Extended Abstract Track

Structured In-Context Task Representations

Editors: List of editors' names

Abstract

Representation learning has been central to deep learning's evolution. While interpretable structures have been observed in pre-trained models' representations, an important question arises: Do networks develop such interpretable structures during in-context learning? Using synthetic sequence data derived from underlying geometrically structured graphs (e.g., grids, rings), we provide affirmative evidence that language models develop internal representations mirroring these geometric structures during in-context learning. Furthermore, we demonstrate how in-context examples can override semantic priors by constructing a representation in dimensions other than the one used by the prior. Overall, our study demonstrates that models can form meaningful representations solely from in-context exemplars.

Keywords: Interpretability, In-context learning

1. Introduction

Researchers have found various geometric representations in the activations of language models, such as linear representations for "truthfulness" (Marks and Tegmark, 2024), "refusal" (Arditi et al., 2024), or even "world models" (Li et al., 2022; Nanda et al., 2023), as well as non-linear features including circular representations (Engels et al., 2024) for periodic concepts, or "onion" representations for a simple token repetition task (Csordás et al., 2024) (see Appendix A for a more comprehensive background).

Meanwhile, language models are also able to solve new tasks that are specified solely by inference time exemplars. This ability is often referred to as in-context learning (ICL) (Brown et al., 2020). A natural questions which arises is: *"Do models create task dependent representations solely from in-context exemplars?"*

In this work, we demonstrate that language models can construct geometric representations of activations reflecting the structure of the given in-context tasks. We design synthetic data generating processes (DGP) with an underlying geometrical graph. We then sample tokens from this DGP with a rule defined on the graph. Given enough exemplars, we evaluate whether the model follows the rules and extract the hidden activations of the model to examine the representations of tokens involved in the task.

We intentionally choose tokens that do not contain any relationship that pertains to the graphical structure (ex: "`apple`", "`opera`") and find that with a sufficient number of exemplars, the model not only learns the task (i.e., infers the newly specified relationship amongst the tokens), but more importantly captures the newly specified task geometry in the first few principal components.

Interestingly, we find that if we use tokens that already have a semantic geometry (e.g., days of the week (Engels et al., 2024)), the model can override this semantic prior to perform the task. In this case, we find that the largest principal components still capture the semantic prior, while lower principal components capture the newly specified geometry.

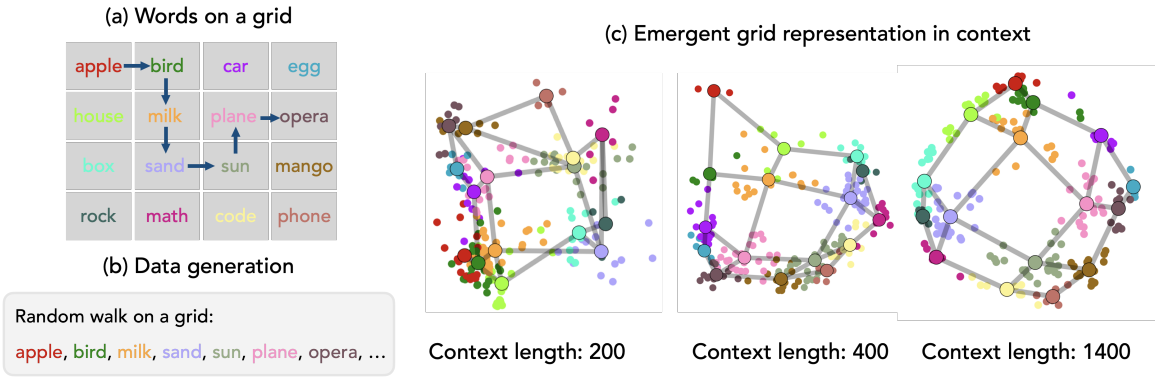

Figure 1: **Llama constructs a grid from observations of a traversal of a 4x4 board.** We randomly order a set of tokens in a 4x4 grid, and randomly traverse the grid, resulting in a sequence of tokens. From observing such a sequence, the model can represent the ground truth shape of the grid in the first two principal components of the tokens' mean activations. We use random tokens that *do not inherently carry the* 4x4 *geometry* to represent nodes in the graph. With more examples and in deeper layers (Fig. 4), we see a clearer representation of the grid appear.

Overall, we discover that large language models can create task specific structured representations from solely in-context exemplars, which can live in lower principal components if there already exists a strong semantic representation.

2. Experiment Setup

For our experiments, we use Llama3.1-8B (Dubey et al., 2024). We experiment with tasks defined on two geometrical graphs: a ring and a grid. For each task, we assume a set of tokens T. In our grid task, we arrange 16 tokens with no strong semantic structure in a 4x4 grid, with edges between horizontal and vertical neighbors. In our ring task, we randomly order 10 tokens on a circle and define edges between neighboring tokens including a link between the first and the last token.

After defining these graphs, we apply a sampling rule on each graph to construct an in-context task. For the grid task, we perform a random walk on the graph, emitting the words on the visited nodes as a sequence (Fig. 1). For the ring task, we simply sample random pairs which are neighbors on the graph (Fig. 2).

For both tasks, the model successfully learns the rules from in context exemplars.

3. Uncovering Geometric Representations

Each of our tasks presents a sequence of tokens that originates from an underlying graph with a certain topology. To find the geometric relationship between tokens, we first compute each token's mean activations from the hidden layers. Namely, for a given layer ℓ and token

t, we collect all activations corresponding to the token t across N timesteps at layer ℓ. We then compute the mean activations per token $t \in T$, notated as $\bar{x}_t^\ell$.

Given mean activations $\bar{x}_t^\ell$ for tokens $t \in T$, we run PCA on our set of mean activations. We then visualize 2-dimensional projections using our two main principal components.

3.1. Grid Representation

Fig. 1 demonstrates the representational geometry for the 4x4 grid task. We observe that Llama-3.1-8B's internal representation indeed forms a 4x4 grid at deeper layers when given enough exemplars, preserving the ground truth neighboring structure. **This demonstrates that models can organize semantically unrelated words into a geometric representation that reflects the in-context task, given enough exemplars**. Interestingly, the corners of the board are collapsed inwards. We believe this is because of a natural under exploration of corner regions (see Fig. 5 in Appendix).

3.2. Ring Representation

Next, we ask: *Will a global structure of a knowledge graph emerge only by observing tiny subgraphs?* We construct a task where the given input sequence is a concatenation of multiple 1-step moves on a ring. The first token of the step is chosen in random, so that each exemplar only reveals one edge of the ring. Fig. 2 shows the results for this task (see Fig. 7 in Appendix for task accuracy). Here, we find again that semantically unrelated token representations organize into a ring structure adhering the specified ordering defined in the task. Recall again that *the tokens have no a priori reason to be organized in a ring.*

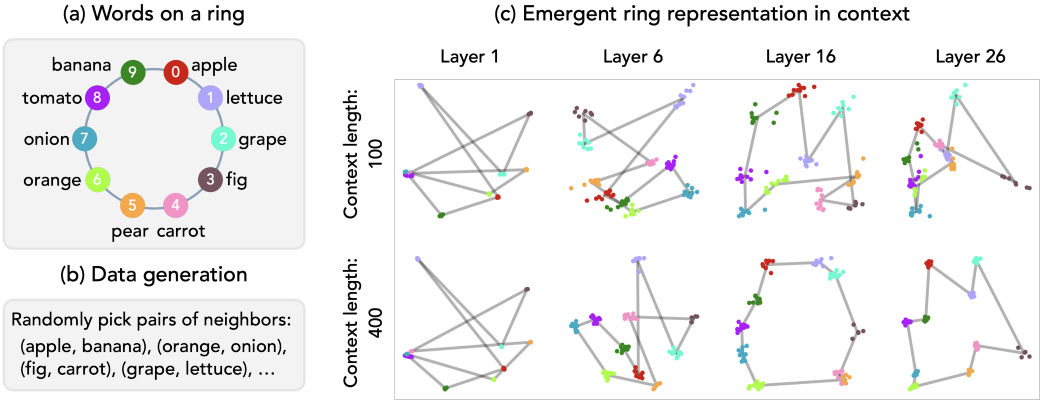

Figure 2: **LLMs can construct a representation of a global graph from many small in-context subgraphs** Given randomly sampled pairs on a ring graph, the model representation reconstructs the ring in its principal components.

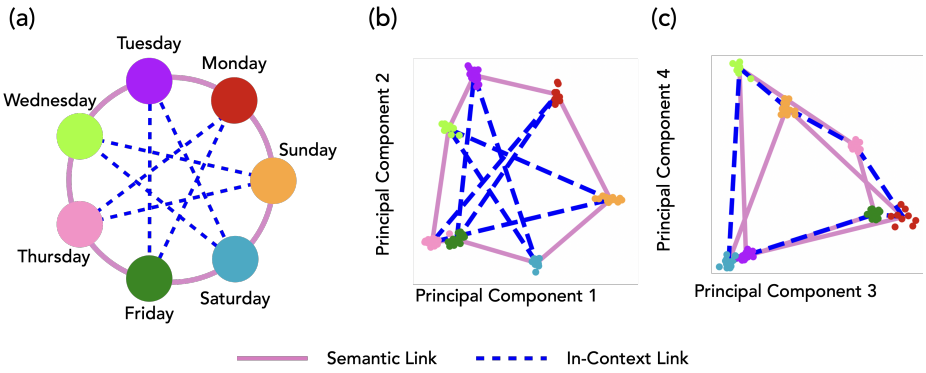

Figure 3: **Models can Override Semantic Priors using Representations in Higher Dimensions** We observe the semantic ring in the first two components, however the model forms the in-context ring in the two subsequent lower components.

3.3. Semantic Prior Inversion

Prior work has found sets of tokens already carrying a representational geometry (Engels et al., 2024). How would this geometry interfere with an in-context task when the task disagrees with a semantic prior? To answer this question, we repeat the previous experiment with a set of tokens that already has a semantically induced circular representation as shown in Engels et al. (2024): {`Mon, Tue, Wed, Thu, Fri, Sat, Sun`}. We then define a new ordering to define the in-context task: {`Mon, Thu, Sun, Wed, Sat, Tue, Fri`}.

The model still solves the task despite the need of more exemplars (Fig. 7, 10). However, in this case, we only see the semantic ring in the first two principal components. Interestingly, the lower third and fourth principal components forms a ring which adheres to the new ordering specified in-context (Fig. 3).

4. Conclusion

In this work, we demonstrate preliminary findings that geometric representations for tasks can also be formed in-context. Interestingly, when tokens with a semantic prior is used as the in-context prompt, we see the larger principal components encode the original semantic prior, while lower principal components encode the geometry specified in the context. While our work adds to our understanding of neural network representations, we also pose an interesting question: which representations are induced by model weights, and which are induced in-context? We view this line of work as an exciting area for further study.

5. Limitations

A primary limitation is the lack of a causal study. Currently it is unclear if the representations have a correlational or causal relationship with the model's predictions. Our results could also be made robust by testing on different language models and different sets of tokens.

References

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. *arXiv preprint arXiv:2406.11717*, 2024.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. *arXiv preprint arXiv:2212.03827*, 2022.

Róbert Csordás, Christopher Potts, Christopher D Manning, and Atticus Geiger. Recurrent neural networks learn to store and generate sequences using non-linear representations. *arXiv preprint arXiv:2408.10920*, 2024.

Abhimanyu Dubey, TRIGGER, LATEX, ET, and AL. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Joshua Engels, Isaac Liao, Eric J. Michaud, Wes Gurnee, and Max Tegmark. Not all language model features are linear, 2024. URL https://arxiv.org/abs/2405.14860.

Jaden Fiotto-Kaufman, Alexander R Loftus, Eric Todd, Jannik Brinkmann, Caden Juang, Koyena Pal, Can Rager, Aaron Mueller, Samuel Marks, Arnab Sen Sharma, Francesca Lucchetti, Michael Ripa, Adam Belfki, Nikhil Prakash, Sumeet Multani, Carla Brodley, Arjun Guha, Jonathan Bell, Byron Wallace, and David Bau. Nnsight and ndif: Democratizing access to foundation model internals, 2024. URL https://arxiv.org/abs/2407.14561.

Roee Hendel, Mor Geva, and Amir Globerson. In-context learning creates task vectors. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9318–9333, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.624. URL https://aclanthology.org/2023.findings-emnlp.624.

Andrew Lee, Xiaoyan Bai, Itamar Pres, Martin Wattenberg, Jonathan K Kummerfeld, and Rada Mihalcea. A mechanistic understanding of alignment algorithms: A case study on dpo and toxicity. *arXiv preprint arXiv:2401.01967*, 2024.

Kenneth Li, Aspen K Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Emergent world representations: Exploring a sequence model trained on a synthetic task. *arXiv preprint arXiv:2210.13382*, 2022.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=aLLuYpn83y.

Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets, 2024. URL https://arxiv.org/abs/2310.06824.

Neel Nanda, Andrew Lee, and Martin Wattenberg. Emergent linear representations in world models of self-supervised sequence models. *arXiv preprint arXiv:2309.00941*, 2023.

Kiho Park, Yo Joong Choe, Yibo Jiang, and Victor Veitch. The geometry of categorical and hierarchical concepts in large language models. *arXiv preprint arXiv:2406.01506*, 2024a.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models, 2024b. URL https://arxiv.org/abs/2311.03658.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering llama 2 via contrastive activation addition. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15504–15522, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.828. URL https://aclanthology.org/2024.acl-long.828.

Eric Todd, Millicent L Li, Arnab Sen Sharma, Aaron Mueller, Byron C Wallace, and David Bau. Function vectors in large language models. *arXiv preprint arXiv:2310.15213*, 2023.

Appendix A. Related Work

Researchers have recently discovered numerous representations for human concepts. Park et al. (2024b) suggests that a language models consist of linear representations, and indeed numerous researchers have found concrete examples. These include "truthfulness" (Marks and Tegmark, 2024; Burns et al., 2022; Li et al., 2023), "refusal" (Arditi et al., 2024), toxicity (Lee et al., 2024), sycophancy (Rimsky et al., 2024), or even "world models" (Li et al., 2022; Nanda et al., 2023). Park et al. (2024a) finds that hierarchical concepts are represented with a tree-like structure consisting of orthogonal vectors.

A relevant line of work includes that of Todd et al. (2023) and Hendel et al. (2023). Both papers find that one can compute a vector from in-context exemplars that encode the task, such that adding such a vector during test time for a new input can correctly solve the task.

Language models do not always form linear representations. Perhaps most relevant to our work, Engels et al. (2024) finds circular feature representations for periodic concepts, such as days of the week or months of the year, using a combination of sparse autoencoders and PCA. An important distinction to make is that while they find non-linear, circular

representations for concepts with a circular semantic prior, we demonstrate that even **random tokens that have no reason to be organized as a ring a priori** can form ring structures when given a circular relationship in-context.

Other non-linear feature representations include that of Csordás et al. (2024), in which they finds that recurrent neural networks trained on token repetition can either learn an "onion"-like representation or a linear representation, depending on the model's width.

Unlike prior work, we find that in-context tasks with a specified structural pattern can be induced in-context.

Appendix B. Experimental Details

B.1. Activation collection

We generate synthetic sequence data from custom code. To run Llama-3.1 models, we use nnsight and compute provided by NDIF (Fiotto-Kaufman et al., 2024).

B.2. Evaluation

Accuracy We evaluate the accuracy by calculating the next token probabilities on the restricted set of tokens we operate on. In App. C, we show the rescaled accuracy, summing up all valid token probabilities.

Average Activation calculation We calculate the average activation of each token by collecting activations from a 200 token window for the board task and a 100 token window for the ring task. The labelled context length corresponds to the maximum context length of this window. This window had to be large to allow every token's activation to be collected.

2D projection of activations We use principcal component analysis (PCA) to find the dimensions ordered by their variances.

Appendix C. Additional Result

C.1. Additional Results on the grid

Fig. 4 shows the result in Fig. 1 over different layers.

Fig. 5 shows the visit count at each board position indice. The red lines denote corner grid points, which are visited less. We suspect this as the reason for seeing collapsed corners in representation space.

C.2. Additional Results on Ring

We show the result over different layers for the ring task in Fig. 6

We show the rescaled accuracy (rule following) of the ring task in Fig. 7.

C.3. Additional Results on the Semantic Ring

Fig. 8 shows the semantic ring of days of week extracted from Llama-3.1-8B.

We show the result over different layers for the semantic ring task in Fig. 9.

The accuracies on the semantic ring task is in Fig. 10.

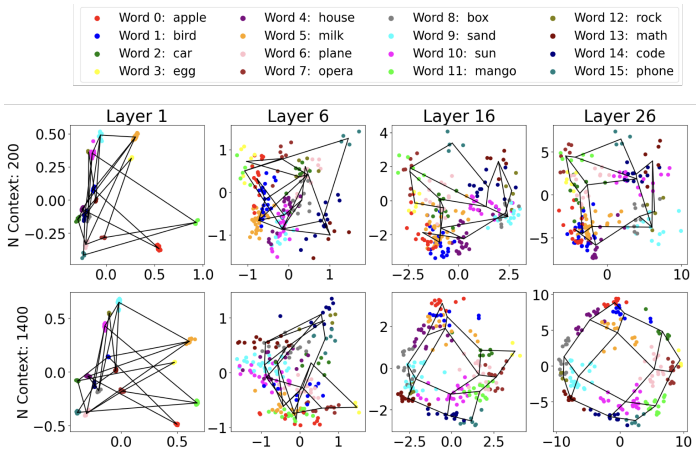

Figure 4: **Detailed depiction of PCA projections with increasing context size (rows) and deeper layers (columns).** As we increase the number of examples, and in deeper layers, we see a more distinct grid representation show up. Note that the legend indicates the ordering of the groundtruth grid – i.e., `apple` (red) is in the top left of the grid, with `house` (brown) and `bird` (blue) as its neighbors.

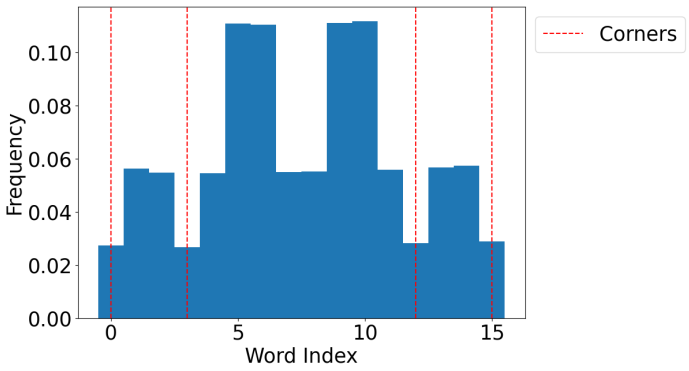

Figure 5: **Board position visit histogram**

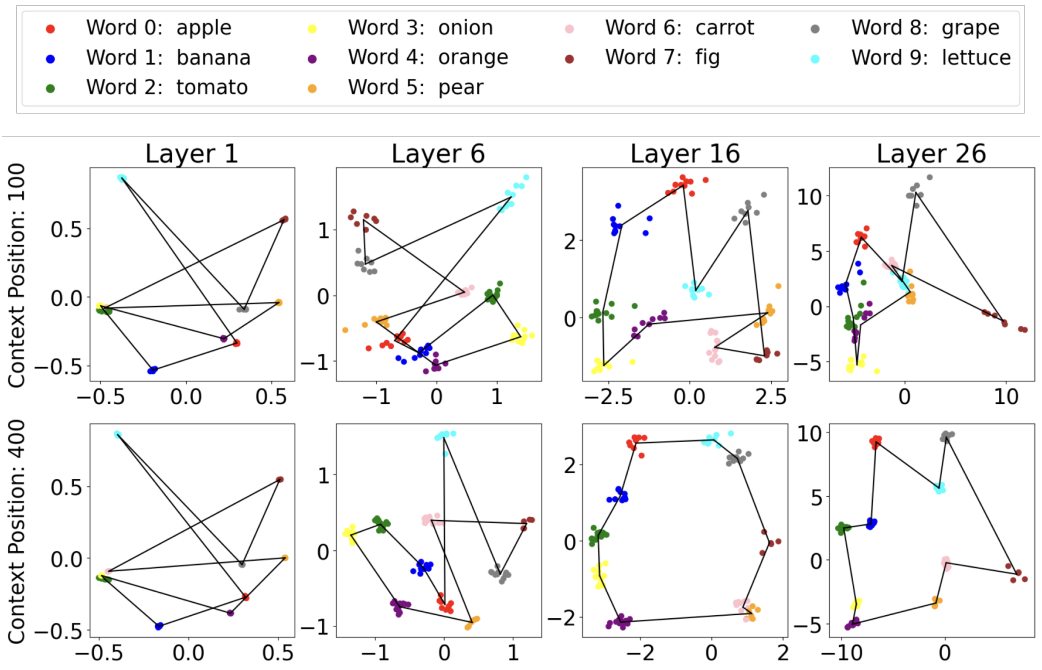

Figure 6: **Results on the Ring task at different layers**

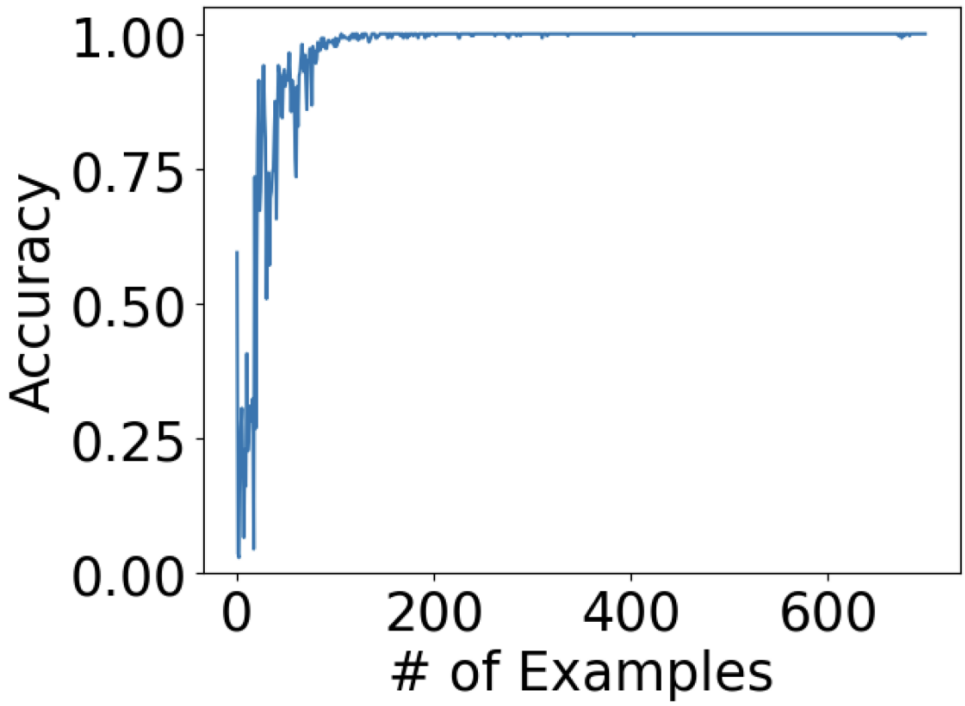

Figure 7: **Rescaled next token accuracy on a ring graph.** We plot the rescaled accuracy in all panels. The rescaled next token accuracy sums up the predicted probabilities of all valid output tokens defined in the DGP. Accuracy on a synthetically constructed ring of 10 words. After 100 tokens, the model learns almost perfectly to follow the rule and output tokens from neighbors on the ring

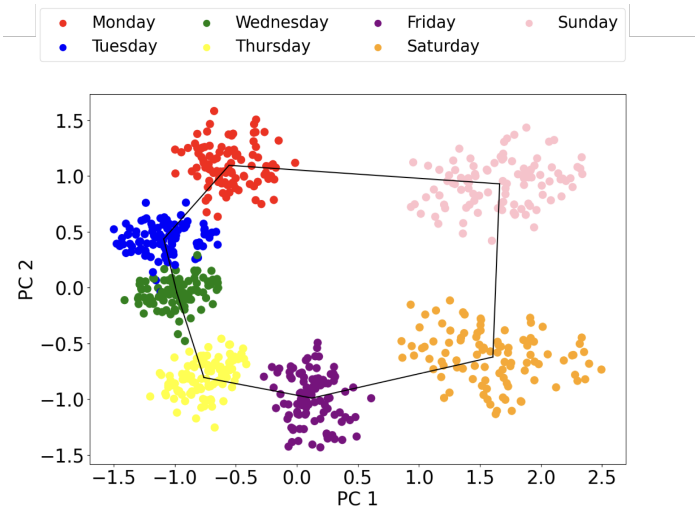

Figure 8: **Semantic Ring from Llama-3.1-8B** A ring of days of the week discovered in Llama3.1-8B(Dubey et al., 2024)'s representation space (layer 10 output), similar to Engels et al. (2024)

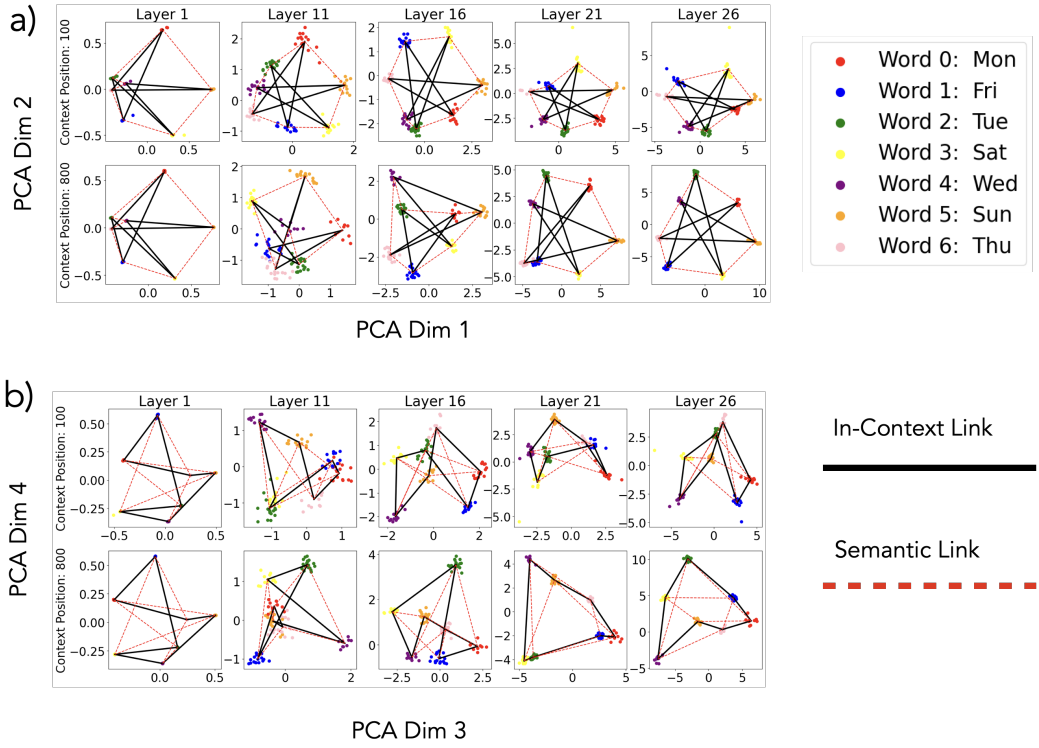

Figure 9: **Semantic ring override task results across multiple layers** (a) PCA dimension 1,2 (b) PCA dimension 3,4

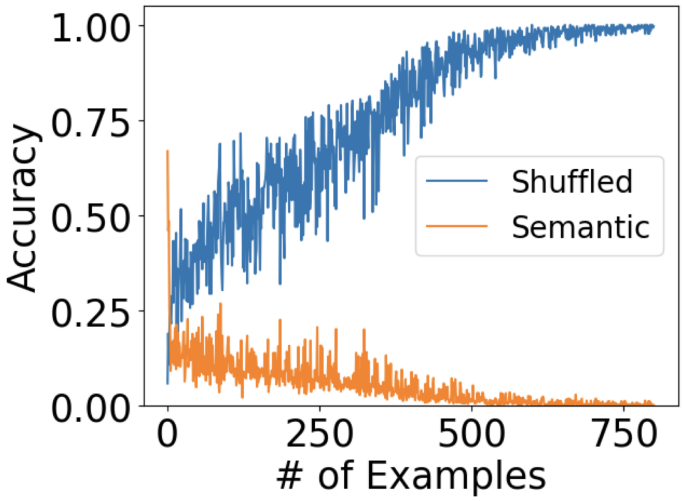

Figure 10: **Rescaled next token accuracy on a semantic ring graph.** We plot the rescaled next token accuracy on the semantic ring task. The rescaled next token accuracy sums up the predicted probabilities of all valid output tokens defined in the DGP. The semantic accuracy shows the next token accuracy on the semantic continuation, e.g. (" Tue" following " Mon"). The semantic accuracy quickly drops as the model figures that the given task is not a simple semantic continuation. The shuffled accuracy, evaluated on the task given in-context slowly rises and reaches nearly 100% give 500 exemplars.