

A APPENDIX

A.1 SUPPLEMENTARY FIGURES

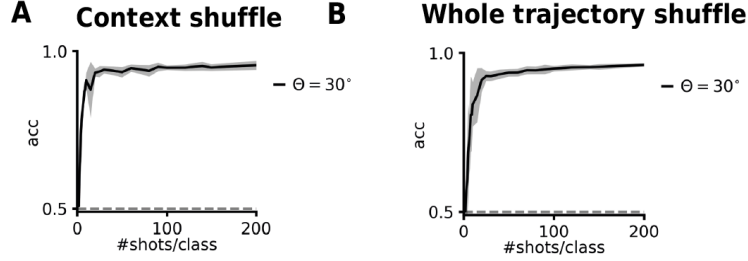


Figure 5: **Influence of the ordering of examples in-context** **A**: Llama3-8B on the $\theta = 30^\circ$ task shown in Fig. 1, but the examples shown in context are shuffled (from the same pool of training examples from a fixed trajectory). **B**: Same as A, but the entire trajectory is changed across simulations (different training examples altogether, not only their ordering in-context).

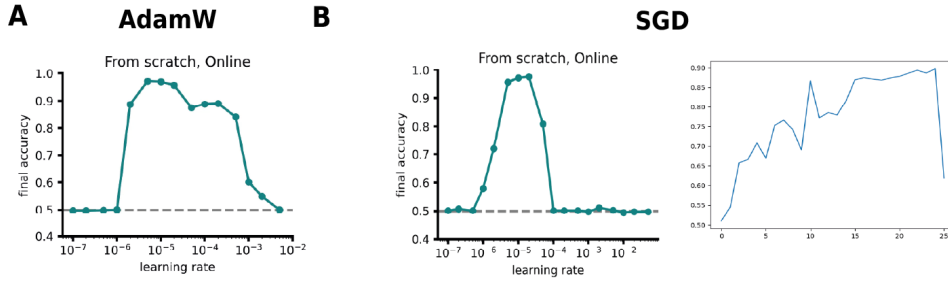


Figure 6: **Hyperparameter choices for SFT** **A**: Sweep of the learning rate (final held-out accuracy) for AdamW + cosine schedule. **B**: Left, same as A, but for vanilla SGD with constant learning rate. Right: example of an unstable SFT training with SGD: held-out accuracy as a function of the number of shots per class.

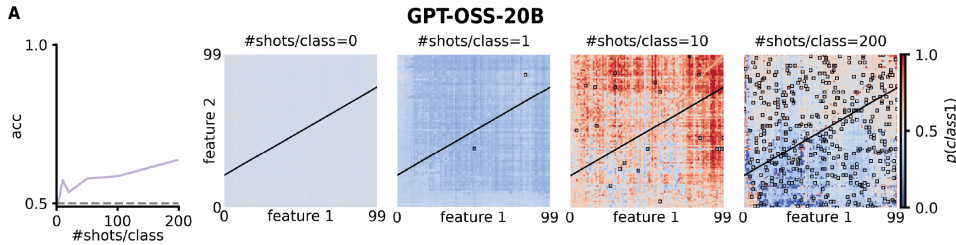


Figure 7: **Additional model comparison on an identical ICL trajectory**. **A**: Gpt-oss-20B on the $\theta = 30^\circ$ task shown in Fig. 1&4. The plots obey the same structure as in Fig. 4A-D.

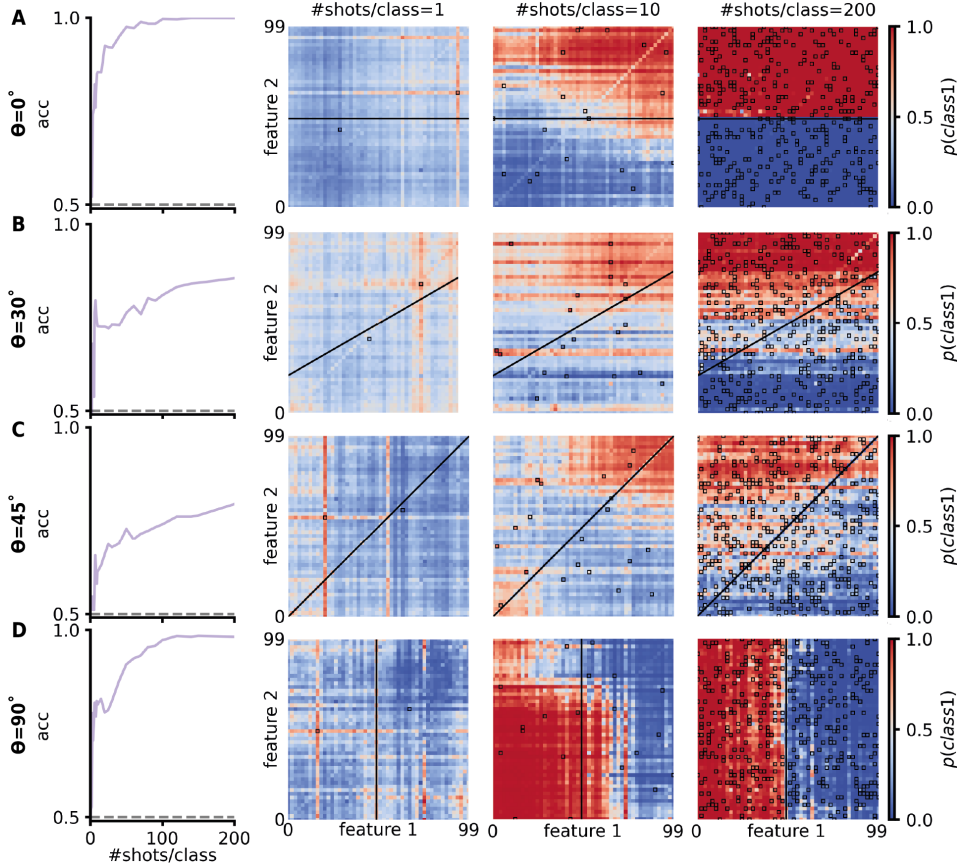


Figure 8: **Semantic 2-D linear classification.** A-D: From left to right: accuracy computed on all 10,000 possible inputs for the task as a function of the number of shots per class; visualization of the decision boundary of the model for increasing number of shots; probability associated with the logit of class 1 for all possible task inputs (same as in Fig. 1B). The probabilities are normalized for decision making such that $p(\text{class } 1) + p(\text{class } 2) = 1$.

A.2 SEMANTICALLY UNRELATED LABELS

The typically used "Foo" and "Bar" labels for semantically unrelated ICL already have a few years of existence in the literature (Min et al., 2022; Wei et al., 2023), so it can be assumed that they are part of the pre-training set somehow. Moreover, "Bar"—unlike "Foo"—is an english word which creates a bias towards "bar", at least in the few shots regime. Since our work studies the unfolding the learning dynamics of ICL, we chose two other sequences of letters with single token representations for most open-source tokenizers (with the exception of mistral) and more balanced default priors.

A.3 SFT TRAINING DETAILS

After a parameter sweep, we used AdamW with learning rate $1e-5$, 100 epochs, and a cosine learning rate schedule (warmup ratio 0.05 and final learning rate $1e-7$). We initially favoured a simpler learning rule (vanilla SGD), but this led to some optimizations unexpectedly blowing up and making results inconsistent (Supp. Fig. 6).

A.4 SEMANTIC VERSION OF THE 2-D LINEAR CLASSIFICATION TASK

We use a list of 50 adjectives of increasing valence as a replacement for integers in an otherwise identical classification task with the same single token output tokens described above.

A.5 ADDITIONAL MODELS

The poor results on gpt-oss-20b (OpenAI, 2025) are likely due to the fact that this model explicitly expects OpenAI’s harmony format for chat (i.e. is not a purely base model).