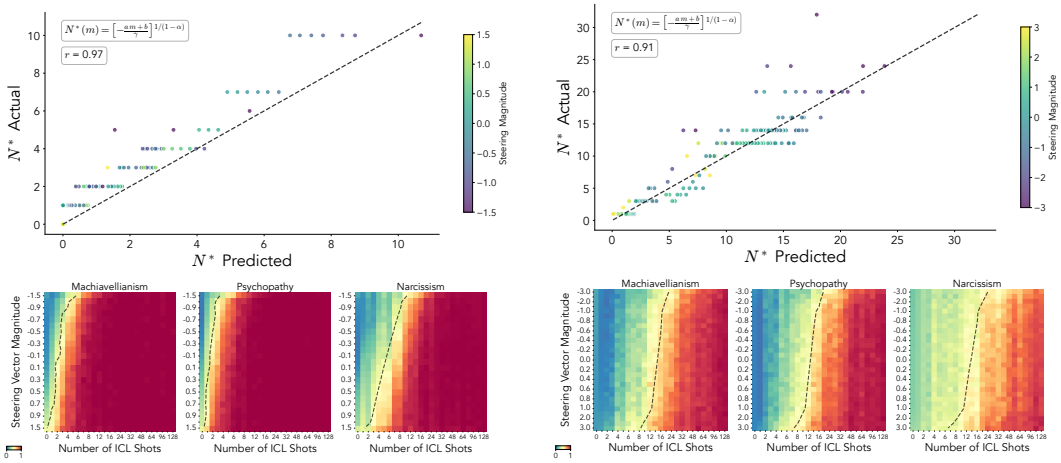


(a) Gemma-2-9B

(b) Qwen-2.5-7B

Figure 11: The belief dynamics model captures cross-over points N^* across different language models.

C Full Steering Range

The results discussed in our main text focus on the case where the Linear Representation Hypothesis (LRH) holds. However, we empirically find that with larger enough steering magnitudes, the linear effect of steering on $\log o(c|x)$ begins to break down and the sigmoidal response function we show in Fig. 5 converges towards 0 (Fig. 12 and Fig. 13). This is similar to the findings of Panickssery et al. (2024) which shows that LLM behavior begins to break down and become incoherent with very large magnitude steering vectors. We find that behavior converges towards chance ($p(y|x) = 0.5$), even with very large context lengths.

Different datasets have different thresholds for m which cause behavior to break down (Fig. 13). For Llama-3.1-8b, this magnitude threshold is larger for Narcissism than other datasets. As shown in Fig. 4, Narcissism has less effect from steering with small magnitudes compared to the other 4 datasets, and also has a later transition point N^*. These results may be together explained by Narcissism having a weaker signal for the target concept c through the likelihood $p(x|c)$, which results in both in-context learning and steering having comparatively less impact on belief compared to datasets with a stronger signal.

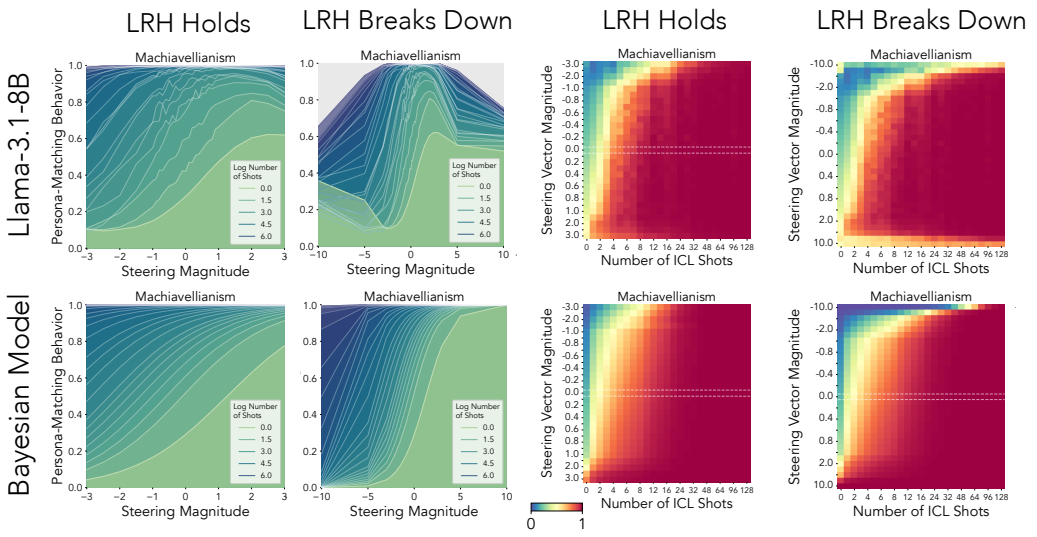
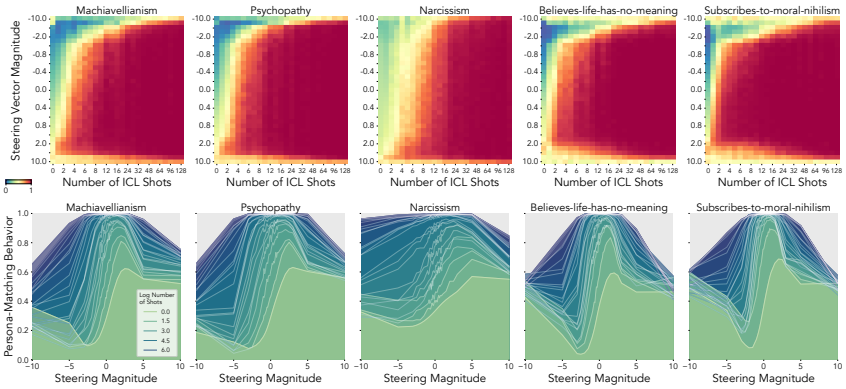
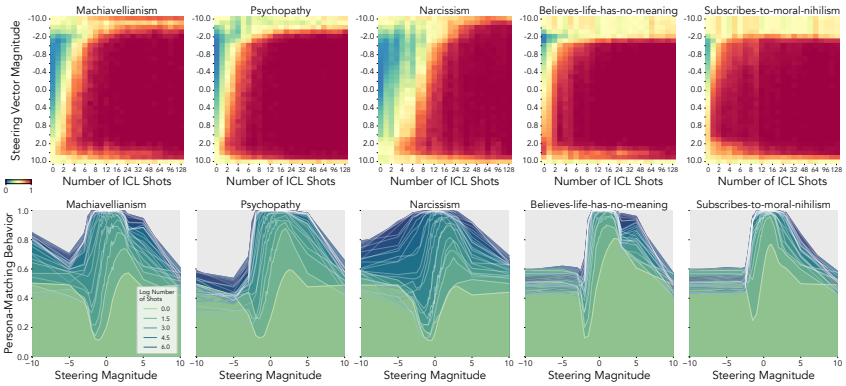


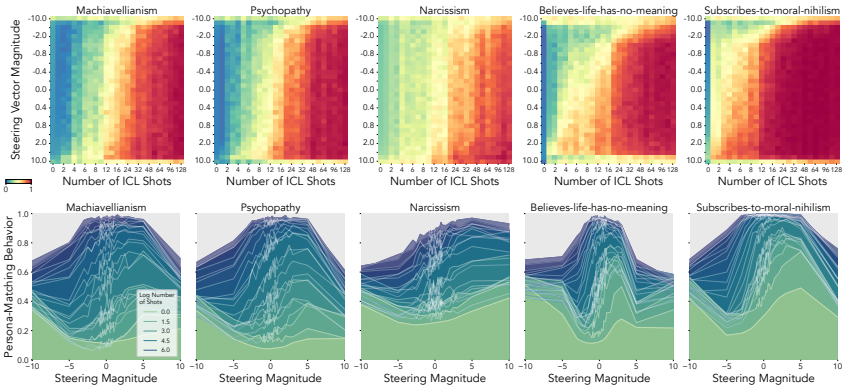
Figure 12: **With large enough magnitudes, the Linear Representation Hypothesis breaks down** Our belief dynamics model is able to explain model behavior within a limited range of m. When steering magnitudes exceed this range, behavior begins to break down and converges to chance ($p(y|x) = 0.5$).



(a) Llama-3.1-8B



(b) Gemma-2-9B



(c) Qwen-2.5-7B

Figure 13: **Different datasets have different thresholds for steering breaking down.** Different datasets have different thresholds for what steering magnitudes m will predictably steer model behavior.