LING5702: Lecture Notes 18
Transformers and Memory

Contents

18.1 Transformer problem w. veridical context memory [Clark et al., 2025] 1
18.2 Transformer problems with procedural memory [Oh & Schuler, 2023]. 3

18.1 Transformer problem w. veridical context memory [Clark et al., 2025]

Transformers have perfect memory of context, but people don't.

Would Transformers work better if they had memory decay/interference?

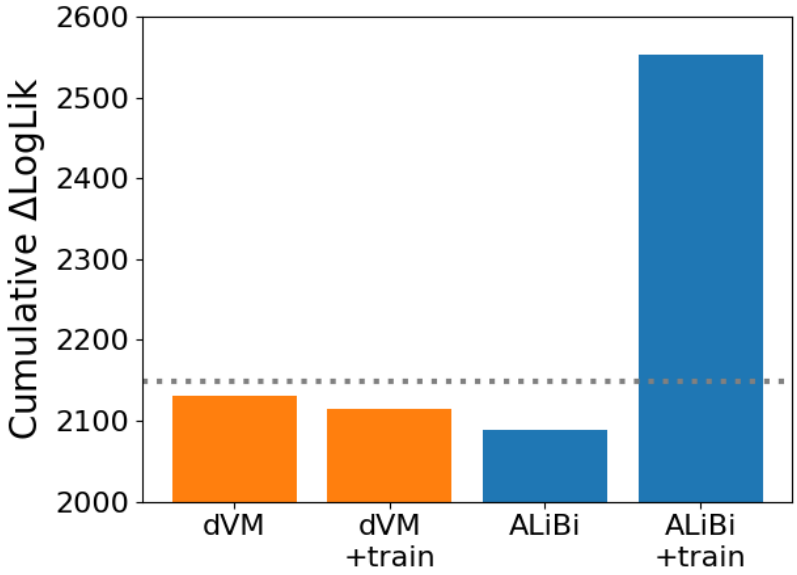
We regressed AliBi [Press et al., 2022], which downweights far attentions:

$$\mathbf{A}'_\ell = m \cdot \begin{array}{|c|c|c|}\hline 0 & & \\ \hline -1 & 0 & \\ \hline -2 & -1 & 0 \\ \hline \end{array} + \frac{1}{\sqrt{d}} \cdot \begin{array}{|c|c|c|}\hline q_1k_1 & & \\ \hline q_2k_1 & q_2k_2 & \\ \hline q_3k_1 & q_3k_2 & q_3k_3 \\ \hline \end{array}$$

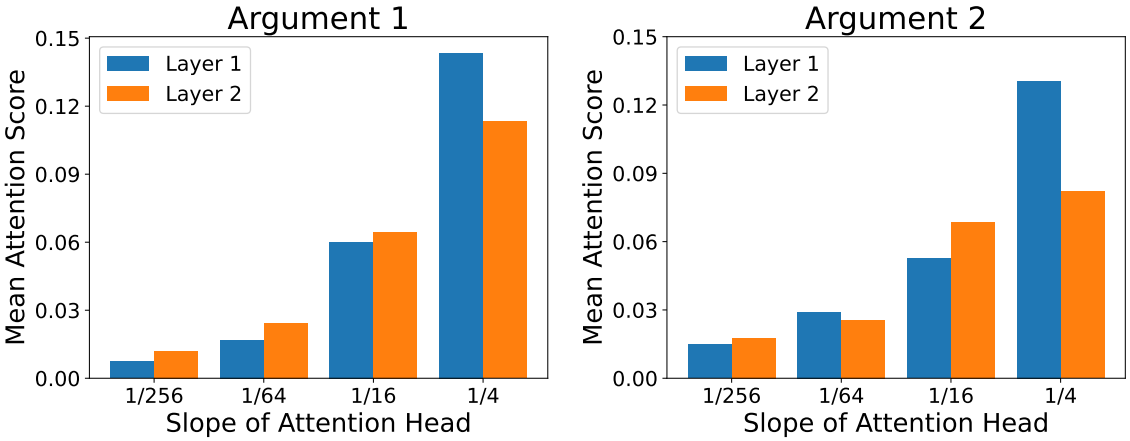
where different heads h get different values of $m = 2^{-8h/H}$.

Evaluated ALiBi recency bias w. 2-layer 4-head 256-dimension Pythia on:

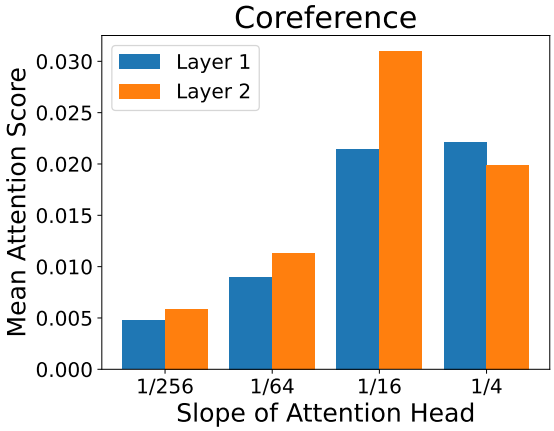
- Brown [Smith & Levy, 2013]: SPR, 35 subj, 7,188 wds.
- Natural Stories [Futrell et al., 2021]: SPR, 181 subj, 10,256 wds.
- UCL [Frank et al., 2013]: SPR, 117 subj, 4,957 wds.
- GECO [Cop et al., 2017]: eye-track (go-past), 14 subj, 56,441 wds.
- Dundee [Kennedy et al., 2003]: eye (go-past), 10 subj, 51,501 wds.
- Provo [Luke & Christianson, 2018]: eye (go-past), 84 subj, 2,746 wds.



The fastest-decaying heads attend to arguments (more interference?):



The second fastest attend to coreference (less interference?):

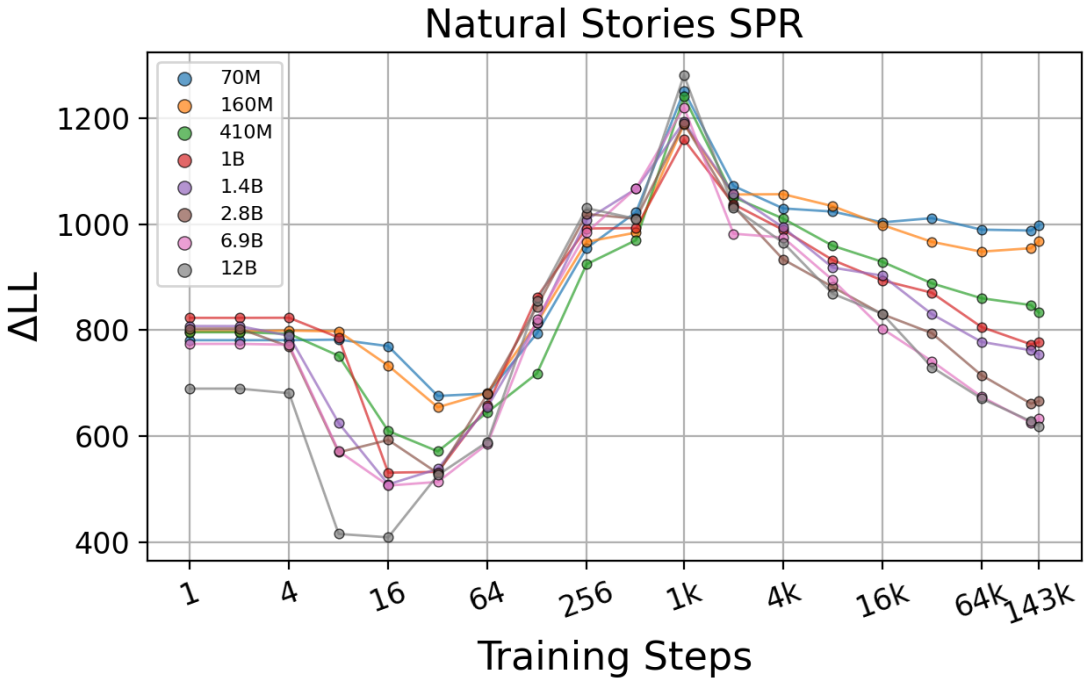


18.2 Transformer problems with procedural memory [Oh & Schuler, 2023]

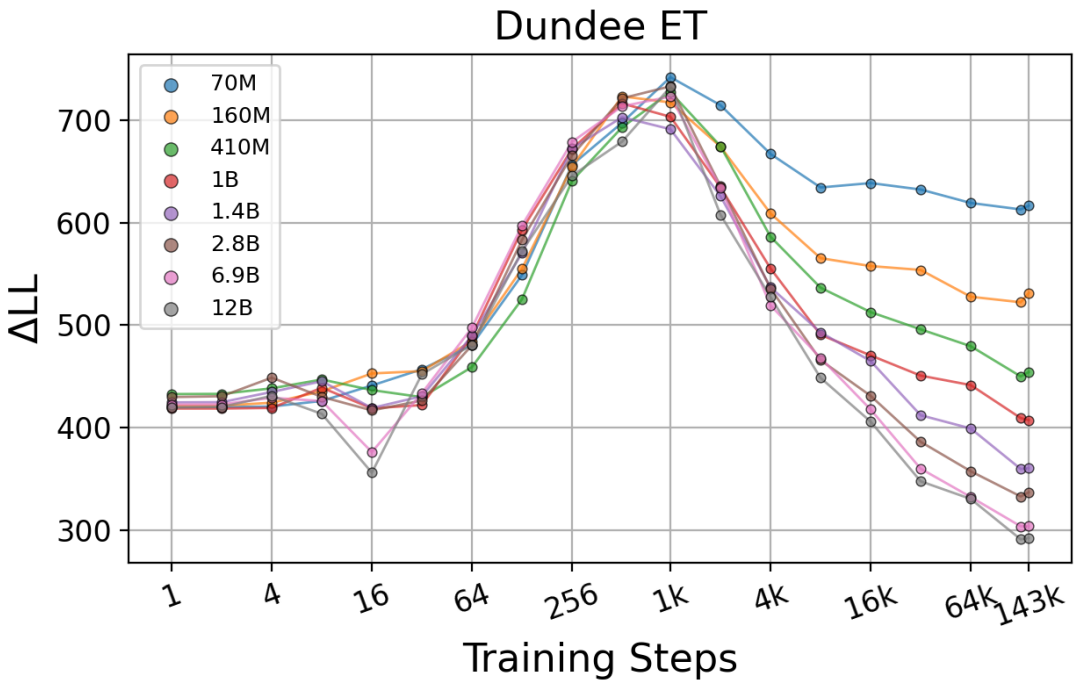
Transformers are trained on billions of words (= hundreds of years).

Would undertrained models be more human-like?

Use undertrained Pythia checkpoints [Biderman et al., 2023]:



LMER, SPR (each step is 2 million tokens).



LMER, eye-tracking go-past (each step is 2 million tokens).

References

[Biderman et al., 2023] Biderman, S., Schoelkopf, H., Anthony, Q. G., Bradley, H., O'Brien, K., Hallahan, E., Khan, M. A., Purohit, S., Prashanth, U. S., Raff, E., Skowron, A., Sutawika, L., & Van Der Wal, O. (2023). Pythia: A suite for analyzing large language models across training and scaling. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, & J. Scarlett (Eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research* (pp. 2397–2430).: PMLR.

[Clark et al., 2025] Clark, C., Oh, B.-D., & Schuler, W. (2025). Linear recency bias during training improves transformers' fit to reading times. In O. Rambow, L. Wanner, M. Apidianaki, H. Al-Khalifa, B. D. Eugenio, & S. Schockaert (Eds.), *Proceedings of the 31st International Conference on Computational Linguistics* (pp. 7735–7747). Abu Dhabi, UAE: Association for Computational Linguistics.

[Cop et al., 2017] Cop, U., Dirix, N., Drieghe, D., & Duyck, W. (2017). Presenting geco: An eyetracking corpus of monolingual and bilingual sentence reading. *Behavior Research Methods*, 49(2), 602–615.

[Frank et al., 2013] Frank, S. L., Fernandez Monsalve, I., Thompson, R. L., & Vigliocco, G. (2013). Reading time data for evaluating broad-coverage models of English sentence processing. *Behavior Research Methods*, 45, 1182–1190.

[Futrell et al., 2021] Futrell, R., Gibson, E., Tily, H. J., Blank, I., Vishnevetsky, A., Piantadosi, S., & Fedorenko, E. (2021). The Natural Stories corpus: A reading-time corpus of English texts containing rare syntactic constructions. *Language Resources and Evaluation*, 55, 63–77.

[Kennedy et al., 2003] Kennedy, A., Pynte, J., & Hill, R. (2003). The Dundee corpus. In *Proceedings of the 12th European conference on eye movement*.

[Luke & Christianson, 2018] Luke, S. & Christianson, K. (2018). The provo corpus: A large eye-tracking corpus with predictability norms. *Behavior research methods*, 50(2), 826–833.

[Oh & Schuler, 2023] Oh, B.-D. & Schuler, W. (2023). Transformer-based language model surprisal predicts human reading times best with about two billion training tokens. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023* (pp. 1915–1921). Singapore: Association for Computational Linguistics.

[Press et al., 2022] Press, O., Smith, N., & Lewis, M. (2022). Train short, test long: Attention with linear biases enables input length extrapolation. In *International Conference on Learning Representations*.

[Smith & Levy, 2013] Smith, N. J. & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, 128, 302–319.