H.3 Logit lens between weights learned by RePS and language modeling objectives

Figure 18 shows the logit lens [Nostalgebraist, 2020] results for the tokens ranked highest or lowest by the lens. Our results suggest that SVs trained with RePS and those trained with a language modeling objective yield similar logit lens behavior.

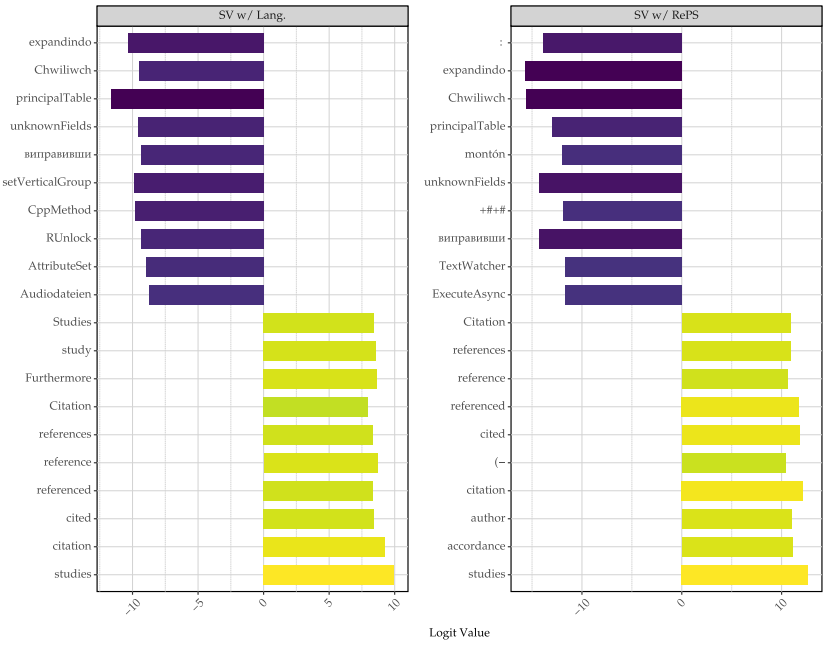

Figure 18: Logits lens rankings of output tokens with methods trained on `Gemma-2-2B` L20.

H.4 *Concept detection* with preference-based vectors

Figure 19 shows the average area under the ROC curve (AUROC) for each method across all concepts using steering vectors trained on `Gemma-2` models. Our results suggest that steering vectors trained with the language modeling loss are better at detecting concepts in the inputs. This validates our hypothesis that the language modeling loss yields better directions for detecting the low-level semantics encoded in embeddings.

I AxBench analyses

AxBench provides a training dataset, Concept500, in which each subset contains 500 steering concepts collected from three distinct domains: *text*, *code*, and *math*. As shown in fig. 20, steering scores across these three genres differ significantly. We hypothesize that this is because AxBench samples instructions from public datasets based on genre. For instance, math-related instructions are drawn from math datasets such as `GSM8K` for training, whereas evaluation instructions come from `Alpaca-Eval`. This discrepancy could lead to an out-of-distribution generalization problem for methods requiring training, while training-free baselines such as prompting are more robust.

To validate our hypothesis and further strengthen the performance of our intervention-based methods on `Gemma-3` models, we augment the training datasets such that their instructions are sampled from the original instruction pool for *text* genre. Figure 21 shows the score distributions across genres after using our augmented training data.

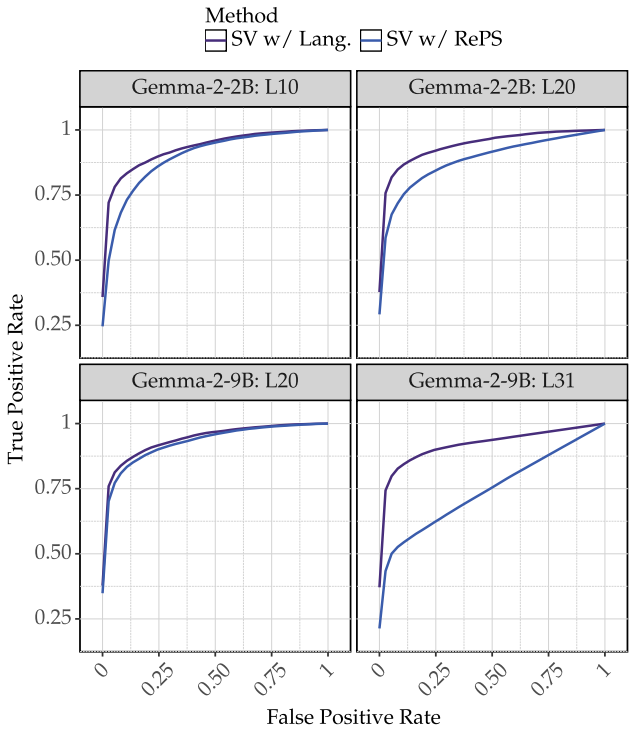

Figure 19: Mean ROC curves over all concepts with steering vectors trained on `Gemma-2` models.

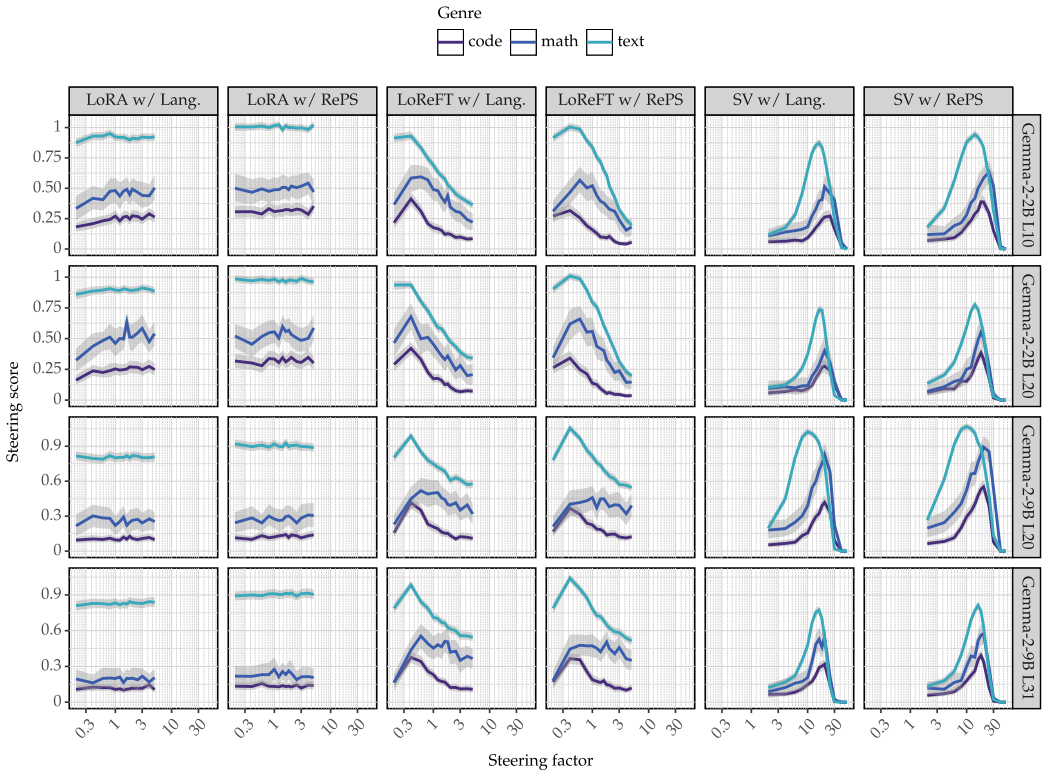

Figure 20: Steering factor vs. scores for concepts with different genres with the training data from AXBENCH.

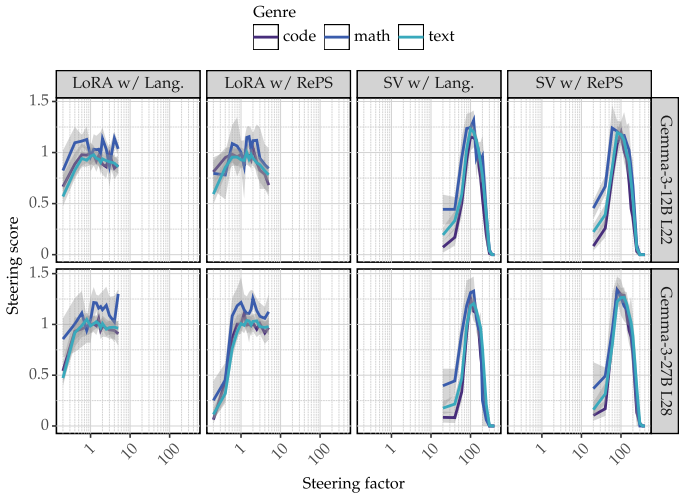

Figure 21: Steering factor vs. scores for concepts with different genres with new training data created for `Gemma-3` models without genre-based instruction sampling procedure.