Does Editing Provide Evidence for Localization?

Zihao Wang [1] **Victor Veitch** [1 2]

Abstract

A basic aspiration for interpretability research in large language models is to "localize" semantically meaningful behaviors to particular components within the LLM. There are various heuristics for finding candidate locations within the LLM. Once a candidate localization is found, it can be assessed by editing the internal representations at the corresponding localization and checking whether this induces model behavior that is consistent with the semantic interpretion of the localization. The question we address here is: how strong is the evidence provided by such edits? To assess localization, we want to assess the effect of the optimal intervention at a particular location. The key new technical tool is a way of adapting LLM alignment techniques to find such optimal localized edits. With this tool in hand, we give an example where the edit-based evidence for localization appears strong, but where localization clearly fails. Indeed, we find that optimal edits at *random* localizations can be as effective as aligning the full model. In aggregate, our results suggest that merely observing that localized edits induce targeted changes in behavior provides little to no evidence that these locations actually encode the target behavior.

1. Introduction

A basic goal of interpretability research for large language models is to map semantically meaningful behavior to particular subcomponents of the model. Semantically meaningful encompasses a wide range of things, e.g., "when asked for directions to the Eiffle tower, the model gives directions to Paris", "the model responds truthfully", or "the model will refuse to response". The aim is to find, e.g., neurons, circuits, or regions of representation space that control these behaviors. If we could find such localizations, we could use them as building blocks to understand complex model behaviors. Many interpretability approaches can be understood in terms of the following idealized template (e.g., Zou et al., 2023; Arditi et al., 2024; Wang & Shu, 2023; Chen et al., 2024; Wei et al., 2024; Li et al., 2024; Meng et al., 2022; Vig et al., 2020; Geiger et al., 2021; Soulos et al., 2019; Finlayson et al., 2021; Wang et al., 2022; Chan et al., 2022; Hanna et al., 2024; Conmy et al., 2023; Todd et al., 2023; Hendel et al., 2023):

1. We use some heuristic to find a candidate location in the model that is conjectured to be responsible for a particular behavior.
2. We then run the model with some set of inputs, and collect the model's internal representations for each input.
3. Then, we edit each of these representations at the candidate location, and generate new outputs according to the edited representations.
4. If the edit changes the model's behavior in the manner that would be expected from changing the target behavior, we take this as evidence in support of localization.

For example, if editing a particular location in the network shifts the model to give truthful answers, we may take this as evidence that the location meaningfully encodes truthfulness in some sense. Or, if editing a location causes the model to act as though the Eiffel tower is in Rome, we may take this as evidence that the location encodes the concept of the Eiffel tower. The basic question in this paper is: how strong is this evidence? That is, to what extent can we conclude that a particular location in the model is responsible for a particular behavior based on the success of editing at that location?

Our core contribution is an example where editing-based evidence appears very strong, but where localization clearly fails. The example replicates the setup of Inference-Time-Interference (ITI) (Li et al., 2024), where the target concept is truthfulness, and the localization is in a small subset of 16 attention heads. Following ITI, we use logit-linear probing to identify candidate heads. We then search for the optimal

[1]Department of Statistics, University of Chicago, Chicago, IL, USA [2]Data Science Institute, University of Chicago, Chicago, IL, USA. Correspondence to: Zihao Wang <wangzh@uchicago.edu>.

ICML 2024 Workshop on Mechanistic Interpretability, Vienna, Austria. PMLR 235, 2024.

localized edit to apply at these heads. Remarkably, we find that the optimal edit induces truthfulness behavior that is essentially as good as finetuning the entire model to be truthful. That is, the localized edit is as effective as can possibly be expected. Intuitvely, this appears to be strong evidence that the locations found by the heuristic (probing) are indeed closely linked to the target concept (truthfulness). However, we then show that this evidence is misleading. We find that applying optimal edits to *random heads* are just as effective as when applied to the localized heads. Accordingly, the edit-based evidence provides no support for the localization hypothesis.

A possible out here is that 16 attention heads is too many, leaving us with significant leeway to induce any behavior we want with editing. We further strengthen the example by showing that it is possible to find a *single* head in the model where editing at that head is as effective as finetuning the entire model. This appears to be the strongest edit-based evidence for localization possible. However, we show that there are in fact multiple such heads. That is, there is simply no single privileged location that can be identified as responsible for the target behavior.

Our results suggest that the evidence provided by editing is weak, and that the success of editing at a particular location is not a reliable indicator of the location's importance for the target behavior. This seems to significantly constrain what can be learned from interpretability methods. It also points to the need for a more rigorous development of such techniques, including both precise statements of what the goals are, and well-grounded standards for evidence that these goals have been met.

The technical development in this paper relies on finding the optimal intervention at a specified location. To that end, we develop a method for localizing LoRA type finetuning to specific locations. This then allows us to frame the search for optimal edits as a finetuning-type optimization problem. This method may also be of independent interest.

2. Background and results from ITI

We replicate the setup of ITI (Li et al., 2024).

Dataset and Model Architecture We use TruthfulQA (Lin et al., 2021) as our dataset. It contains 817 questions that humans might answer incorrectly due to misconceptions. Each question contains an average of 3.2 truthful answers and 4.1 false answers. We use 60% of the questions for training, and the rest for validation and testing.

We use an Alpaca-7B (Taori et al., 2023) model that is finetuned from the Llama-7B base model. The model consists of $L = 32$ layers, each consisting of a Multi-head Attention (MHA) layer, and a Multilayer Perceptron (MLP) layer. We focus on the MHA layer, which has $H = 32$ attention heads, with each head having dimension $H = 128$ (the hidden dimension is $DH = 4096$).

Ignoring MLP and layer normalization, the computation at layer l can be written as:

$$\boldsymbol{o}_h^l := \text{Attn}_h^l(\boldsymbol{r}^l) \in \mathbb{R}^D \tag{2.1}$$

$$\boldsymbol{o}^l := [(\boldsymbol{o}_1^l)^T, \ldots, (\boldsymbol{o}_H^l)^T]^T \in \mathbb{R}^{DH} \tag{2.2}$$

$$W^l := [W_1^l, \ldots, W_H^l] \in \mathbb{R}^{DH \times DH} \tag{2.3}$$

$$\boldsymbol{r}^{l+1} := \boldsymbol{r}^l + W^l \boldsymbol{o} = \boldsymbol{r}^l + \sum_{h=1}^{H} W_h \boldsymbol{o}_h \in \mathbb{R}^{DH} \tag{2.4}$$

where $r^l \in \mathbb{R}^{DH}$ is the residual stream before layer l, Attn_h^l is the h-th attention module at layer l, with $\boldsymbol{o}_h^l$ being its output. $\boldsymbol{o}^l$ is the concatenated head outputs. W^l is the project-out matrix, that applies H independent linear transformations to the corresponding head outputs. Finally $\boldsymbol{r}^{l+1}$ is residual stream output after layer l.

Localization and intervention using activation statistics To localize, we collect representations for positive and negative examples, and use probing to find where the truthfulness concept is represented. To intervene, we find the direction best separating activations for positive and negative examples, and apply this direction to the representation.

Each example is of the form, $(x, y, x_{\text{random}})$, concatenating a question x, a corresponding answer y, and another random question x_{random}. For positive examples, we use a truthful response $y = y_+$, and for negative examples, we use an untruthful response $y = y_-$. To collect the representations, we feed the positive and negative examples through the model, and collect the activations of the attention heads, $\{\boldsymbol{o}_h^l\}_{h \in [H], l \in [L]}$, *at the last token*.

For each of the $L \times H$ head locations, we train a logistic regression probe on the D-dimensioanl activations to predict whether it's a positive or negative example. Then we pick the attention heads with the highest probing accuracies as the localized heads.

For the selected head at (l, h), we find the direction u_h^l that is "best" at separating the activations of positive and negative examples. There are several variants, but according to (Li et al., 2024), the best option is the mass mean shift, which is the difference between the average positive and negative activations. Then we estimate the standard deviation of activations along the direction to be σ_h^l, and use the weighted direction $\theta_h^h := \sigma_h^l u_h^l$ as the intervention vector, which we add to the corresponding head during inference autoregressively.

More specifically, the applied intervention is:

$$\boldsymbol{r}_{\text{ITI}}^{l+1} := \boldsymbol{r}^{l} + W^{l}(\boldsymbol{o} + \alpha\boldsymbol{\theta}^{l}) \tag{2.5}$$

$$= \boldsymbol{r}_{\text{orig}}^{l+1} + \alpha W^{l}\boldsymbol{\theta}^{l} = \boldsymbol{r}_{\text{orig}}^{l+1} + \alpha \sum_{h=1}^{H} W_{h}^{l}\boldsymbol{\theta}_{h}^{l} \tag{2.6}$$

where $\boldsymbol{\theta}_l$ is the concatenated intervention vectors across all heads at layer l, and α is the intervention strength. This intervention is repeated for each next token prediction autoregressively until the whole answer is completed.

Evaluation Metrics Since the goal is to assess model's generation quality, it's natural to use truthfulness score and informativeness score of generations as the evaluation metrics. They use GPT-judge models (Lin et al., 2021) to evaluate the model's generations for truthfulness and informativeness, and use Info*Truth (the product of scalar truthful and informative scores) as the main metric.

We also report other metrics as in the ITI paper: KL divergence of the model's next-token prediction distribution post- versus pre-intervention, and multiple-choice accuracy (MC) which is determined via comparing the conditional probabilities of candidate answers given the question.

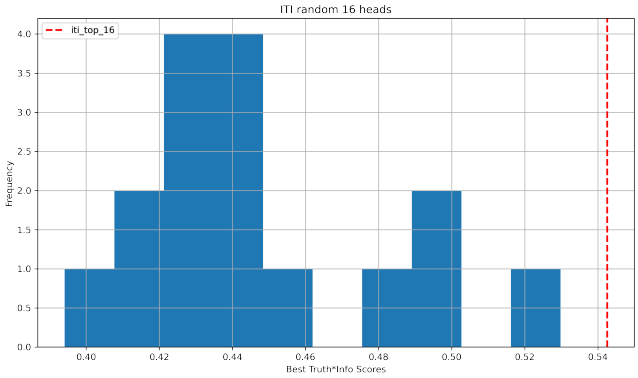
3. Editing Localized Heads Modifies the Output as Expected

In ITI, the authors find that editing on 16 localized heads (out of a total of 1024 heads) successfully steers model generations to be more truthful while still being informative. They also find intervening on all attention heads doesn't make model generations more truthful than intervening just at the localized heads. This seems to suggest that the truthfulness concept is indeed encoded in the localized heads.

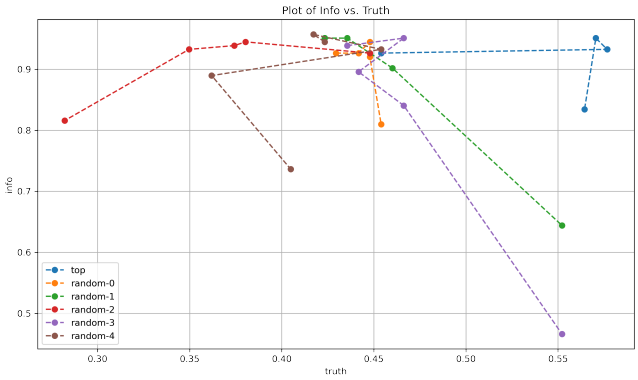
We now strengthen this evidence further. Similar to Hase et al. (2024), we check if interventions at random heads can also make model generations more truthful. More specifically,

1. Randomly select 16 heads, and compute intervention vectors $\boldsymbol{\theta}^{l}$'s accordingly.
2. Apply varying intervention strength α, collect model generations, and compute scores for truthfulness and informativeness using GPT-judge across all intervention strengths.
3. Repeat for 16 times.

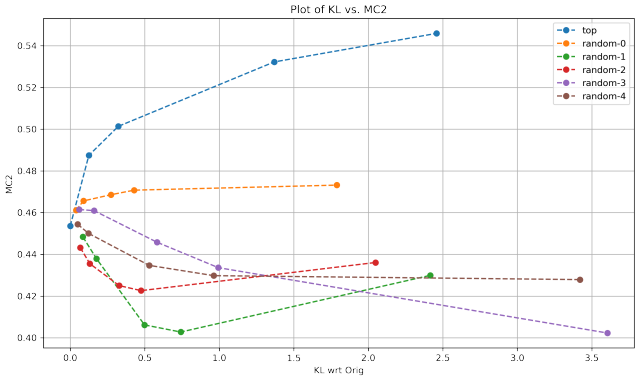
We find that interventions at the localized heads are more effective than interventions at random heads. In fig. 1a we report the Info*Truth score (average truthfulness score times average informativeness score). We find that using localized heads have significantly higher Truth*Info scores than using random heads (p-value 1.6×10^{-8}). In fact, using random heads often doesn't have noticeable effect on the truthfulness at all, as shown in fig. 1b, fig. 1c .



(a) Info*Truth Scores



(b) Truth vs Info Scores



(c) KL vs MC Scores

Figure 1. Localized heads perform much better than random when using ITI interventions. We observe better Truth*Info scores, better truth-info score tradeoff, as well as better MC-KL tradeoff.

This appears to add further evidence that the localized heads are "special" for the truthfulness concept. However, this strong association could be because the intervention and localization are "correlated", since both use statistics of the same activations (determined by the design of the data, etc). E.g. for heads with very low probing accuracy, the

estimated intervention vectors could be very noisy, and thus the interventions could be less effective.

4. Finding "optimal" interventions

To test whether a particular behavior is localized to specific location, we would like to assess the effect of the *optimal* intervention at that location. In the case of our running example, we want the localized edit to the representation space that does the best job of steering the model's generations to be more truthful while maintaining informativeness. Then, the questions are: what is the best we could hope to achieve? (I.e., what is "optimal"?) And, (how) can we find a localized edit that achieves it?

Fitting the alignment objective gives optimal interventions

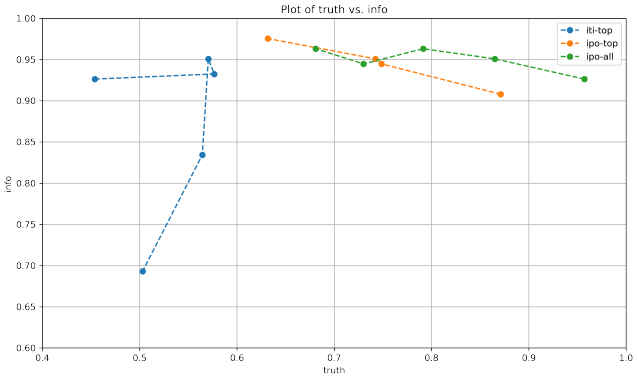


Figure 2. IPO interventions achieve much better performance than using ITI. Using IPO interventions at localized heads give nearly optimal info-truth tradeoff as well.

The key observation is that the dataset used to construct positive and negative examples can be restructured as paired "preference" data $\{(x_i, y_i^+, y_i^-)\}_i$, where x_i is the question, y_i^+ is the truthful answer, and y_i^- is the untruthful answer. Since the goal is to make model generations more truthful, we can directly adopt contrastive alignment methods for biasing the model towards the truthful answers. In this case, we use the IPO (Azar et al., 2024) learning objective, where the goal is to upweight probabilities for y_i^+ and downweight probabilities for y_i^- (up to some threshold):

$$\text{argmax}_\phi \sum_i \left[\log \left(\frac{\pi_\phi(y_i^+|x_i)}{\pi_0(y_i^+|x_i)} \Big/ \frac{\pi_\phi(y_i^-|x_i)}{\pi_0(y_i^-|x_i)} \right) - \frac{\tau^{-1}}{2} \right]^2$$

where $\pi_\phi(.|x)$ is the model's generation probability, $\pi_0(.|x)$ is the original model's generation probability, and τ decides the threshold. Ideally, the optimized $\pi_{\phi^*}(.|x)$ should generate responses that are more truthful than the original model, while minimally affecting the off-target aspects of the generation (in this case, the informativeness of the responses).

To test the effectiveness of IPO alignment, we finetune the weights for project-out matrices W^l's defined in eq. (2.3) using (rank 1) LoRA (Hu et al., 2021). The finetuned model gives nearly perfect trade-off between truthfulness and informativeness, that is far better than ITI interventions fig. 2. This also suggests that ITI heuristics are very far from optimal, and contrasts with ITI results that intervening on all heads doesn't make model generations more truthful.

Now we treat this result as the overall best performance that we can achieve with interventions. We want to see if optimal interventions at localized heads can achieve the same performance, and if random heads can achieve the same performance.

Connect weight updates to representation editing

The connection to IPO lets us search for the best possible update to the model's *weights*. However, we are interested in localized edits to model *representations*. To continue, we need to connect the weight editing to representation editing.

Rank-1 LoRA Directly applying rank-1 LoRA to W^l, we can view the effect of adding in the modified LoRA weight matrix as an edit to the representation as follows:

$$\boldsymbol{r}^{l+1}_{\text{LoRA}} := \boldsymbol{r}^l + (W^l + \boldsymbol{b}^l(\boldsymbol{a}^l)^T)\boldsymbol{o}^l = \boldsymbol{r}^{l+1}_{\text{orig}} + \langle \boldsymbol{a}^l, \boldsymbol{o}^l \rangle \boldsymbol{b}^l, \tag{4.1}$$

where $\boldsymbol{a}^l, \boldsymbol{b}^l$ are the LoRA weights to optimize. Comparing with eq. (2.5), we see that $\boldsymbol{b}^l$ plays the role of the added $W^l\boldsymbol{\theta}^l$, and $\langle \boldsymbol{a}^l, \boldsymbol{o}^l \rangle$ is the intervention strength but is adapted to the representation $\boldsymbol{o}$.[1]

This formulation connects weight edits to representation edits. However, it doesn't yet allow us to localize edits to specific heads — while $\boldsymbol{\theta}^l$ can be read as concatenation of headwise intervention vectors, the projected $W^l\boldsymbol{\theta}^l$ have no corresponding interpretions. Therefore, we can't restrict the edits to specific heads by imposing structure on $\boldsymbol{b}^l$'s.

Rank-1 LoRA with reparameterization We can make more direct connections by reparameterizing $\boldsymbol{b}^l$ with $W^l\boldsymbol{b}^l$ (without changing expressiveness):

$$\boldsymbol{r}^{l+1}_{\text{LoRA-reparam}} := \boldsymbol{r}^{l+1}_{\text{orig}} + \langle \boldsymbol{a}^l, \boldsymbol{o}^l \rangle W^l\boldsymbol{b}^l \tag{4.2}$$

$$= \boldsymbol{r}^{l+1}_{\text{orig}} + \langle \boldsymbol{a}^l, \boldsymbol{o}^l \rangle \sum_{h=1}^{H} W^l_h \boldsymbol{b}^l_h \tag{4.3}$$

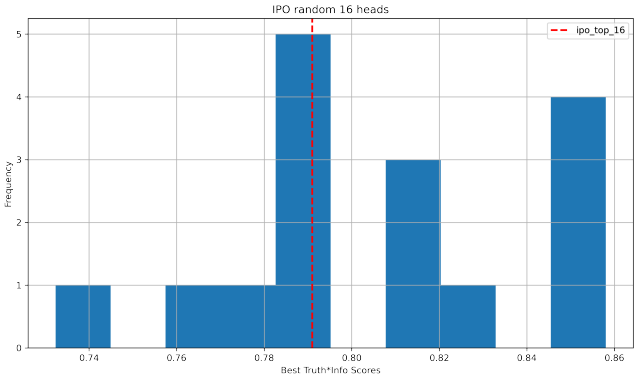
Here $\boldsymbol{b}^l_h$ plays the role of the intervention vector $\boldsymbol{\theta}^l_h$, and $\boldsymbol{a}^l$ decides the intervention strength adaptively.

[1] One could replace $\langle \boldsymbol{a}^l, \boldsymbol{o}^l \rangle$ with a constant intervention strength, but allowing the extra flexibility is closer to the ideal of best-possible-localized-intervention.

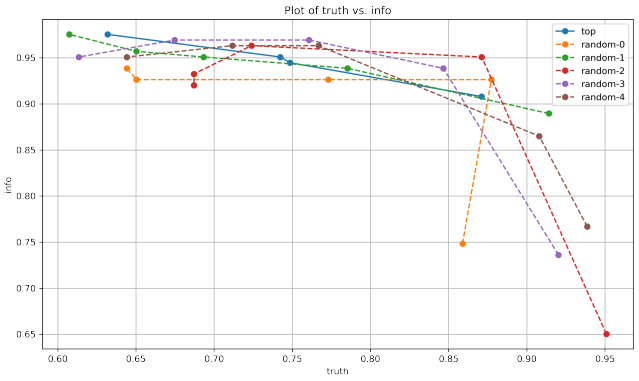
Now we have the algorithm to find the optimal interventions for the chosen set of heads:

1. Finetune the model weights using reparameterized LoRA with the IPO objective.
2. And, restrict b^l to be nonzero only for the chosen set of heads.

5. Optimal interventions at localized heads are nearly optimal, but so are random heads



(a) Info*Truth Scores



(b) Truth vs Info Scores

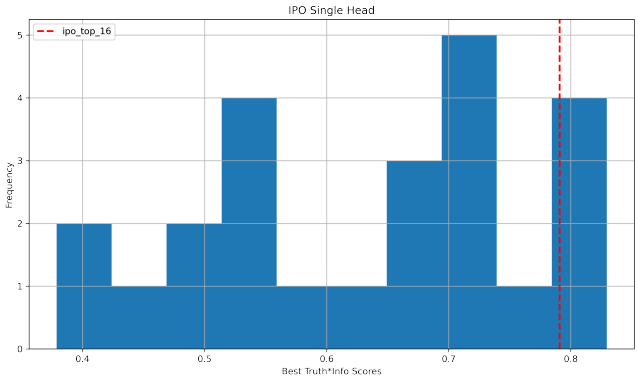
Figure 3. Using IPO optimal localized interventions, randomly selected heads perform nearly optimally for steering model generations. In particular, random heads are as good as the conjectured localized heads. The random heads are the same as those in fig. 1b.

Optimal Edits at Conjectured Localization We can now search for the best possible interventions at the localized heads. Figure 2 shows the result. We find that the optimal interventions strongly outperform the heuristic ITI interventions. Moreover, the localized interventions are about as effective as full IPO alignment! This appears to be the strongest edit-based evidence for localization that we could hope for.

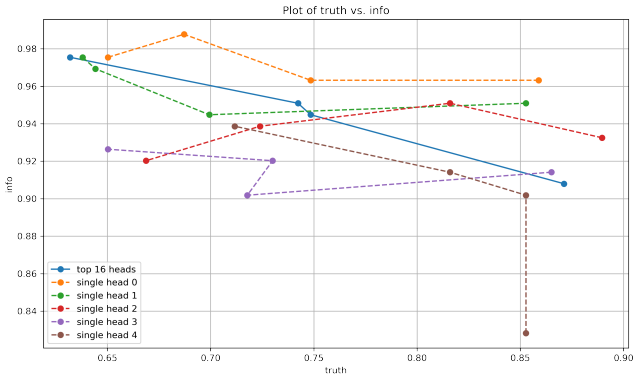
Optimal Edits at Random Localization Now, we apply the same optimal edit procedure to 16 *randomly selected* heads. Figure 3 shows the results. In short: the optimal interventions at random heads are often just as effective as the optimal interventions at the localized heads. Accordingly, the fact that editing at the localized heads was effective at steering generations provides no evidence that the truthfulness concept is localized to those heads.

Further, the random heads we use here are the same random heads used in section 3. Using the ITI heuristic intervention, the selected heads looked highly different from these random heads. But we now see that this appears to be an artifact of the suboptimal interventions and choice of metric, rather than a meaningful difference in how the heads relate to truthfulness.

6. Intervening a single head is just as effective



(a) Info*Truth Scores



(b) Truth vs Info Scores

Figure 4. Using a single-head is as effective, and there are multiple of them!

It is now clear that edit-based evidence does not provide strong evidence for localization in the 16 head setup. However, a possible way of saving localization would be to argue that 16 heads is too many, giving too much leeway to induce any behavior we want with editing. For example, if we edited half the heads of the model, it would not be surprising if we could make the model do anything we wanted. Accordingly, we might hope that there is still a valid syllogism of

the form "the localized edit is extremely constrainted" and "edits at this location optimally control the target behavior" implies "the target behavior is localized to this location".

To test this, we now focus on the single head case. The procedure is simple: we randomly sample 24 single heads, one at a time, and search for optimal interventions. The distribution of the best Truth*Info scores is shown in fig. 4a. We find 5 single-heads that are as effective fig. 4b, and none of them has high probing accuracy. Notice that, still, none of these heads can be understood as localizing the truthfulness concept. The reason is that there are multiple distinct locations that work equally well! That is, even in the extreme case of a very localized edit that replicates the target behavior essentially optimally, we still cannot conclude that there is evidence supporting localization.

7. Are the Probing-Localized Heads Anything Special?

So far what we mean by localization, is that we can change model generation on target concept by an edits at this location. And our experiments show no evidence for this type of localization, and probing-localized heads play no special role.

So, are the probing-localized heads anything special at all?

Probing-localized heads seems special for MC scores We do observe that these heads achieve slightly better Multiple-Choice (MC) scores compared to randomly selected heads (see fig. 5), although this advantage is not as pronounced as with the ITI interventions (see fig. 1c). Thus, these heads may be special in terms of changing model probabilities on the given fixed dataset, which is what MC measures.

The gap between what the model "knows" and what it generates It's important to note that the model's probabilities for fixed responses, do not directly correspond to what the model actually generates. Even if the model assigns a higher probability to a truthful response than an untruthful one, it may still not generate the truthful response if the fixed dataset is off-policy (i.e. both probabilities are low). This highlights the well-known gap between what a model "knows" (which is the motivation behind probing) and what it ultimately generates (Joshi et al., 2023; Wang et al., 2020; Kadavath et al., 2022; Saunders et al., 2022; Burns et al., 2022).

Implications It's possible that while probing-localized heads are not special at all for controlling model generations , they are special in changing what the model "knows". Though we caution that the results here are not rigorous evidence for localization even in this sense. Even if there

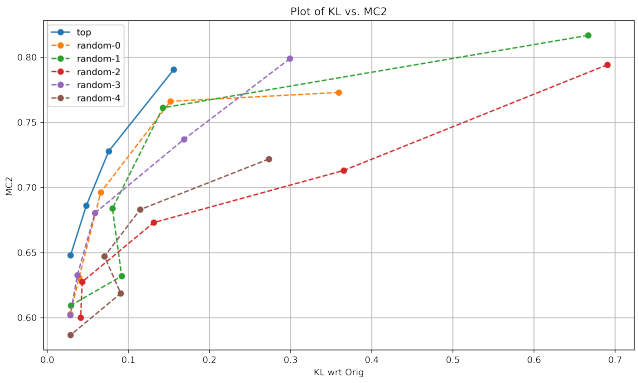


Figure 5. Probing-localized heads seem somewhat special in MC scores.

is a knowledge localization in some sense, it is clear that this does not inform steering, and does not give a way of monitoring model behavior (because changes in completely unrelated locations can change the behavior). This points to the need for making the goal of localization precise.

8. Discussion

The main idea in this paper is that to assess the localization of a behavior we should study the effect of the *optimal* intervention at the conjectured localization. The main obstacle is that, in general, it is not clear how to define or find the optimal intervention. To overcome this, we map the problem of finding the optimal intervention to the problem of finding the optimal weight update, which can be solved using existing LLM alignment methods.

The main result is an example where, naively, the evidence for localization appears strong, but when we use optimal interventions, the evidence disappears.

The particular example—truthfulness and ITI-based evidence—was selected simply because the data used to define the heuristic happens to also allow us to set up a contrastive alignment problem. The most limited read of the results here is that ITI interventions do not provide evidence for localization, and that truthfulness does not appear to be localizable. However, the broader point is that by giving an example where editing-based evidence doesn't support localization, we see that in general such edits—by themselves—cannot provide evidence for localization. This is true irrespective of the particular behavior or heuristic being evaluated.

Thus far, we've been a bit vague about what localization means. Editing does tautological evidence for localization in the sense of "it's possible to modify model behavior on such-and-such a behavior by an edit at this location". On the opposite end, the strongest possible standard would be to show that the location is unique, or at least necessary.

This is the standard that would be required if our aim was, e.g., to establish that LLM truthfulness can be monitored by examining a small set of heads. Potentially, there are interesting and useful notions of localization in between these two extremes. However, we can see no useful sense of localization that is consistent with the location being only as good as a *randomly selected* alternative. As we have seen, heuristic edit-based evaluation cannot even rule out this case.

Our findings add to a growing body of work that assesses the validity of interpretability results. Niu et al. (2024) argue that the Knowledge Neuron thesis, which suggests that facts are stored in MLP weights, is an oversimplification and does not adequately explain the process of factual expression in language models.Makelov et al. (2023) demonstrate that subspace activation patching can lead to an illusory sense of interpretability, as the effects may be achieved through dormant parallel pathways rather than the hypothesized subspaces. Most relevant to our work, Hase et al. (2024) find that localization conclusions from causal tracing do not provide insight into which model MLP layer would be best to edit to override an existing stored fact.

Overall, the results here point to the need for precise statements of what the objectives are in interpretability. With clear objectives, it may be possible to develop theoretically grounded methods for evaluation. Precise, falsifiable, statements and clear standards of evidence would suffice to prevent the kind of failure we observe in this paper.

Acknowledgements

This work is supported by ONR grant N00014-23-1-2591 and Open Philanthropy.

References

Arditi, A., Obeso, O., Syed, A., Paleka, D., Rimsky, N., Gurnee, W., and Nanda, N. Refusal in language models is mediated by a single direction. *arXiv preprint arXiv:2406.11717*, 2024.

Azar, M. G., Guo, Z. D., Piot, B., Munos, R., Rowland, M., Valko, M., and Calandriello, D. A general theoretical paradigm to understand learning from human preferences. In *International Conference on Artificial Intelligence and Statistics*, pp. 4447–4455. PMLR, 2024.

Burns, C., Ye, H., Klein, D., and Steinhardt, J. Discovering latent knowledge in language models without supervision. *arXiv preprint arXiv:2212.03827*, 2022.

Chan, L., Garriga-Alonso, A., Goldowsky-Dill, N., Greenblatt, R., Nitishinskaya, J., Radhakrishnan, A., Shlegeris, B., and Thomas, N. Causal scrubbing: A method for rigorously testing interpretability hypotheses. In *AI Alignment Forum*, pp. 1828–1843, 2022.

Chen, Z., Sun, X., Jiao, X., Lian, F., Kang, Z., Wang, D., and Xu, C. Truth forest: Toward multi-scale truthfulness in large language models through intervention without tuning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 20967–20974, 2024.

Conmy, A., Mavor-Parker, A., Lynch, A., Heimersheim, S., and Garriga-Alonso, A. Towards automated circuit discovery for mechanistic interpretability. *Advances in Neural Information Processing Systems*, 36:16318–16352, 2023.

Finlayson, M., Mueller, A., Gehrmann, S., Shieber, S., Linzen, T., and Belinkov, Y. Causal analysis of syntactic agreement mechanisms in neural language models. *arXiv preprint arXiv:2106.06087*, 2021.

Geiger, A., Lu, H., Icard, T., and Potts, C. Causal abstractions of neural networks. *Advances in Neural Information Processing Systems*, 34:9574–9586, 2021.

Hanna, M., Liu, O., and Variengien, A. How does gpt-2 compute greater-than?: Interpreting mathematical abilities in a pre-trained language model. *Advances in Neural Information Processing Systems*, 36, 2024.

Hase, P., Bansal, M., Kim, B., and Ghandeharioun, A. Does localization inform editing? surprising differences in causality-based localization vs. knowledge editing in language models. *Advances in Neural Information Processing Systems*, 36, 2024.

Hendel, R., Geva, M., and Globerson, A. In-context learning creates task vectors. *arXiv preprint arXiv:2310.15916*, 2023.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Joshi, N., Rando, J., Saparov, A., Kim, N., and He, H. Personas as a way to model truthfulness in language models. *arXiv preprint arXiv:2310.18168*, 2023.

Kadavath, S., Conerly, T., Askell, A., Henighan, T., Drain, D., Perez, E., Schiefer, N., Hatfield-Dodds, Z., DasSarma, N., Tran-Johnson, E., et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Li, K., Patel, O., Viégas, F., Pfister, H., and Wattenberg, M. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, 36, 2024.

Lin, S., Hilton, J., and Evans, O. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*, 2021.

Makelov, A., Lange, G., Geiger, A., and Nanda, N. Is this the subspace you are looking for? an interpretability illusion for subspace activation patching. In *The Twelfth International Conference on Learning Representations*, 2023.

Meng, K., Bau, D., Andonian, A., and Belinkov, Y. Locating and editing factual associations in gpt. *Advances in Neural Information Processing Systems*, 35:17359–17372, 2022.

Niu, J., Liu, A., Zhu, Z., and Penn, G. What does the knowledge neuron thesis have to do with knowledge? *arXiv preprint arXiv:2405.02421*, 2024.

OpenAI. Openai api, 2020. URL https://openai.com/blog/openai-api/. Accessed: 2021-08-19.

Saunders, W., Yeh, C., Wu, J., Bills, S., Ouyang, L., Ward, J., and Leike, J. Self-critiquing models for assisting human evaluators. *arXiv preprint arXiv:2206.05802*, 2022.

Soulos, P., McCoy, T., Linzen, T., and Smolensky, P. Discovering the compositional structure of vector representations with role learning networks. *arXiv preprint arXiv:1910.09113*, 2019.

Taori, R., Gulrajani, I., Zhang, T., Dubois, Y., Li, X., Guestrin, C., Liang, P., and Hashimoto, T. B. Alpaca: A strong, replicable instruction-following model. *Stanford Center for Research on Foundation Models. https://crfm. stanford. edu/2023/03/13/alpaca. html*, 3(6):7, 2023.

Todd, E., Li, M. L., Sharma, A. S., Mueller, A., Wallace, B. C., and Bau, D. Function vectors in large language models. *arXiv preprint arXiv:2310.15213*, 2023.

Vig, J., Gehrmann, S., Belinkov, Y., Qian, S., Nevo, D., Sakenis, S., Huang, J., Singer, Y., and Shieber, S. Causal mediation analysis for interpreting neural nlp: The case of gender bias. *arXiv preprint arXiv:2004.12265*, 2020.

Wang, C., Liu, X., and Song, D. Language models are open knowledge graphs. *arXiv preprint arXiv:2010.11967*, 2020.

Wang, H. and Shu, K. Backdoor activation attack: Attack large language models using activation steering for safety-alignment. *arXiv preprint arXiv:2311.09433*, 2023.

Wang, K., Variengien, A., Conmy, A., Shlegeris, B., and Steinhardt, J. Interpretability in the wild: a circuit for indirect object identification in gpt-2 small. *arXiv preprint arXiv:2211.00593*, 2022.

Wei, B., Huang, K., Huang, Y., Xie, T., Qi, X., Xia, M., Mittal, P., Wang, M., and Henderson, P. Assessing the brittleness of safety alignment via pruning and low-rank modifications. *arXiv preprint arXiv:2402.05162*, 2024.

Zou, A., Phan, L., Chen, S., Campbell, J., Guo, P., Ren, R., Pan, A., Yin, X., Mazeika, M., Dombrowski, A.-K., et al. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*, 2023.

A. Experiment Details

Dataset and Model Architecture We use the TruthfulQA dataset (Lin et al., 2021) and the Alpaca-7B model (Taori et al., 2023) for our experiments. The dataset contains 817 questions with truthful and untruthful answers. We turn them into pairs, and use 60% for training (6560 paired data) and the rest for validation and testing. The model consists of 32 layers, each with 32 attention heads and a hidden dimension of 4096.

Training Details We use IPO objective (Azar et al., 2024) and use hyperparameter $\tau = 0.1, 0.2, 0.3, 0.4, 0.5$. We train for two epochs with a cosine scheduler, with a batch size of 4. We use "paged_adamw_32bit" optimizer. For training with different numbers of heads, we find a smaller number of heads benefit from a higher learning rate. For all-heads, we use a learning rate of 1×10^{-4}, and for 16 heads, we use 5×10^{-4}. For single-head, we use 2×10^{-3}.

Evaluation Metrics We reuse code from ITI (Li et al., 2024) for evaluation when possible. For GPT-judge models, we follow (Lin et al., 2021) and finetune on truthfulness and informativeness dataset using OpenAI API (OpenAI, 2020). Our finetuned model achieves similar validation error as in (Lin et al., 2021).