A Detailed analysis

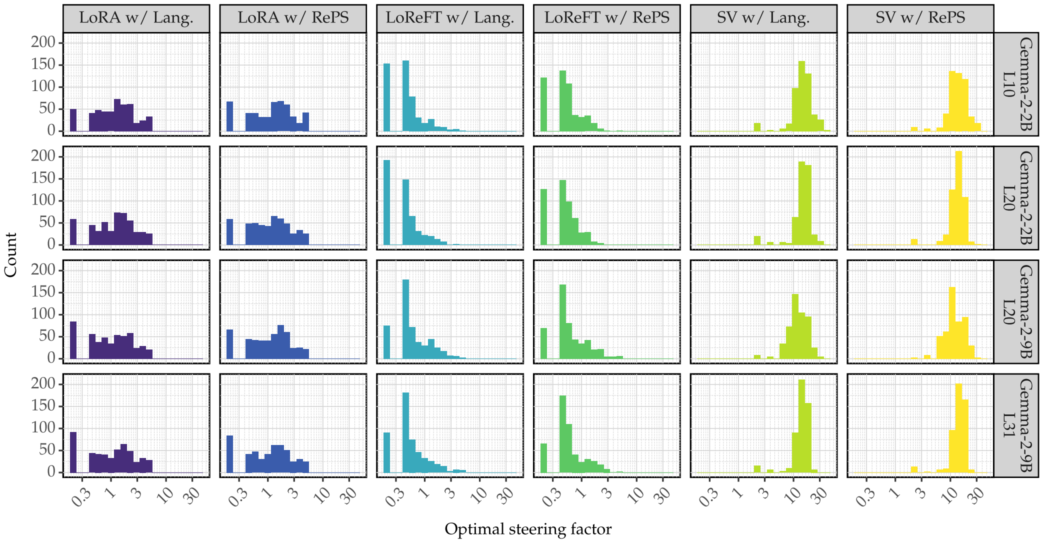

Figure 2: Mean score breakdown for all methods on our unseen testing instruction set after selecting the optimal factor (based on the Overall Score) on our evaluation instruction set for `Gemma-2` models.

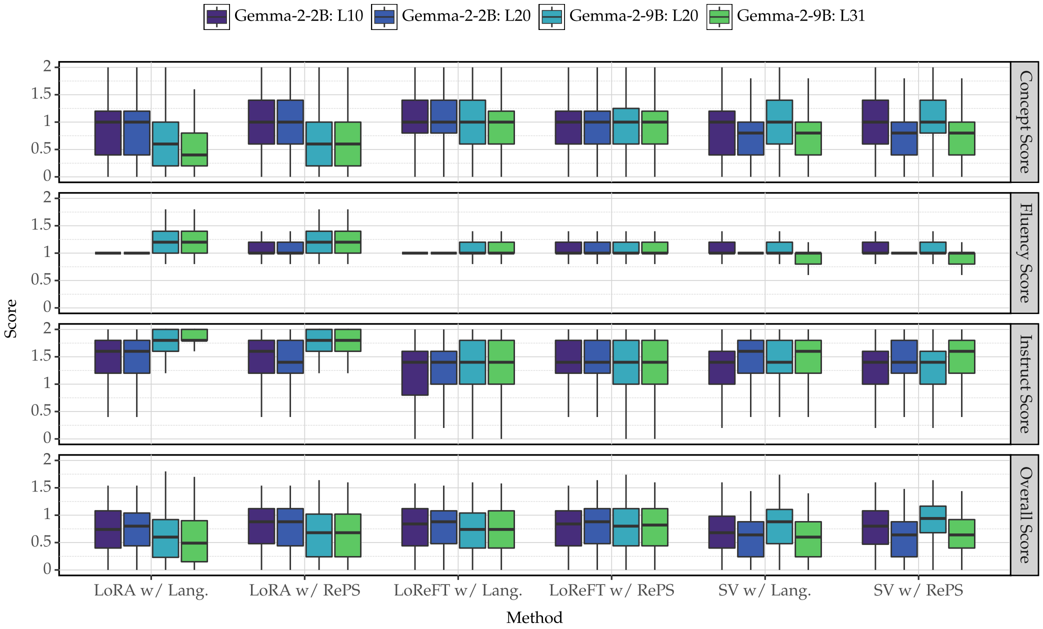

Figure 3: Distribution of optimal steering factors for each intervention-based methods (LoRA, ReFT and SV) with two objectives (Lang. and RePS) across the 4 tasks with `Gemma-2` models.

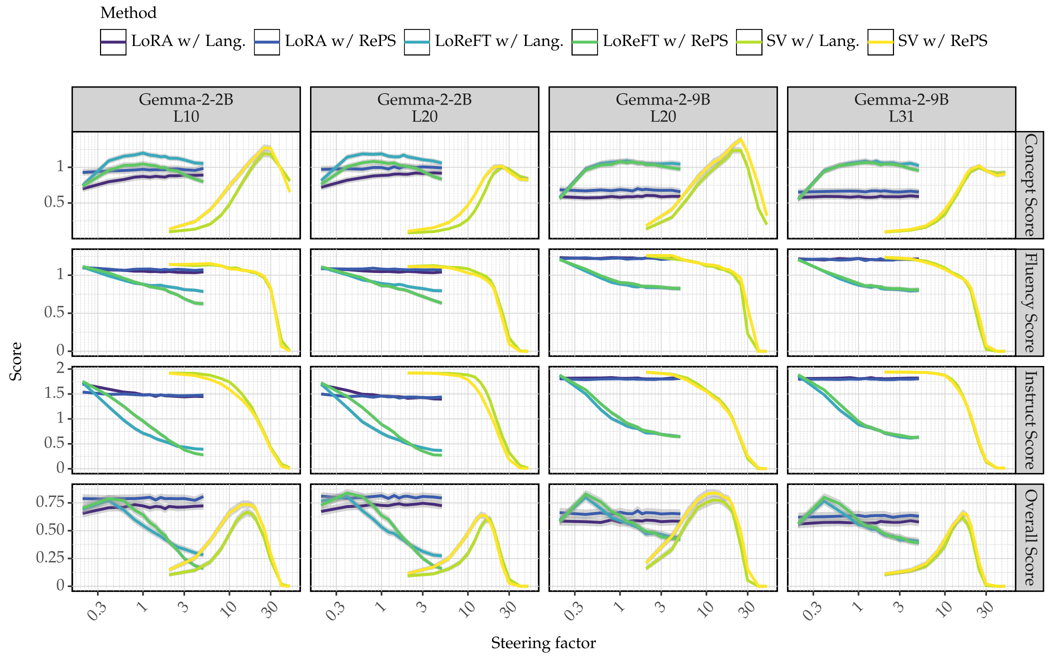

Figure 4: Steering factor vs. scores for `Gemma-2` models.

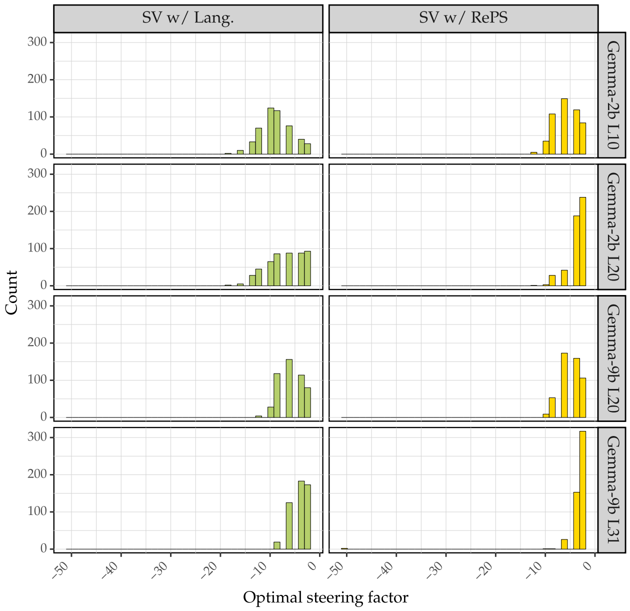

Figure 5: Distribution of optimal suppression factors for each intervention-based methods (LoRA, ReFT and SV) with two objectives (Lang. and RePS) across the 4 tasks with `Gemma-2` models.

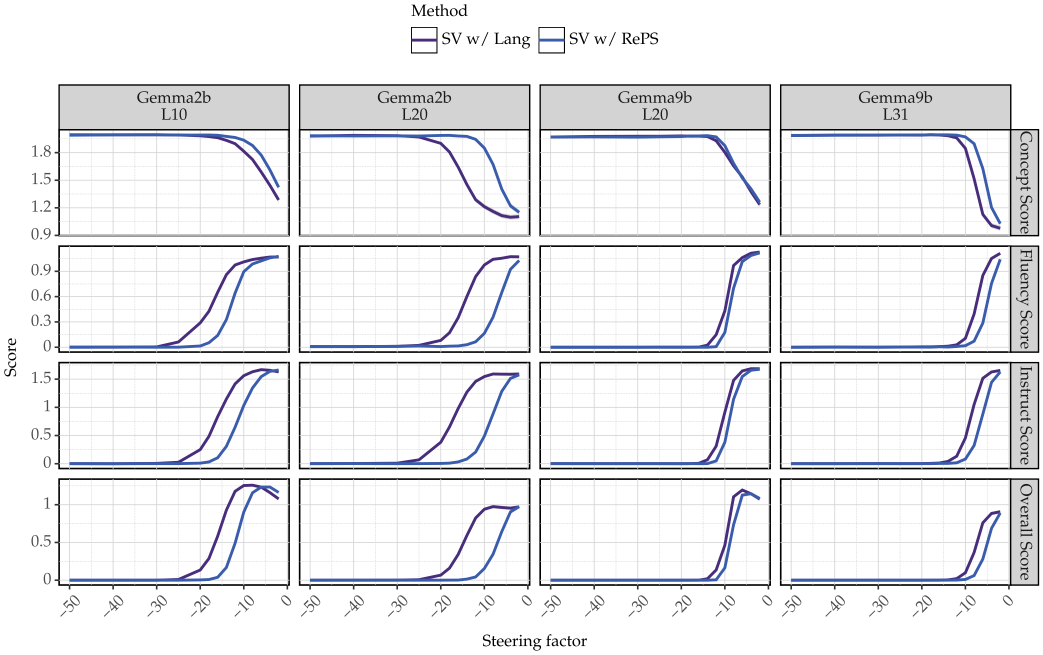

Figure 6: Suppression factor vs. scores for Gemma-2 models.