Training Fails to Promote Translation Equivariance in Transformers

Kathleen Kenealy, Michael C. Mozer
{KKENEALY,MCMOZER}@GOOGLE.COM
Google DeepMind

Abstract

We present a result that would seem to have remarkable implications for the design of transformers. We observe that in a trained Gemma3 model, 27% of the variance in individual key-query attention coefficients can be attributed to the absolute position of the query in the context window, roughly the same as an untrained model. Training thus produces no move toward translation equivariance.

The elements of language—words, phrases, and sentences—have intrinsic meaning apart from their position in a text. The phrase "San Diego" can be understood whether it is at the beginning or end of a sentence. "The conference is in San Diego" states a fact that does not depend on whether it is presented in isolation or following "I am looking forward to the conference." Robust understanding by language models requires *translation equivariance*: shifting the position of a text string in the context window should not intrinsically alter how that text is interpreted. Of course, relevant preceding context can shade meaning, but merely moving a sentence in the window from, say, position 47 to position 98 should not alter how it is understood. Without translation equivariance, models would need to learn what "San Diego" means from scratch for every position within the context window and language understanding would be extremely brittle.

Transformers appear to have an inductive bias that favors translation equivariance. The query-key attention mechanism itself is agnostic to the order of preceding tokens. Position codes are incorporated to represent token ordering and distance. Commonly used position codes encode *relative* position—the number of positions between tokens—not *absolute* position—the index of a token within the context window (Ke et al., 2020; Huang et al., 2020; Shaw et al., 2018; Su et al., 2024).

We analyze translation equivariance by examining attention patterns in a transformer. Specifically, we characterize the *persistence* of a token on influencing subsequent processing in terms of the attention allocated to its key from queries that follow. To formalize, we use the notation $a_{t,\Delta t,l,h}$ to denote the attention coefficient associated with a key at position t and a query at position $t + \Delta t$ in layer l and head h of the transformer. If a model exhibits translation equivariance, $a_{t,\Delta t,l,h}$ will not depend on t.

We conducted analyses using a pretrained Gemma3-500m model (Gemma Team et al., 2025) run on C4 eval sequences with a 1024-token context window. Figure 1a plots the persistence of a key as a function of Δt, averaged across test sequences, t, l, and h. The curve is well fit by a power function: the coefficient of determination (R^2) is 0.933. There is of course variability across tokens, layers, and heads: Figure 1b shows typical curves for individual heads of a randomly selected key and layer. The decrease in attention with Δt is still evident even for single heads and single instances.

Although the mean persistence curve is well characterized by a power function, when we condition the analysis on the absolute position of the key (t), Figure 1c indicates that the early positions in the context window are best fit by a power function (purple) and later positions by an exponential (red).

This mysterious pattern has a simple explanation. Suppose that a model distributes its causal attention *uniformly*. The query at position $t+\Delta t$ will attend to itself and the $t+\Delta t$ preceding tokens (assuming zero indexing), leading to an attention coefficient of $1/(t + \Delta t + 1)$ for each key. We refer to the function obtained by conditioning on t and varying Δt as the *uniform causal prior*. When we compare

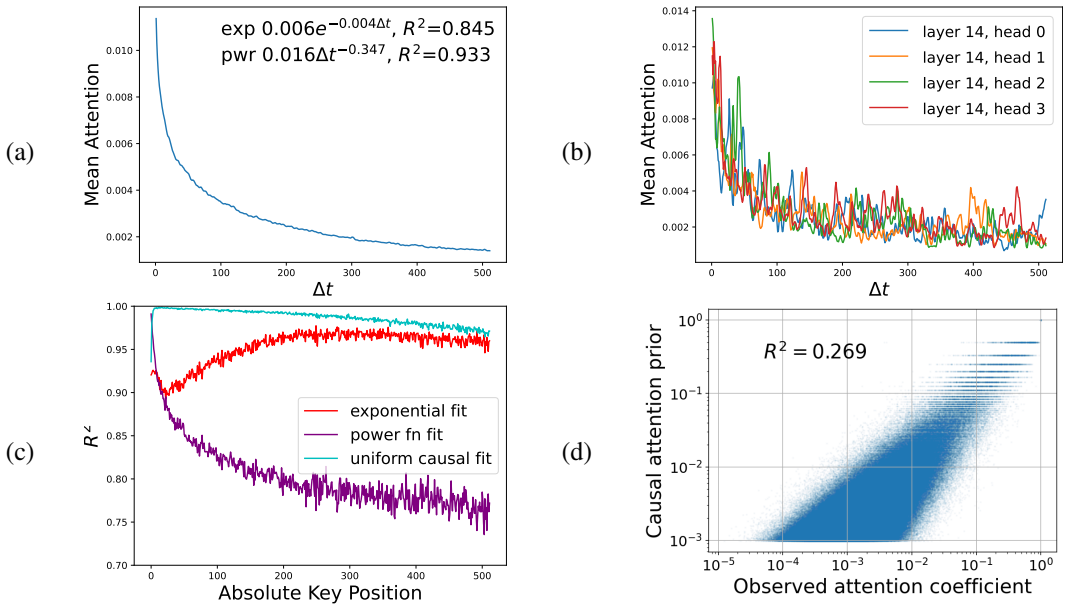

Figure 1: (a) Attention coefficient as a function of key-query distance (Δt), averaged over documents, key position t, layers l, and heads h. (b) Attention coefficient as a function of Δt for a speciific document, layer l, and head h, averaged over key positions t. (c) Quality of fit (coefficient of determination, R^2), as a function of t, between persistence function and exponential, power function, and the uniform causal prior, $(t + \Delta t + 1)^{-1}$. (d) Scatterplot of observed attention coefficients versus uniform causal prior.

this parameter-free curve to our persistence function, we obtain a better fit for every key position t (cyan curve of Figure 1c) than either exponential or power functions, both having two free parameters.

This result indicates that even after training, the transformer still distributes its attention quite widely, at least for aggregated attention across layers, heads, and tokens. Figure 1b suggests that the same pattern is observed for individual instances (i.e., individual layers, heads, and tokens). To obtain direct evidence, we measure the fraction of variance explained in the attention coefficients by the query position ($t + \Delta t$) for individual instances. Figure 1d is a scatterplot of attention coefficients, $a_{t,\Delta t,l,h}$, sampled over all documents, t, Δt, l, and h, versus the causal attention prior, $1/(t+\Delta t+1)$. To capture the dynamic range, we use a log-log plot and obtain a coefficient of determination, R^2, of 0.269 for the log values. (The non-log values produce $R^2 = 0.568$, but because this value is inflated by a few outliers, we adopt the more conservative log-based R^2.) To underscore this result, *if you pick a random token in a random layer and a random attention head, over one quarter of the variance in the attention coefficient is due to the absolute position of the query, independent of token content.*

Contrasting with an *untrained* model with zero weight initialization, we would obtain $R^2 = 1.0$; with actual Gemma initialization, we obtain $R^2 = 0.269$, the same as the trained model. Thus, the training process barely has an effect on narrowing the distribution of attention to a query. One might well have imagined that a model would learn to focus on the local context most of the time, which would make the absolute query position irrelevant—exact translation equivariance—and would yield an R^2 closer to 0.

Our findings have yet to be evaluated on other transformer architectures. We conjecture they will behave similarly because training must start from a roughly uniform attention distribution in order to initially learn to detect relevant context. And with a sufficiently overparameterized model, it appears that solutions are found without narrowing the focus of attention. It remains to be determined whether this property is detrimental to transformer generalization, and if so, what can be done to mitigate the problem. Solutions include local attention windows (currently in Gemma, though the locality spans the 1024 tokens context we studied here) and top-k attention mechanisms, possibly decreasing k over the course of training, and attention normalization schemes (e.g., Miller, 2023).

References

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. Gemma 3 technical report, 2025. URL https://arxiv.org/abs/2503.19786.

Zhiheng Huang, Davis Liang, Peng Xu, and Bing Xiang. Improve transformer models with better relative position embeddings. *arXiv preprint arXiv:2009.13658*, 2020.

Guolin Ke, Di He, and Tie-Yan Liu. Rethinking positional encoding in language pre-training. *arXiv preprint arXiv:2006.15595*, 2020.

Evan Miller. Attention is off by one. https://www.evanmiller.org/attention-is-off-by-one.html, July 2023. Accessed: 2025-08-25.

Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations. *arXiv preprint arXiv:1803.02155*, 2018.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568, 2024.