- ALiBi (Press et al., 2022) uses separate decay rates across attention heads
  - For four attention heads: [1/4, 1/16, 1/64, 1/256]
- Cognitive models (e.g., Lewis and Vasishth, 2005) typically use a single decay rate
- Experiment 3 accordingly simplified ALiBi to only use one decay rate




no recency bias

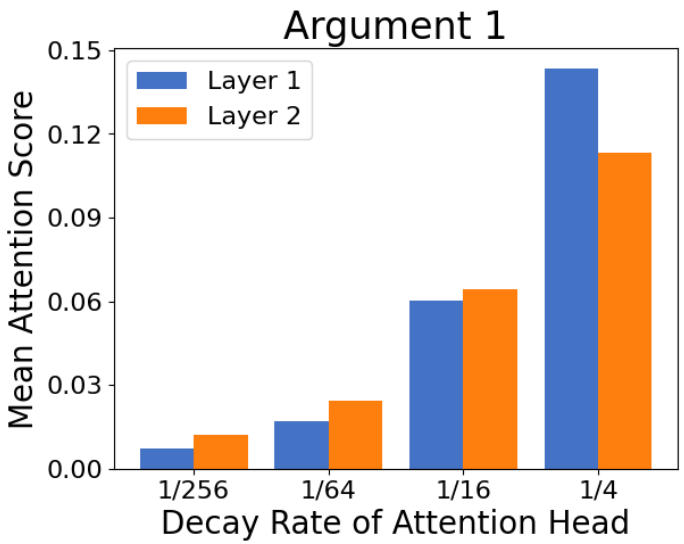
# Experiment 4: Analysis of ALiBi Attention Heads

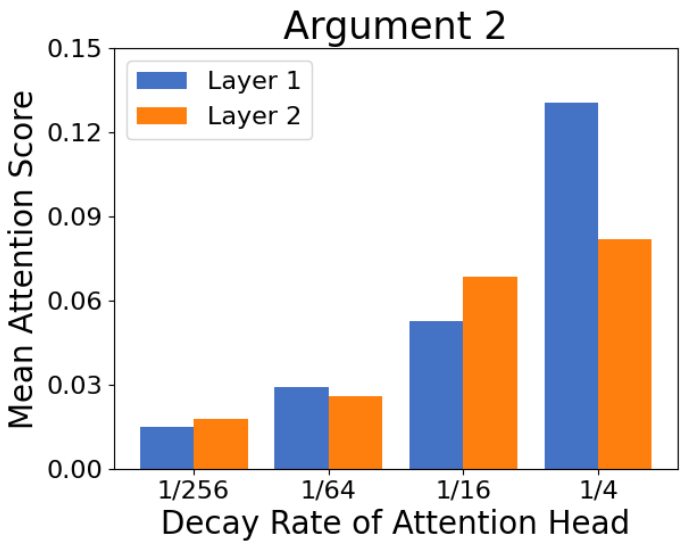
- We hypothesize that mixed decay rates enable tracking different dependencies
- Individual attention heads in a Transformer with ALiBi were tested for three types of semantic dependencies:
  - First arguments (e.g., *The dog bit the man*)
  - Second arguments (e.g., *The man was bitten by the dog*)
  - Coreference (e.g., *The dog chased its tail*)

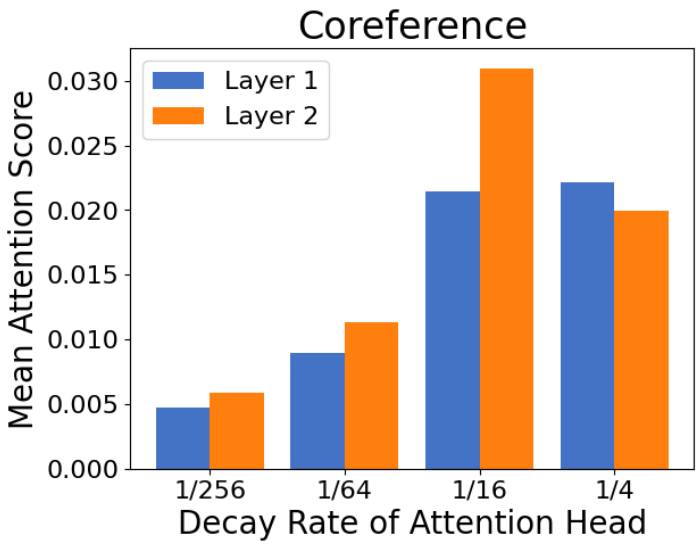
# Experiment 4: Analysis of ALiBi Attention Heads

- Argument and coreference dependencies were identified in Natural Stories (Futrell et al., 2021)
- Dependencies were extracted from existing annotations (Shain et al., 2018)
- Mean attention score was calculated between head and dependent word
- Separately tested each attention head in each layer ($H = 4$, $L = 2$)

# Experiment 4: Results




- Arguments are tracked most by attention heads with faster decay (1/4)
- Coreference is tracked by heads with slower decay (1/16)
- May reflect longer dependency distance for coreference

# Outline

Background

Regression Experiments

Discussion

# Outline

Background

Regression Experiments

Discussion

# Discussion

- Despite the similarity between Transformer attention and cue-based retrieval, Transformers makes other unrealistic assumptions about memory
- Adding a recency bias to address this improves reading-time predictions
- Specifics matter:
  - ALiBi improves predictions; de Varda and Marelli (2024) bias does not
  - Only adding recency bias during inference does not work
  - Mixed decay rates help; uniform decay does not

# Discussion

- What to make of the mixed decay rates in ALiBi?
  - Necessary for improved reading time predictions
  - Contrasts with cognitive models using one decay parameter (e.g., Lewis and Vasishth, 2005)
- Might enable tracking different dependencies, e.g., nonlocal vs. local
- Might reflect different retrieval operations for different dependencies (Yoshida et al., 2025)
- Mixed decay rates might approximate interference during comprehension

- Starting point for broad-coverage models relating memory and expectation

# Conclusion

Thanks for listening!


This material is based upon work supported by the National Science Foundation under Grant Number 1816891. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

# References


Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pages 2397–2430.

Noam Chomsky and George A Miller. 1963. Introduction to the formal analysis of natural languages. In *Handbook of Mathematical Psychology*, pages 269–321. Wiley, New York, NY.

Uschi Cop, Nicolas Dirix, Denis Drieghe, and Wouter Duyck. 2017. Presenting GECO: An eyetracking corpus of monolingual and bilingual sentence reading. *Behavior Research Methods*, 49(2):602–615.

Andrea Gregor de Varda and Marco Marelli. 2024. Locally biased transformers better align with human reading times. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 30–36.

Stefan L. Frank, Irene Fernandez Monsalve, Robin L. Thompson, and Gabriella Vigliocco. 2013. Reading time data for evaluating broad-coverage models of English sentence processing. *Behavior Research Methods*, 45(4):1182–1190.

Richard Futrell, Edward Gibson, and Roger P. Levy. 2020. Lossy-context surprisal: An information-theoretic model of memory effects in sentence processing. *Cognitive Science*, 44(3).

Richard Futrell, Edward Gibson, Harry J. Tily, Idan Blank, Anastasia Vishnevetsky, Steven Piantadosi, and Evelina Fedorenko. 2021. The Natural Stories corpus: A reading-time corpus of English texts containing rare syntactic constructions. *Language Resources and Evaluation*, 55:63–77.

# References


Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. 2020. The Pile: An 800GB dataset of diverse text for language modeling. *arXiv preprint*, arXiv:2101.00027.

Gemini Team. 2024. Gemini 1.5: Unlocking multi-modal understanding across millions of tokens of context. *arXiv preprint*, arXiv:2403.05530.

Edward Gibson. 2000. The Dependency Locality The-ory: A distance-based theory of linguistic complexity. In *Image, language, brain: Papers from the first Mind Articulation Project Symposium*, pages 95–126. MIT Press, Cambridge, MA.

John Hale. 2001. A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics*.

Alan Kennedy, Robin Hill, and Joël Pynte. 2003. The Dundee Corpus. In *Proceedings of the 12th European Conference on Eye Movement*.

Roger Levy. 2008. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.

Richard L. Lewis and Shravan Vasishth. 2005. An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science*, 29(3):375–419.

Steven G. Luke and Kiel Christianson. 2018. The Provo Corpus: A large eye-tracking corpus with predictability norms. *Behavior Research Methods*, 50(2):826–833.

Danny Merkx and Stefan L. Frank. 2021. Human sentence processing: Recurrence or attention? In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 12–22.

George A. Miller and Stephen Isard. 1964. Free recall of self-embedded English sentences. *Information and Control*, 7:292–303.

# References


Byung-Doh Oh and William Schuler. 2022. Entropy- and distance-based predictors from GPT-2 attention patterns predict reading times over and above GPT-2 surprisal. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9324–9334.

Byung-Doh Oh and William Schuler. 2023. Transformer-based language model surprisal predicts human reading times best with about two billion training tokens. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1915–1921.

Ofir Press, Noah Smith, and Mike Lewis. 2022. Train short, test long: Attention with linear biases enables input length extrapolation. In *International Conference on Learning Representations*.

Soo Hyun Ryu and Richard L. Lewis. 2021. Accounting for agreement phenomena in sentence comprehension with Transformer language models: Effects of similarity-based interference on surprisal and attention. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 61–71.

Cory Shain, Marten van Schijndel, and William Schuler. 2018. Deep syntactic annotations for broad-coverage psycholinguistic modeling. In *Workshop on Linguistic and Neuro-Cognitive Resources*.

Nathaniel J. Smith and Roger Levy. 2013. The effect of word predictability on reading time is logarithmic. *Cognition*, 128:302–319.

William Timkey and Tal Linzen. 2023. A language model with limited memory capacity captures interference in human sentence processing. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8705–8720.

Ethan Gotlieb Wilcox, Jon Gauthier, Jennifer Hu, Peng Qian, and Roger P. Levy. 2020. On the predictive power of neural language models for human real-time comprehension behavior. In *Proceedings of the 42nd Annual Meeting of the Cognitive Science Society*, pages 1707–1713.

Ryo Yoshida, Shinnosuke Isono, Kohei Kajikawa, Taiga Someya, Yushi Sugimito, and Yohei Oseki. 2025. If attention serves as a cognitive model of human memory retrieval, what is the plausible memory representation? *arXiv preprint arXiv:2502.11469*.

# Perplexity vs. Psycholinguistic Predictive Power