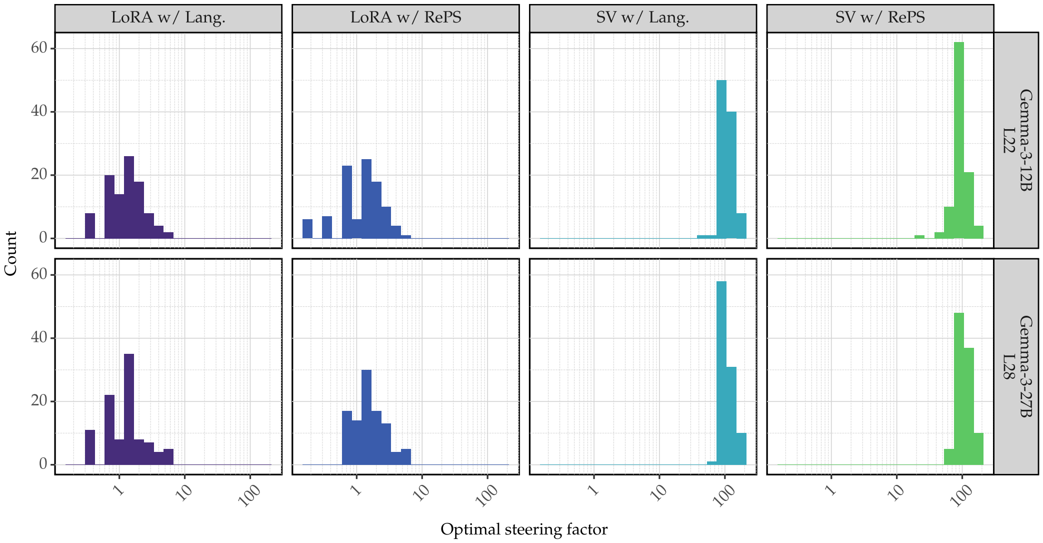

Figure 7: Mean score breakdown for all methods on our unseen testing instruction set after selecting the optimal factor (based on the Overall Score) on our evaluation instruction set for `Gemma-3` models.

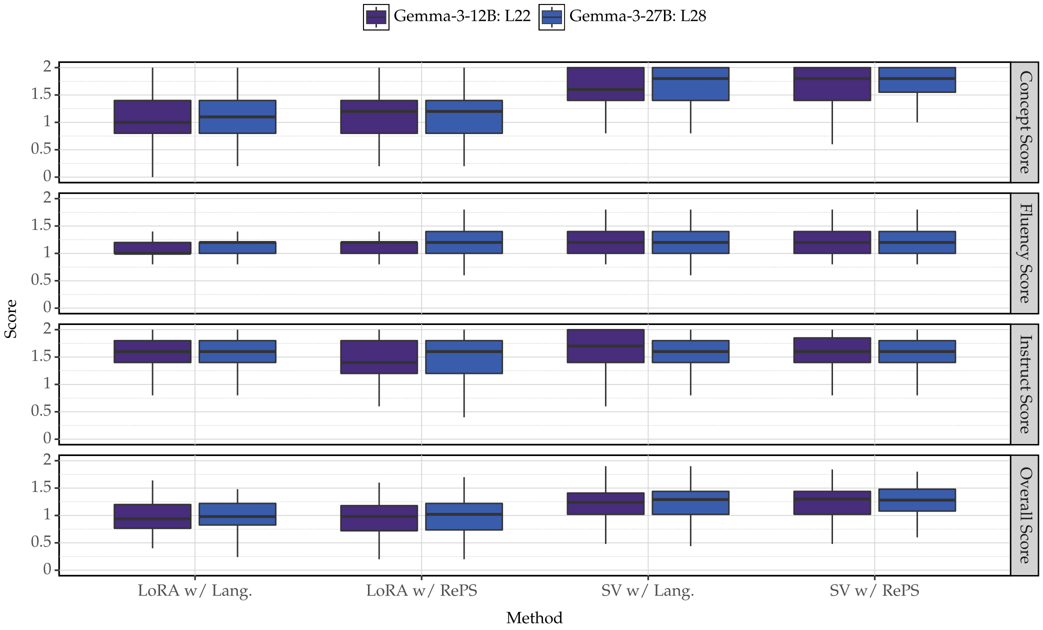

Figure 8: Distribution of optimal steering factors for each intervention-based methods (LoRA, ReFT and SV) with two objectives (Lang. and RePS) across the 4 tasks with `Gemma-3` models.

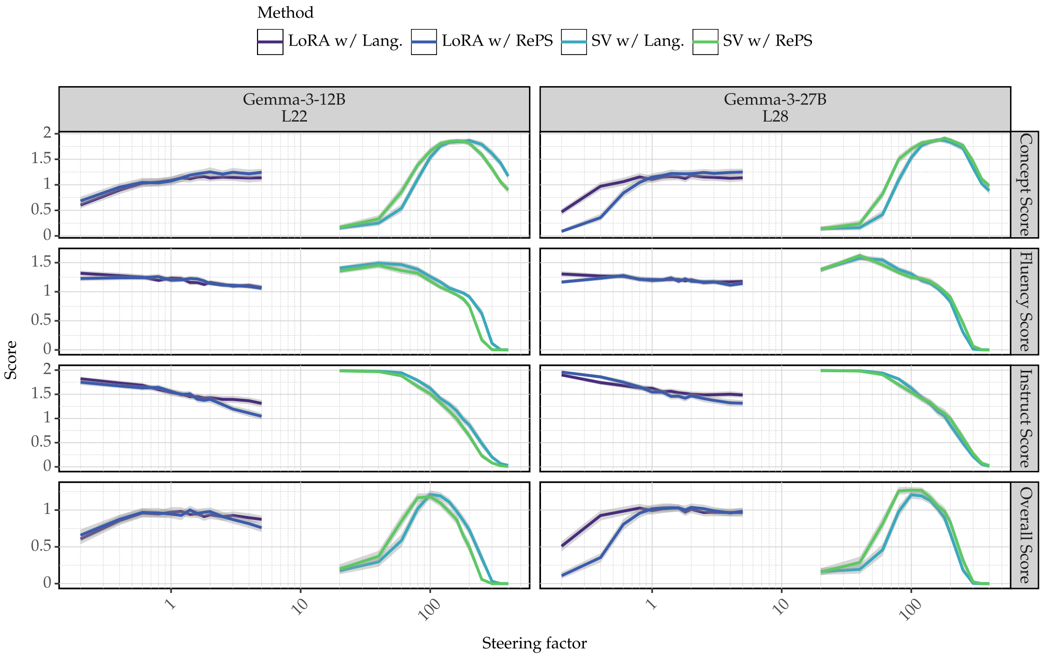

Figure 9: Steering factor vs. scores for `Gemma-3` models.

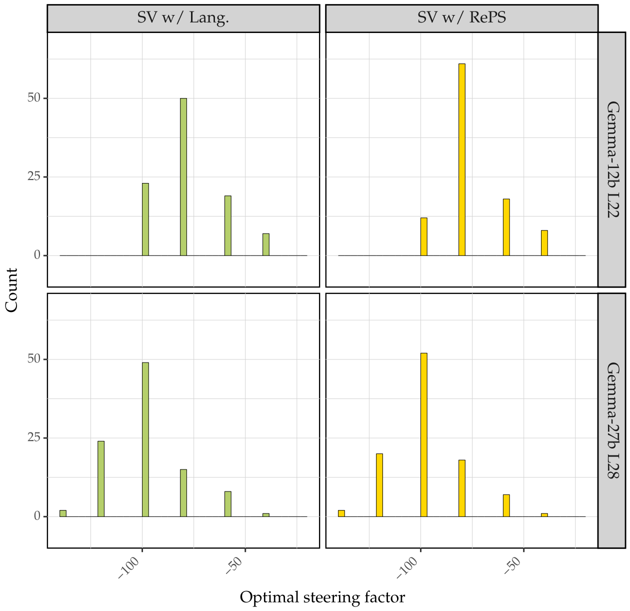

Figure 10: Suppression factor vs. scores for `Gemma-3` models.

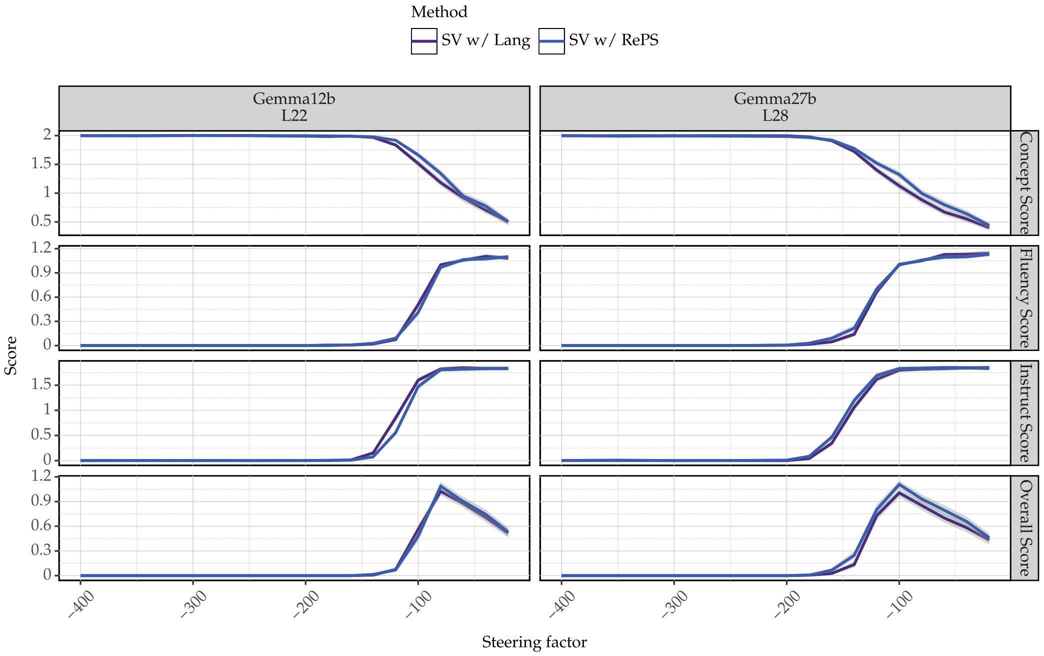

Figure 11: Suppression Mean score breakdown for all methods on our unseen testing instruction set after selecting the optimal factor (based on the Overall Score) on our evaluation instruction set for `Gemma-3` models.