Not all induction heads are created equal

Tankred Saanum[1,2,3,†] **Eric Schulz**[2]
[1]Harvard University [2]Helmholtz Munich [3]Max Planck Institute for Biological Cybernetics
[†]tankredsaanum@fas.harvard.edu

Abstract

Induction heads are specialized Transformer heads thought to be central for in-context learning. These heads work by having a token x attend to tokens that *succeeded* x in the past, allowing Transformers to predict repetitive structures in the input prompt. When training large-scale Transformer models on text corpora, multiple such heads emerge. In this paper we show that induction heads in a Large Language Model (Qwen2.5-1.5B) exhibit diverse and context-dependent strategies for attending to these successor tokens when there are multiple successor candidates that can be attended to. Some heads prefer to attend to the very *last* successor token, others to the very first. Some heads even "learn" to incorporate second-order contextual cues (e.g. what tokens preceded x) to attend to the successors that are actually predictive of future tokens. Overall, our findings show that induction heads are more sophisticated than previously believed, implement context-dependent computations for predicting future tokens based on patterns observed in the past.

Keywords: transformers; induction heads; attention

Introduction

Induction heads have been argued to account for some of the most fundamental in-context learning abilities of Transformers. An induction head may be defined as a head that, when presented with token x at position t, typically attends to whatever token succeeded a previous instance of token x at a position t' where $t' < t$ (see Fig. 1 for illustration). This simple mechanism can give rise to a range of abilities, including copying patterns in text (Olsson et al., 2022), learning Markovian dynamics (Demircan, Saanum, Jagadish, Binz, & Schulz, 2024; Edelman, Tsilivis, Edelman, Malach, & Goel, 2024), and inducing semantic relationships (Ren et al., 2024). However, attending uniformly to whatever succeeded past occurrences of x may not be sufficient for in-context learning in more sophisticated settings.

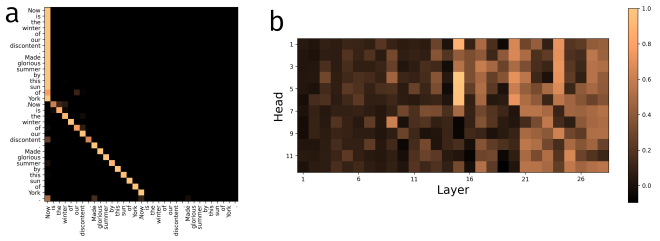

Figure 1: **a**) A characteristic token $\times$ token attention matrix for an induction head, specifically head 4 layer 15. **b**) computing all other attention head's similarity this induction head reveals a cluster of induction heads in layer 15. We analyze these in our study.

For instance, if a token x appears at multiple occasions in the input text, which successor of x should the induction head pay attention to when it needs to predict what follows x in a new context? Maybe all of them? Maybe just the last one? More generally, if the Large Language Model (LLM) needs to predict the successor token of x, but x has had *multiple* different successor tokens in the past, do induction heads use the immediate context preceding x to attend differently to different possible outcomes?

In this paper we show that five induction heads in Qwen2.5-1.5B show a range of diverse and systematic attention profiles when presenting the LLM with different repetitive patterns. Specifically, we find that induction heads are generally context sensitive, and use the context preceding x to attend to the correct successor tokens, ignore the "wrong" ones. While induction heads are undeniably implicated in in-context learning, our results suggest that there is more to the story: The representations that induction heads convert into attention maps must already contain information that allows them to distinguish informative from uninformative contexts.

Diversity

When there are multiple occurances of a token x in the past, do induction heads distribute their attention differently? We presented Qwen2.5-1.5B (Yang et al., 2024) with a phrase from William Shakespeare's *Richard III*, and repeated it seven times. Inspecting the token $\times$ token attention masks of the induction heads in layer 15, we indeed see diverse strategies for attending to the different repetitions of the phrase (see Fig. 2**a**). Head 1 showed a preference towards attending to the *previous* repetition, Head 4 distributed attention roughly equally among all repetitions, and head 6 preferred the *first* instance of the phrase. Head 7 showed a similar preference to head 4, but instead of only attending to the successor token of x, it attended to the *surrounding* tokens of past instances of x.

Response to disruptions

Next we prompted the LLM with a variation of the repeated Shakespeare phrase. Here we replaced the 3rd repetition with a corrupted version in which the positions of the tokens had been scrambled. This meant that the tokens in the 3rd phrase were not predictive of future tokens. Strikingly, some of the induction heads (Head 4 and 7) learned to ignore the tokens in this repetition, showing adaptive and context dependent characteristics (see Fig. 2**b**).

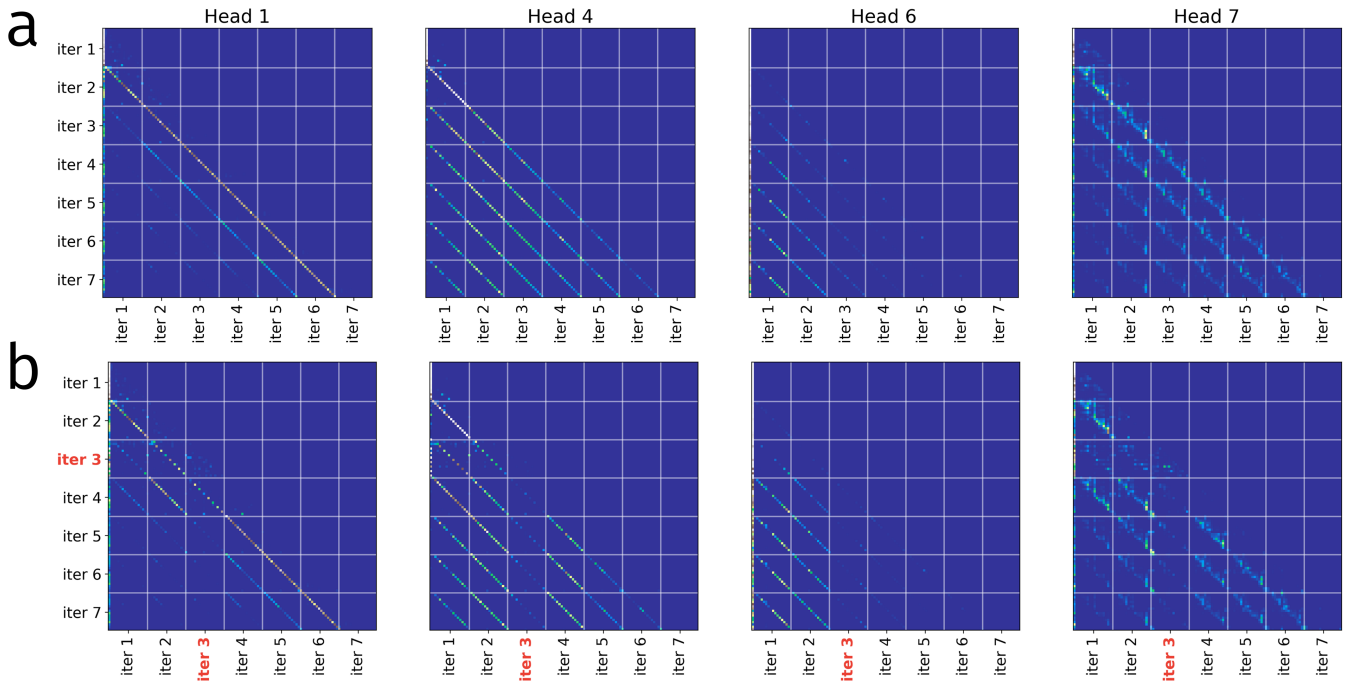

Figure 2: **a**) In response to a simple phrase repeated seven times, the four induction heads highlighted show diverse attention profiles. Head 1 attends mostly to the last repetition, head 4 to all repetitions, and head 6 to the first repetition. Head 7 pays attention to the surrounding tokens. **b**) When *corrupting* the third repetition, head 4 and 7 almost cease to attend to this repetition.

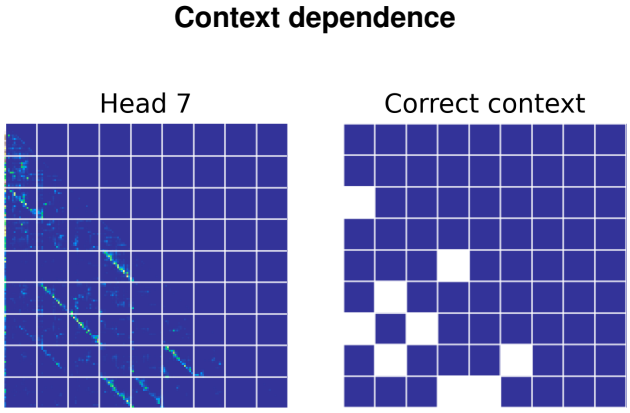

Figure 3: In the first three repetitions, head 7 distributes attention across all previous repetitions, but eventually starts to attend to successor tokens from the "correct" context, suggesting learning.

Does this context-dependence generalize to more complex settings? We created three unique scrambled versions of the original phrase, (A, B, C), and composed a sequence where each scrambled phrase appeared three times in a random order, for instance $A, B, A, C, B, C, A, B, C$. We refer to these three scrambled phrases as latent contexts. Since the contexts consisted of the same tokens in different order, induction heads that are not sensitive to the latent context (e.g. whether the current tokens come from A, B or C), should distribute attention to successor tokens uniformly across all contexts. On the other hand, context sensitive induction heads should only attend to the corresponding successor tokens from identical context. Simply put, tokens from the A context should only attend to successor tokens from past A contexts, and so on.

Here too, most of the induction heads in our analysis prefer to attend to successor tokens from the matching contexts. Illustrating with Head 7, we see that the first three repetitions attend uniformly across contexts. However, after the third repetition, the attention profile closely matches the optimal attention pattern, where tokens only attend to their successor tokens from the matching contexts (see Fig. 3). This suggests that an adaptive learning mechanism, perhaps involving a prediction error signal, is at play.

Discussion

Discovering and predicting repetitive patterns is a hallmark of intelligence (Saanum, Éltető, Dayan, Binz, & Schulz, 2023; Sayood, 2017; Kumar et al., 2022). Induction heads have been proposed as a computational mechanism underpinning this ability in LLMs. Our results show that there is not only a substantive diversity in the attention strategies that induction heads exhibit when predicting repetitive patterns, but also that these strategies are adaptive, and change systematically depending on the context, suggesting that the computations performed by induction heads are more sophisticated than previously believed.

References

Demircan, C., Saanum, T., Jagadish, A. K., Binz, M., & Schulz, E. (2024). Sparse autoencoders reveal temporal difference learning in large language models. *arXiv preprint arXiv:2410.01280*.

Edelman, E., Tsilivis, N., Edelman, B., Malach, E., & Goel, S. (2024). The evolution of statistical induction heads: In-context learning markov chains. *Advances in Neural Information Processing Systems*, *37*, 64273–64311.

Kumar, S., Correa, C. G., Dasgupta, I., Marjieh, R., Hu, M. Y., Hawkins, R., ... others (2022). Using natural language and program abstractions to instill human inductive biases in machines. *Advances in Neural Information Processing Systems*, *35*, 167–180.

Olsson, C., Elhage, N., Nanda, N., Joseph, N., DasSarma, N., Henighan, T., ... others (2022). In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*.

Ren, J., Guo, Q., Yan, H., Liu, D., Zhang, Q., Qiu, X., & Lin, D. (2024). Identifying semantic induction heads to understand in-context learning. *arXiv preprint arXiv:2402.13055*.

Saanum, T., Éltető, N., Dayan, P., Binz, M., & Schulz, E. (2023). Reinforcement learning with simple sequence priors. *Advances in Neural Information Processing Systems*, *36*, 61985–62005.

Sayood, K. (2017). *Introduction to data compression*. Morgan Kaufmann.

Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., ... others (2024). Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.