Exploring length generalization in the context of indirect object identification task

Mechanistic interpretability aims to explain how neural networks learn at the circuit level. So far only a handful of circuits with compelling evidence have been discovered in relatively small language models [1,2,3]. It remains unclear whether these circuits (a) persist once formed after training further, and (b) have similar explanatory power in larger models or models with different architecture. A recent case study on reverse engineering the circuit behind indirect object identification (IOI) task has uncovered an interesting circuit with a fairly specialized division of labor [3]. While the authors performed extensive experiments and ablation analyses to validate the specific circuit components (i.e., specialized attention heads), how the model performance varies under different perturbations is underexplored. Here we study the performance on IOI task and its variants in the original GPT-2 small and other models of equivalent size. This analysis makes some progress toward addressing the following open research questions curated by Neel Nanda: 5.34 and 5.35 [4]. Please refer to the following Github repo for additional details: https://github.com/cx0/mech-interpretability

We have compiled 8 large language models that have approximately the same model size (number of parameters) as the GPT-2 small used in the original study [3] (see **Table 1**). Notably, these models have the same architecture (layers, attention heads, MLP) and vocabulary size. Some models have different activation functions (GeLU vs ReLU) and larger context window size but these differences are not relevant for our analysis. These models are conveniently implemented in TransformerLens package [5], a highly versatile and capable suite for mechanistic interpretability research.

Table 1. Large language models used in this analysis. Models have the same architecture. Differences between models are highlighted in bold. Please see TransformerLens model summary page for all the models.

Model name	n_params	n_layers	d_model	n_heads	act_fn	n_ctx	d_vocab	d_head	d_mlp
GPT-2 small	85M	12	768	12	GeLU	1024	50257	64	3072
OPT small	85M	12	768	12	**ReLU**	**2048**	**50272**	64	3072
GPT Neo small	85M	12	768	12	GeLU	**2048**	50257	64	3072
CRFM-small-a	85M	12	768	12	GeLU	1024	50257	64	3072
CRFM-small-b	85M	12	768	12	GeLU	1024	50257	64	3072
CRFM-small-c	85M	12	768	12	GeLU	1024	50257	64	3072

CRFM-small-d	85M	12	768	12	GeLU	1024	50257	64	3072
CRFM-small-e	85M	12	768	12	GeLU	1024	50257	64	3072

We calculated the IOI task performance for each model listed above using the original task prompt – “After John and Mary went to the store, John gave a bottle of milk to” – and the expected response “Mary”. IOI task performance showed high variability among models of approximately the same size (see **Table 2)**. Stanford CRFM small models (nicknamed a to e here) are replications of the original GPT-2 small models trained on 5 different seeds. Notably, the best and worst performing CRFM models show a substantial 28 percentage-point difference indicating the importance of training dynamics for the performance on the IOI task.

The algorithm implemented by the discovered circuit has specific heads attending to identifying duplicate tokens, moving this information around, and finally removing the duplicate tokens. This algorithm keeps track of object count and makes use of this information when completing the task. Therefore, if the model actually employs this algorithm for IOI task, the model performance should not be affected by either the order of names (John and Mary vs Mary and John) and/or the indirect object of interest (correct answer token). Clearly, each token has different base probabilities depending on the training corpus, and some names are more frequent than others. However, the difference in base probabilities for different names is arguably negligible from the perspective of this particular algorithm implementation.

To test whether the model performance is sensitive to name order and choice of indirect object of interest, we introduced two variants of the original prompt:

- Original prompt: “After **John** and **Mary** went to the store, **John** gave a bottle of milk to” (expected response: **Mary**)
- Alternate prompt with swapped names: “After **Mary** and **John** went to the store, **John** gave a bottle of milk to” (expected response: **Mary**)
- Alternate prompt with swapped indirect object: “After **John** and **Mary** went to the store, **Mary** gave a bottle of milk to” (expected responses: **John**)

IOI task performance on these modified prompts shows remarkable difference among models as well across prompts for the same model (see **Table 2**). For example, when the name order is swapped (i.e., Mary and John), CRFM-small-a model outputs a 94.11% probability for the correct token when the indirect object is kept the same while the same model’s output drops to 50.69% when the indirect object of interest is swapped as well. Performance changes among models do not exhibit any consistent behavior across prompts to speculate any potential source of bias. It is important to note that these modifications yield semantically equivalent prompts and these small models do not show strong performance robustness.

Table 2. IOI task performance of GPT-2 small model and other models of equivalent model size. Top row indicates the order of names in the prompt used in our analysis. Indirect object of interest is underscored and highlighted boldly. Task performance was measured as correct token probability. Average performance across all models was shown at the bottom row.

	"After **John** and **Mary** went to the store"		"After **Mary** and **John** went to the store"	
Model name	John to **Mary**	Mary to **John**	John to **Mary**	Mary to **John**
GPT-2 small	70.07%	83.83%	87.90%	60.07%
OPT small	74.87%	50.92%	66.09%	57.25%
GPT Neo small	41.19%	60.94%	41.98%	52.36%
CRFM-small-a	79.65%	86.33%	94.11%	50.69%
CRFM-small-b	72.24%	67.36%	82.33%	70.47%
CRFM-small-c	59.06%	51.60%	61.94%	48.85%
CRFM-small-d	87.71%	58.10%	88.17%	61.89%
CRFM-small-e	74.78%	82.26%	84.77%	76.38%
Average	69.94%	67.66%	75.91%	59.74%

To further probe the stability of task performance while keeping in mind that the IOI circuit attends to duplicate tokens, we introduced a new prompt modification that introduces an extra copy of each name. Similar to the previous modification setup, an extra copy of each name in the prompt should not impact the model performance unless, for example, the inhibition head behaves differently in the context of more than one copy. This prompt modification was chosen to mimic length generalization which is a formidable capability of large language models to showcase the extent of their reasoning and learning skills. The new prompt appends a short sentence including both names relevant to the task:

Original prompt: "After **John** and **Mary** went to the store, **John** gave a bottle of milk to" (expected response: **Mary**)
Alternate prompt with extra name copies: "**John** and **Mary** are friends. After **John** and **Mary** went to the store, **John** gave a bottle of milk to" (expected response **Mary**)

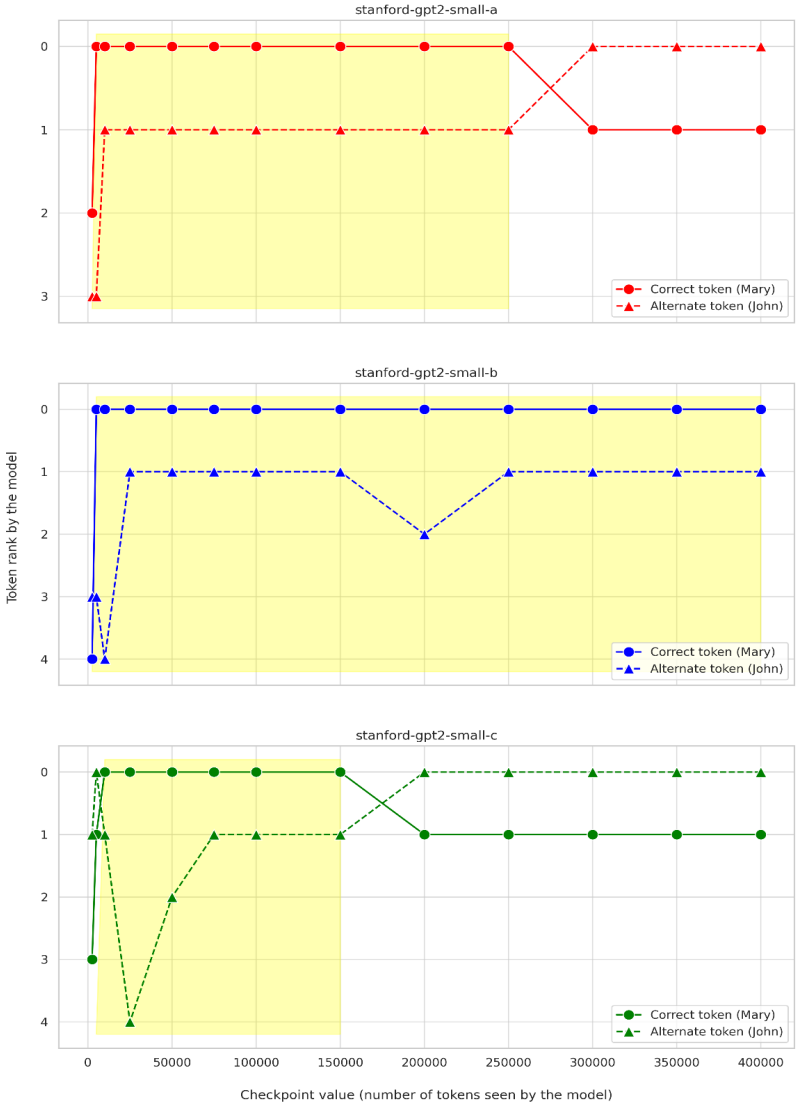
Table 3. IOI task performance when all prompt modifications were included. Indirect object of interest (task goal) is highlighted in bold and underscored for each scenario. Correct token probabilities are shown in parentheses and highlighted in purple.

	"**John** and **Mary** are friends. After **John** and **Mary** went to the store"		"**John** and **Mary** are friends. After **Mary** and **John** went to the store"	
Model name	John to **Mary**	Mary to **John**	John to **Mary**	Mary to **John**
GPT-2 small	86.23%	90.68%	53.59%	93.40%
OPT small	91.90%	81.16%	76.85%	84.83%
GPT Neo small	66.77%	70.58%	37.91%	82.16%
CRFM-small-a	88.37%	94.40%	9.99% (74.84%)	94.61%
CRFM-small-b	90.68%	84.58%	72.88%	92.79%
CRFM-small-c	85.77%	80.25%	24.37% (47.69%)	87.75%
CRFM-small-d	79.84%	73.72%	54.53%	72.68%
CRFM-small-e	94.85%	85.08%	49.00%	89.48%
Average	85.55%	82.55%	n/a	87.21%

This modification significantly boosts the model performance almost across the board and especially improves the performance for OPT model (Meta AI's open source model) except for the scenario where the name order was swapped while keeping the indirect object of interest the same. In particular, CRFM-small-a and CRFM-small-c models generated incorrect responses (John instead of Mary) with high confidence. In order to investigate the remarkably poor performance of these models, we checked the model performance of three comparable models from the same family (CRFM-small-a, CRFM-small-b, and CRFM-small-c) at different checkpoints during their training. These checkpoints are publicly available and convenient to load and probe using TransformerLens package.

We found that al the three selected models learned the correct token in this particular case very early in their training, however, the two poor performing models switched to favoring the alternate incorrect response in the second half of their full training (see **Figure 1**).

Figure 1. Model performance as measured by the top token rank (0-indexed) for each model during various checkpoints. The training region in which the model picks the correct response is highlighted in yellow.



Encouraged by the observation that these models were able to correctly predict the next token early in their training, we densely sampled available checkpoints (number of tokens seen by the model) to investigate the early training period as well as the training dynamics around the time where the two poor performing models began to pick the incorrect token. Figure 2 shows the corresponding model performance starting from checkpoint 50,000 to the end of the training. Interestingly, there are relatively short periods in training (indicated by black arrows in the figure) where the model learned to pick the correct answer shortly before switching back to favoring the incorrect token. In particular, CRFM-small-a provides an opportunity to study and better understand the importance of training dynamics and path dependence in model performance. This model is equivalent to CRFM-small-b in every aspect except trained on a different seed,

and shows remarkably strong performance similar to CRFM-small-b during the first half of its training right before an abrupt drop in performance. If the IOI circuit is acquired early on, it would be interesting to empirically test in this model whether the poor performance can be directly attributed to some deficiency in the circuit (e.g., losing one of the formed heads).

Figure 2. Model performance over densely sampled checkpoints during training of the three selected models (CRFM-small-a, CRFM-small-b and CRFM-small-c). The training region in which the model picks the correct response is highlighted in yellow. Black arrows point at narrow training segments where the model picked the correct response.

Next we applied the name order swap modification to the sentence with the extra copy of names. We found that the same poor performing models from previous analysis (namely, CRFM-small-a and CRFM-small-c) were also susceptible to one of the modified prompts in this new setting (see **Table 4**).

Table 4. IOI task performance when name order swap was applied to the extra copy of names as well. Indirect object of interest is underscored and highlighted boldly. Task performance was measured as percentage performance on the answer token. Average performance across all models was shown at the bottom row.

	"**Mary** and **John** are friends. After **John** and **Mary** went to the store"		"**Mary** and **John** are friends. After **Mary** and **John** went to the store"	
Model name	John to **Mary**	Mary to **John**	John to **Mary**	Mary to **John**
GPT-2 small	93.91%	61.46%	91.58%	76.15%
OPT small	95.16%	57.87%	93.74%	87.70%
GPT Neo small	75.06%	36.43%	59.01%	61.53%
CRFM-small-a	94.60%	8.44% (76.89%)	93.89%	58.87%
CRFM-small-b	91.63%	50.69%	85.57%	87.65%
CRFM-small-c	93.47%	17.47% (60.84%)	90.40%	76.98%
CRFM-small-d	89.68%	32.26% (32.27%)	88.17%	68.11%
CRFM-small-e	89.30%	52.08%	75.44%	87.95%
Average	90.35%	n/a	84.72%	75.61%

It is difficult to speculate whether these two specific modifications pose a similar type of challenge during training for the selected models. The only sentence structure shared between these prompts is the immediate repetition of the same name. In the first prompt, the name order in the scenario where the two models performed poorly is John-Mary-Mary-John-John-[Mary], and the second prompt's structure is Mary-John-John-Mary-Mary-[John]. It is conceivable that the IOI circuit may be particularly vulnerable to order effects when multiple token copies appear in succession. This may limit the generalization of IOI circuit and similar tokens for tasks involving longer context (therefore, more likely to involve multiple token copies). On the other hand, we tested whether these larger model variants were susceptible to similar performance issues and found that CRFM-medium models were able to correctly guess the answer token when the same modified prompt was presented (see **Appendix Table 1**).

To control for the importance of multiple duplicate copies, we created a new prompt by replacing one of the copies with the corresponding pronoun:

Original prompt: "After **John** and **Mary** went to the store, **John** gave a bottle of milk to" (expected response: **Mary**)

Alternate prompt with an extra name copy removed: "**John** and **Mary** are friends. After **they** went to the store, **John** gave a bottle of milk to" (expected response **Mary**)

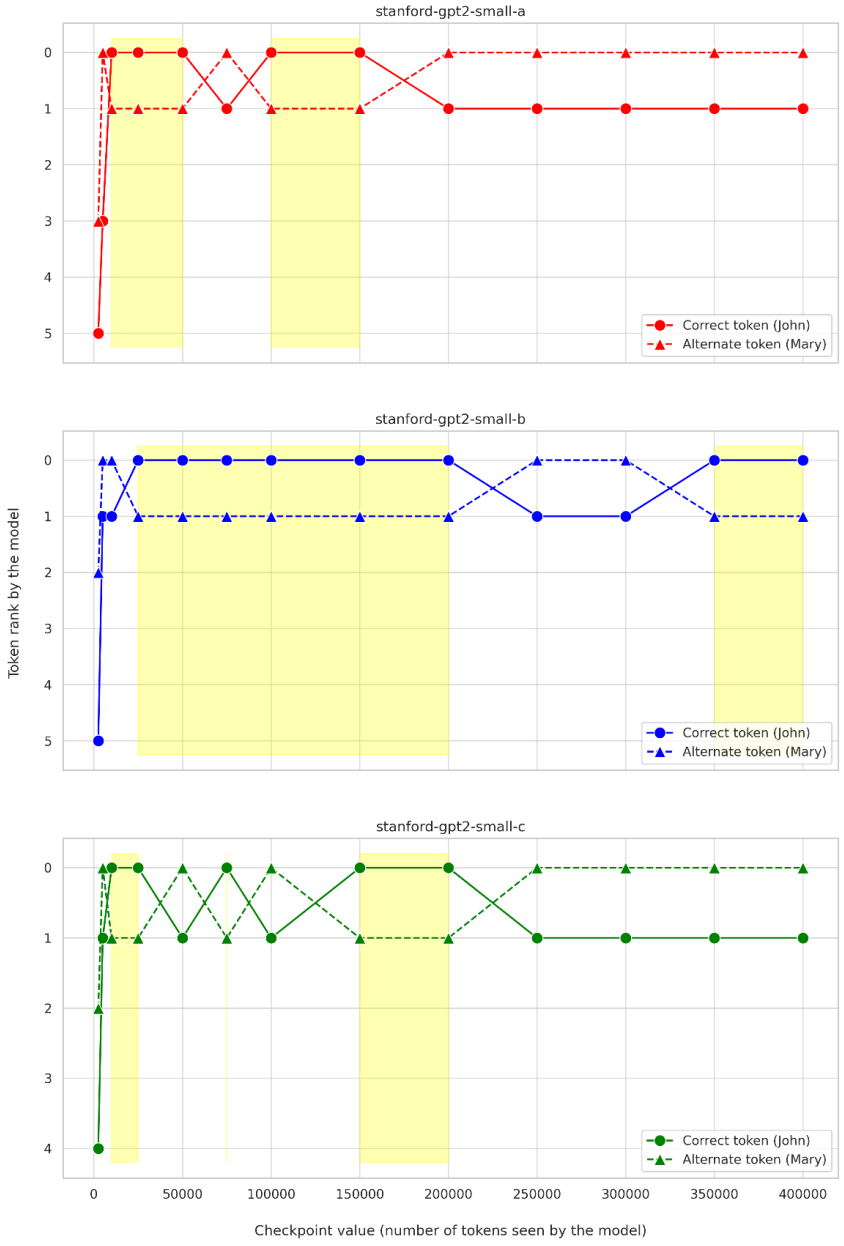
When we used this new prompt with an extra name copy removed, all the models including the two poor performing CRFM-small models were able to pick the correct response (see **Table 5**).

T**able 5.** IOI task performance when the extra name copy was removed to match the same number of name tokens as the original prompt included in IOI circuit paper.

	"**John** and **Mary** are friends. After **they** went to the store"		"**Mary** and **John** are friends. After **they** went to the store"	
Model name	John to **Mary**	Mary to **John**	John to **Mary**	Mary to **John**
GPT-2 small	84.98%	89.49%	93.84%	72.68%
OPT small	92.49%	83.12%	97.12%	85.29%
GPT Neo small	65.87%	78.68%	69.40%	62.68%
CRFM-small-a	75.40%	94.00%	95.77%	51.53%
CRFM-small-b	88.91%	89.69%	90.92%	79.93%
CRFM-small-c	80.39%	86.02%	94.21%	62.38%
CRFM-small-d	69.71%	78.79%	88.74%	65.36%
CRFM-small-e	89.32%	87.84%	93.71%	74.98%

Similar to the previous training dynamics analysis, we investigated the model performance of these three models during their training. In this setting, the poor performing models had difficulty at picking the correct response from the start of their training (see **Figure 3**). Interestingly, CRFM-model-b was only able to switch back to the correct token response at the end of their training. These problems in model performance robustness pose a great challenge when evaluating task-specific model capabilities. Identifying circuits and validating their generalization presents a real opportunity to address these shortcomings.

Figure 3. Model performance as measured by the top token rank (0-indexed) for each model during various checkpoints. The training region in which the model picks the correct response is highlighted in yellow.



Appendix

Table S1. Model performance of selected CRFM-medium models on the prompts that their smaller model variants performed poorly.

	"**Mary** and **John** are friends. After **John** and **Mary** went to the store"	
Model name	John to **Mary**	Mary to **John**
CRFM-small-a	94.60%	8.44% **(76.89%)**
CRFM-small-b	91.63%	50.69%
CRFM-small-c	93.47%	17.47% **(60.84%)**
CRFM-small-d	89.68%	32.26% **(32.27%)**
CRFM-small-e	89.30%	52.08%

Table S2. Model performance of selected CRFM-medium models on the prompts that their smaller model variants performed poorly.

	"**Mary** and **John** are friends. After **John** and **Mary** went to the store"	
Model name	John to **Mary**	Mary to **John**
CRFM-medium-a	93.25%	74.61%
CRFM-medium-b	86.82%	77.01%
CRFM-medium-c	74.34%	43.55%
CRFM-medium-d	85.05%	84.46%
CRFM-medium-e	85.12%	55.68%

References

[1] 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. A mathematical framework for transformer circuits. Transformer Circuits Thread, 2021. https://transformer-circuits.pub/2021/framework/index.html

[2] Nanda, Neel, et al. "Progress measures for grokking via mechanistic interpretability." *arXiv preprint* *arXiv:2301.05217* (2023).

[3] Wang, Kevin, et al. "Interpretability in the wild: a circuit for indirect object identification in gpt-2 small." *arXiv preprint* *arXiv:2211.00593* (2022).
[4] 200 Concrete Problems in Interpretability
[5] https://github.com/neelnanda-io/TransformerLens