Next, we define a new term $a = \frac{1}{2}\|d_i\|^2$ and, using our previous theorems, substitute as follows:

$$\begin{aligned}
-\eta &= w_i^T v + b \\
&= d_i^T v + b && \text{Substitute } w \approx d_i \\
&= d_i^T \left(\sum_j \beta_j(v) d_j \right) + b && \text{L.R.H definition} \\
&= \beta_i(v)\, \|d_i\|^2 + b && d_i^T d_j \approx 0,\ \forall i \neq j \\
&= a\, \beta_i(v) + b && \text{Definition of } a
\end{aligned}$$

Finally, we define the log posterior odds when steering v by $m \cdot d_i$ as:

$$\log \frac{p(c_i \mid v + m \cdot d_i)}{p(c_i' \mid v + m \cdot d_i)} = -\eta_{\text{steered}}$$

Steering changes $v \to v + m \cdot d_i$, and thus

$$\begin{aligned}
\eta_{\text{steered}} &= d_i^T (v + m \cdot d_i) + b && \text{Substituting } v \text{ in } \eta \\
&= d_i^T v + m\|d_i\|^2 + b \\
&= d_i^T v + a \cdot m && \text{Definition of } a \\
&= \eta + a \cdot m && \text{Definition of } \eta
\end{aligned}$$

Finally, we obtain

$$\log \frac{p(c_i \mid v + m \cdot d_i)}{p(c_i' \mid v + m \cdot d_i)} = \log \frac{p(c_i \mid v)}{p(c_i' \mid v)} + a \cdot m$$

B Main Results Across Models

We find that our account remains highly predictive across three models, with average correlations of $r = 0.98$ for Qwen-2.5-7B and $r = 0.97$ for Gemma-2-9B computed across the entire heatmap (Fig 8), and correlations of $r = 0.91$ for Qwen-2.5-7B and $r = 0.97$ for Gemma-2-9B for prediction of N^* (the phase boundary; Fig. 11). Note also that all correlation values are computed for held-out predictions. Furthermore, we find that our predictions regarding the influence of in-context learning and steering are corroborated (Fig. 9, Fig. 10).

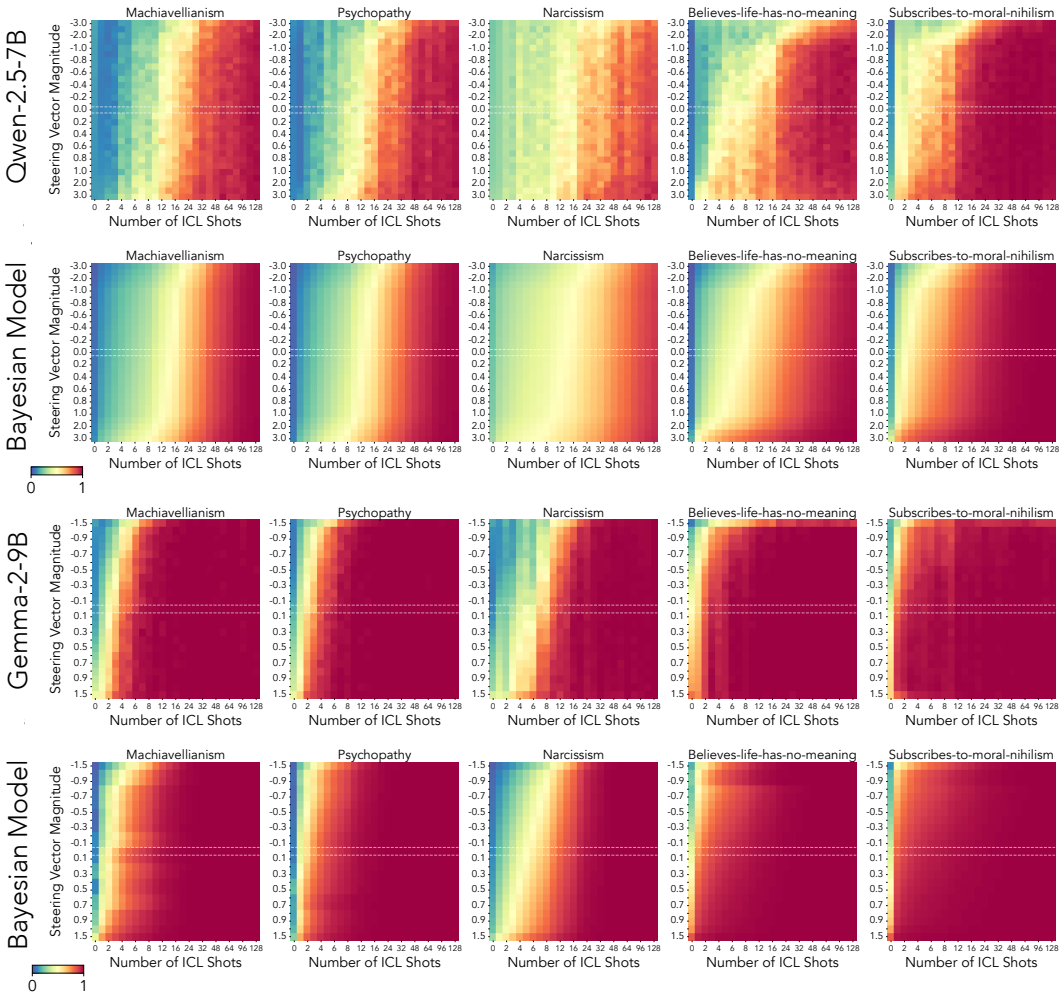

Figure 8: **In-context learning and activation steering jointly affect behavior**. Results presented in Fig. 6 replicate across Qwen-2.5-7B and Gemma-2-9B models, showing the generalizability of the belief dynamics model.

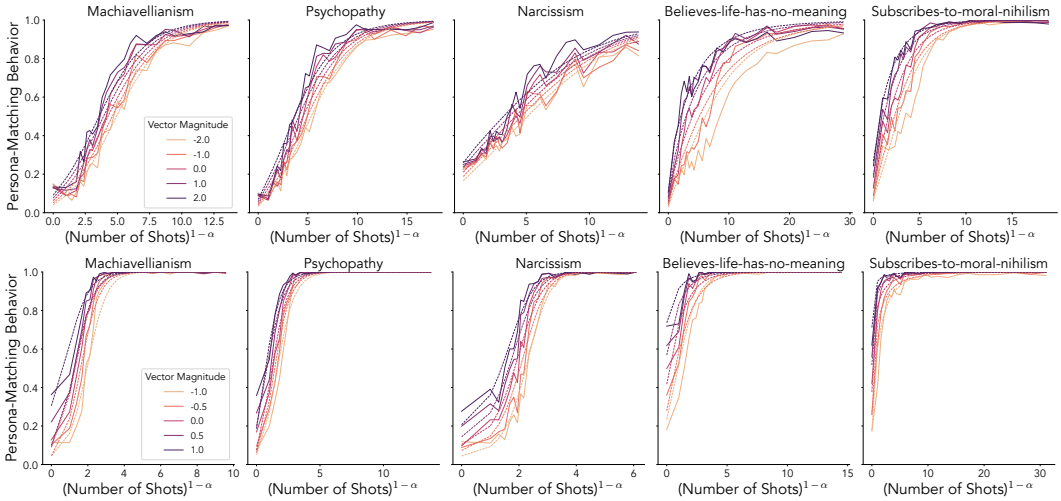

Figure 9: In-context learning curves in Qwen-2.5-7B (top) and Gemma-2-9B (bottom).

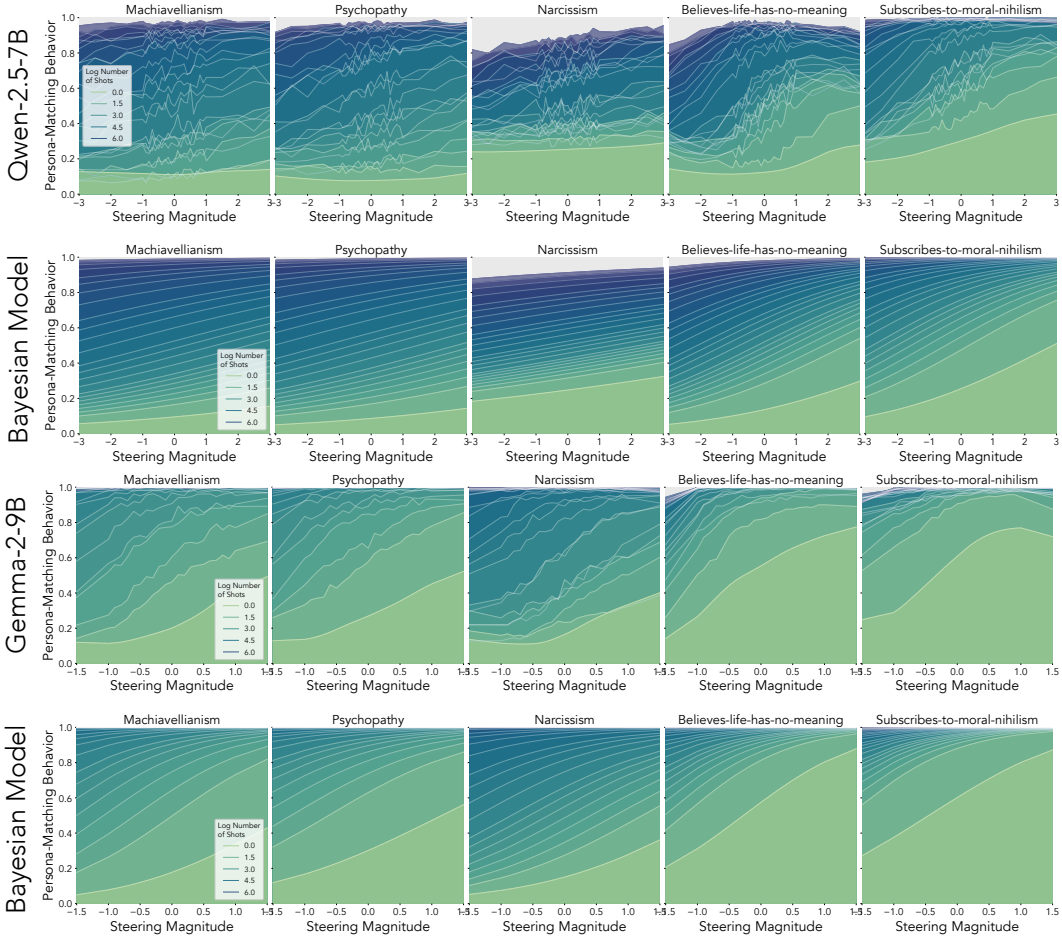

Figure 10: Steering magnitude response function in Qwen-2.5-7B and Gemma-2-9B.