Inducing Induction in Llama via Linear Probe Interventions

Sheridan Feucht Byron C. Wallace David Bau
Northeastern University
{feucht.s, b.wallace, d.bau}@northeastern.edu

Abstract

For induction heads to copy forward information successfully, heads in earlier layers must first load previous token information into every hidden state, a process Olsson et al. (2022) call *key shifting*. While this information is hypothesized to exist, there have been few attempts to explicitly locate it in models. In this work, we use linear probes to identify the subspaces responsible for storing previous token information in Llama-2-7b and Llama-3-8b. We show that these subspaces are causally implicated in induction by using them to "edit" previous token information and trigger random token copying in new contexts.

1 Introduction

What information encoded in a token allows it to be copied via induction? Although the workings of induction heads have been carefully studied (Elhage et al., 2021; Olsson et al., 2022), there has yet to be work explicitly isolating the presumed previous token information that enables induction heads to function as they do. Using linear probes from Feucht et al. (2024), we identify low-dimensional subspaces in Llama-2-7b (Touvron et al., 2023) and Llama-3-8b (Meta, 2024) that contain previous token information used for random token copying. To show this, we take information identified by linear probes for one input and substitute it into a completely different context (Figure 1), which artificially triggers copying in certain layers.

2 Related Work

Elhage et al. (2021) and Olsson et al. (2022) describe induction circuits as consisting of two stages: (1) *key shifting,* where one set of attention heads copies previous token information into succeeding token positions, and (2) *prefix matching,* where another set of heads later attends to that information to copy previously-seen sequences. While they characterize these heads through direct examination of weights and attention patterns, we show causal evidence of this process by editing previous token information in Llama models.

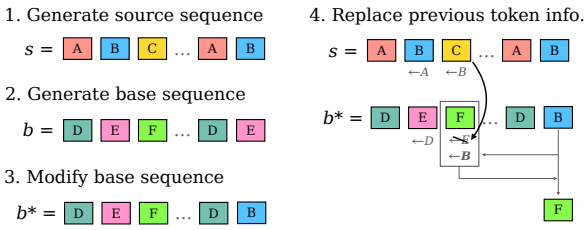

Figure 1: We artificially induce copying behavior in Llama, using probes from Feucht et al. (2024) to modify previous token information.

3 Method

3.1 Random Copying Task

First, we define a simple copying task wherein a sequence of l uniformly sampled tokens is duplicated and separated by a newline. We measure Llama's ability to predict the last token of such a sequence, which it can only achieve via in-context copying (e.g. `land.id.Pale \n land.id` $\rightarrow$ `Pale`). For $l = 30$ with random tokens, Llama-2-7b has a completion accuracy of 87.9% for this task, whereas Llama-3-8b reaches 100% accuracy.

3.2 Probe Intervention

Next, we show that previous token information measured by probes from Feucht et al. (2024) is causally implicated in this random token copying task. Figure 1 shows the setup for this experiment. We first randomly generate source and base sequences s and b, following the same newline-separated template as in Section 3.1. We then replace the last token of b with the last token of s to create b^*, which prevents models from completing the copying task: Llama-2-7b copying accuracy drops to 0% with this change, whereas Llama-3-8b accuracy drops to 13%. Finally, we show that we

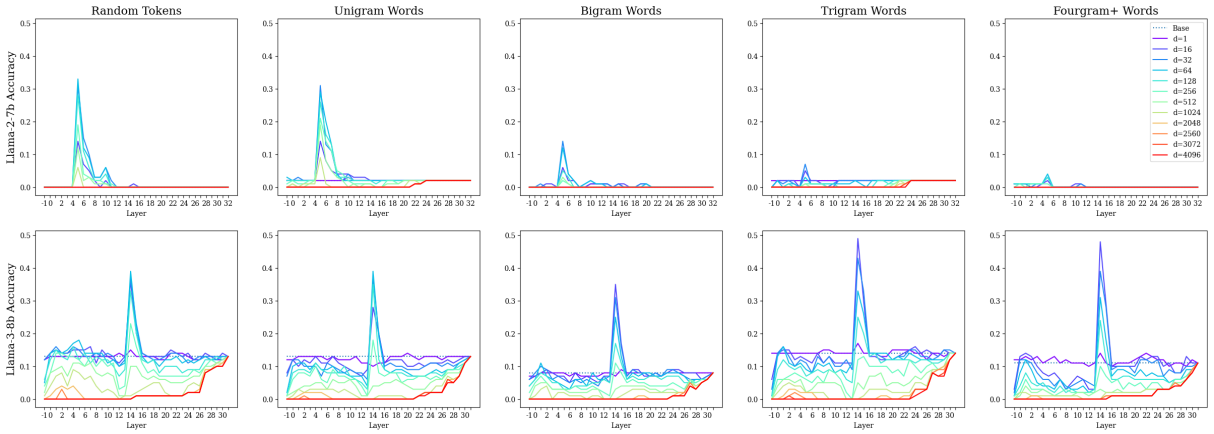

Figure 2: Results for experiment described in Figure 1, which show previous token information implicated in token copying at either layer 5 (Llama-2-7b) or layer 14 (Llama-3-8b). The leftmost column indicates results for sequences consisting of fully-random tokens, whereas the four right-hand columns show results for source sequences ending in multi-token words from Wikipedia. While unigrams act similarly to random tokens as expected, we see a divergence for multi-token words, where interventions being to fail for Llama-2-7b but become shaper for Llama-3-8b. Accuracy is calculated across one hundred examples for each cell.

can restore copying behavior for b^* by modifying the hidden state of the token we want to copy (i.e. `F` in Figure 1). We use linear probes to extract previous token information [← `B`] from the hidden representation for `C` at layer ℓ (Figure 1), and then substitute this "flag" into the hidden state for `F` at the same layer, measuring the resulting accuracy for prediction of `F` over 100 sequences.

To do this, we calculate the SVD of the probe, $P^\ell = U\Sigma V^*$, and then take the top d rows of V^*, which form an orthonormal basis for a subspace A of dimension d. To substitute previous token information from a source hidden state s_t^ℓ into a hidden state from the modified base sequence $b_t^{*\ell}$, we take the projection of s_t^ℓ onto A and add it to the projection of $b_t^{*\ell}$ onto the orthogonal complement of A. Thus, the formula for an intervened hidden state at token position t and layer ℓ is simply $h_t^\ell = AA^T s_t^\ell + (I - AA^T) b_t^{*\ell}$. We perform this intervention for every layer ℓ.

3.3 Entity Copying

In addition to testing on random sequences of tokens, we also test sequences that end in multi-token words (MTWs) taken from Wikipedia (Foundation, 2022). We then modify s so that it ends with a Wikipedia MTW (e.g. `land.id.task.ed \n land.id.task` → `ed`). This means that the previous token information inserted into b^* now comes from the final token of a MTW. Motivated by results from Feucht et al. (2024) showing "erasure" of that information in early layers, we ask whether this will affect induction behavior for those sequences.

4 Results

The leftmost column of Figure 2 shows that this intervention causes a 34% increase in accuracy above baseline for Llama-2-7b and a 26% increase for Llama-3-8b on random sequences of tokens, suggesting that models attend to the artificially-inserted [← `B`] information to promote the output of `F` as hypothesized. Results for MTW copying are more mixed. For Llama-2-7b, interventions cease to be effective with MTWs. However, Llama-3-8b interventions are arguably stronger.

5 Discussion

Results in Figure 2 indicate that there is some mechanism after layer 5 in Llama-2-7b (layer 14 in Llama-3-8b) that refers to previous token information in order to copy tokens from its context. We interpret this result as direct evidence of hypothesized "key shifting" being used for copying in full-scale models. Although this effect is strong for random tokens, MTWs are treated differently. Since prior work shows that previous token information is "erased" for MTWs in both models, we expected interventions to become less effective for MTWs; however, this is only true for Llama-2-7b. For Llama-3-8b, interventions instead become *more* effective, with fewer dimensions required to achieve high accuracy ($d = 16$ instead of $d = 32$). This difference in MTW induction mechanisms may indicate a fundamental difference between how these two models represent multi-token words.

References

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. 2021. A mathematical framework for transformer circuits. *Transformer Circuits Thread*. Https://transformer-circuits.pub/2021/framework/index.html.

Sheridan Feucht, David Atkinson, Byron Wallace, and David Bau. 2024. Token erasure as a footprint of implicit vocabulary items in llms. *Preprint*, arXiv:2406.20086.

Wikimedia Foundation. 2022. Wikimedia downloads.

Meta. 2024. Introducing meta llama 3: The most capable openly available llm to date.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. 2022. In-context learning and induction heads. *Transformer Circuits Thread*. Https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.