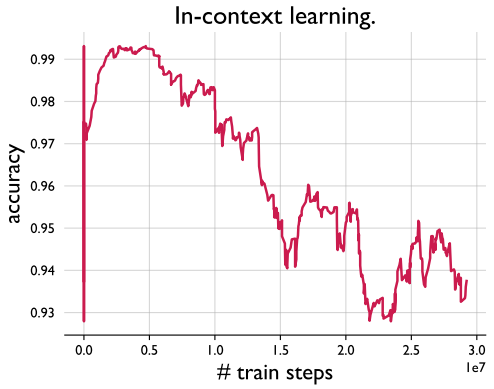

Figure 1: ICL is still transient on a dataset of fixed, rather than learned, embeddings. Omniglot images were fed through an Imagenet pre-trained Resnet50 encoder to obtain embeddings. This experiment was motivated by a concern that the learned Resnet encoder might be memorizing Omniglot images with extended training, making IWL the only possible solution (as exemplars within a class would not be similar, making finding matches impossible). We also believe these fixed embeddings to be closer to word embeddings used by language models (where classes in our setting could correspond to synonyms in a language setting). We find the continued transience of ICL a promising sign that this phenomenon is driven by transformer models (and not learnt resnet embeddings), and also evidence of generalization to other (albeit related) datasets. We have also started experiments with one-hot data with Gaussian noise, and will update the paper once those runs are complete.

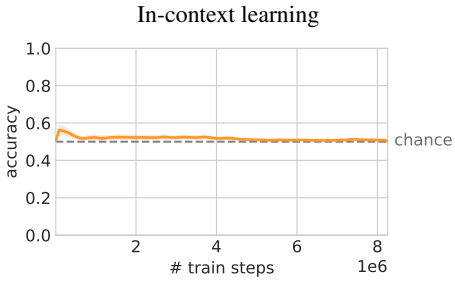

Figure 2: Experiments on longer context sequences – ICL is still transient. We believe longer context sequences may actually be less informative due to recency biases and/or more noise (due to more possible attention targets) in softmax attention. The context here consists of 16 exemplar-label pairs, 6 of which are from the query class. We also generally observed more variance in this setting across seeds (e.g., ICL only emerges in 2 out of 4 seeds. It is transient on both seeds where it emerges).

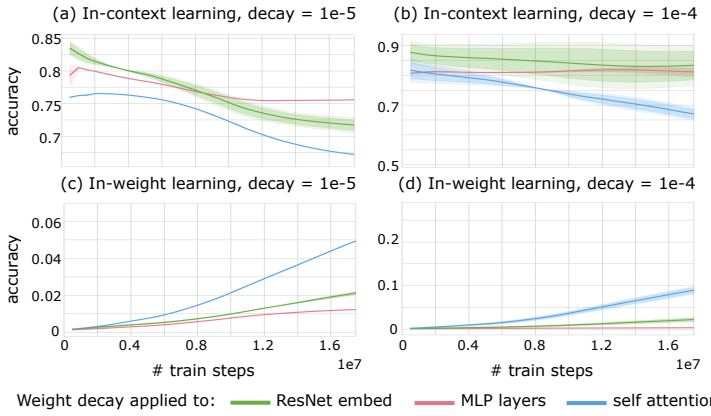

Figure 3: Selectively applying weight decay to different sets of network weights reveals that IWL relies on MLP layers. When we apply weight decay only to self-attention layers (blue), ICL is still transient. When we only apply weight decay to MLP layers, ICL transience is mitigated (red). These results can be interpreted in light of prior work indicating that in-weights information is stored in MLP layers (Geva et al 2021, Meng et al 2023). By selectively penalizing this behavior, we enable ICL to persist, thus providing convergent evidence that ICL normally fades (when no weight decay is applied) due to competition with IWL circuits.

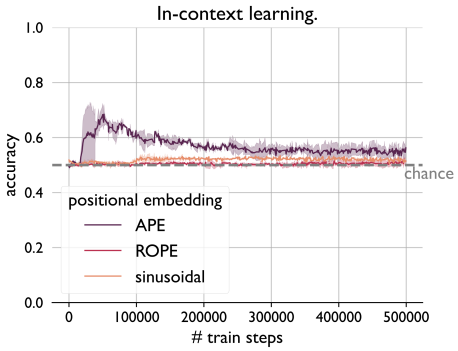

Figure 4: Experiments with alternate position embedding schemes, in a smaller, 2-layer model. APE here corresponds to learnt absolute positional encodings. We see that ICL does not emerge when using ROPE, though this is likely due to the tiny size of the model. In cases where ICL does emerge, we find it is transient – the main message of our work. Further explorations with position embedding schemes are an exciting avenue for future work.

Geva, M., Schuster, R., Berant, J., Levy, O. (2021). Transformer Feed-Forward Layers Are Key-Value Memories. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 5484–5495.

Meng, K., Bau, D., Andonian, A., Belinkov, Y. (2023). Locating and Editing Factual Associations in GPT, arXiv:2202.05262.