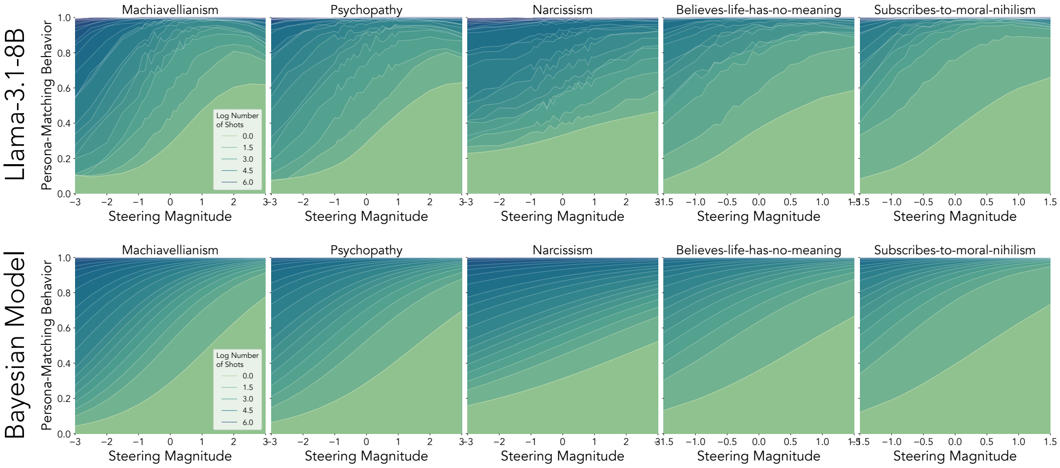

Figure 5: **Change in behavior as a function of steering vector magnitude** As we scale steering vector magnitude (x-axis), we find a sigmoidal response function in behavior (y-axis). With steering magnitudes in the range $[-1, 1]$, we find approximately linear effects of steering, which taper off as magnitude increases. This pattern holds across different numbers of ICL examples (different colors). This pattern is well-captured by our model, which assumes a linear impact of steering on the log prior odds, and hence a sigmoidal impact in probability space.

et al., 2022). Specifically, LRH states that neural network representations encode semantically meaningful concepts in hidden representations in a "linear" manner (Elhage et al., 2022; Arora et al., 2016). Here, "linearity" refers to three related phenomena: (i) concepts are linearly accessible from model representations, e.g., via simple logistic probes (Belinkov, 2022; Tenney et al., 2019); (ii) linear algebraic manipulations of hidden representations along certain directions can steer model outputs (Panickssery et al., 2024; Turner et al., 2024); and (iii) representations are defined as an additive mixture of these directions (Bricken et al., 2023; Templeton et al., 2024). One can unify these notions within a single formal computational model as follows.

$$v = \sum_i \beta_i(v) d_i \quad \text{s.t.} \quad d_i^\intercal d_j \sim 0 \;\forall\; i, j, \tag{5}$$

where $v \in \mathbb{R}^n$ is a hidden representation corresponding to input x, $d_i \in \mathbb{R}^n$ represents some concept c_i, and $\beta_i(v) \in \mathbb{R}$ is a scalar denoting the extent to which d_i is present in v.

LRH argues that if a concept is linearly represented (in the sense described above), then a logistic classifier $\sigma\left(-w^\intercal v - b\right)$ suffices to infer the extent to which concept c is present in the representation v. Since we assume minimal interference between directions that reflect different concepts, a well-trained classifier will have weights in line with d_i (assuming it captures the concept we are interested in). Combining these assumptions, we get the following:

$$p(c_i|x) = p(c_i|v) = \sigma\left(-w^\intercal v - b\right) = \sigma\Bigg(- \beta_i \|d_i\|^2 - \sum_{j \neq i} \beta_j d_i^T d_j - b \Bigg) \approx \sigma\left(-\beta_i(v) a - b\right), \tag{6}$$

where $a = \|d_i\|^2$. Using this, we can express the posterior odds as follows: $\log \frac{p(c_i|v)}{p(c_i'|v)} = \log \frac{p(c_i|v)}{1 - p(c_i|v)} = a\beta_i(v) + b$. Thus, if one steers the model representation along direction d_i, e.g., changing v to $v + m \cdot d_i$[2], the model's *belief* in concept c_i will linearly increase (in log-space) to $a\beta_i(v) + a \cdot m + b$. Relating this back to the unsteered model's log-posterior odds, we get the following:

$$\log \frac{p(c_i|v + m \cdot d_i)}{p(c_i'|v + m \cdot d_i)} = \log \frac{p(v|c_i)}{p(v|c_i')} + \log \frac{p(c_i)}{p(c_i')} + a \cdot m = \log \frac{p(v|c_i)}{p(v|c_i')} + \log \frac{p'(c_i)}{p'(c_i')} \quad . \tag{7}$$

[2] Note that steering boundlessly (e.g., taking $m \to \infty$) will push the representations to a region that lies outside the support over which distribution $P(c|v)$ is defined. We see such effects empirically (e.g., see Fig. 12) and thus focus our discussion in the main paper in a range for m where posterior-belief changes monotonically.

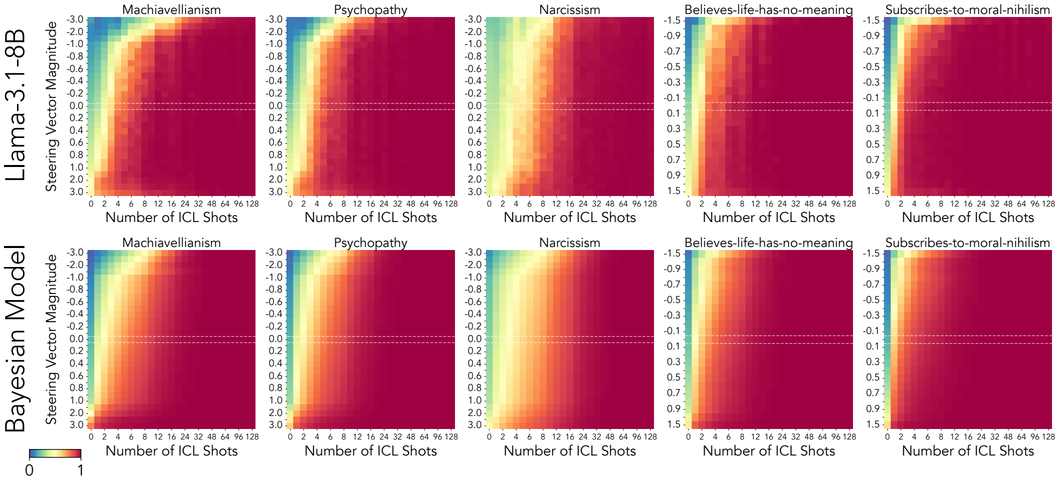

Figure 6: **In-context learning and activation steering jointly affect behavior** The in-context learning dynamics we observe in Fig. 4 and the steering vector magnitude response function in Fig. 5 interact to create a phase boundary (Top). Our belief dynamics model re-constructs this diagram with high fidelity (Bottom).

That is, steering yields a constant shift in the model log-posterior odds that will consistently change model beliefs for both an individual observation x or an entire population $X \sim P_x$. We therefore argue the effects of steering are best described as alteration of a model's prior beliefs in a concept c, updating the log prior odds from $\log \frac{p(c)}{p(c')}$ to $\log \frac{p'(c)}{p'(c')}$ (where $p'(c)$ is an unnormalized prior). Intuitively, this formalizes the claim that steering vectors should be expected to change behavior y regardless of the input x. For example, for the concept C_{happy} we should expect the steering vector $\hat{d}_c$ to make an LM behave more *happy* even if it is given inputs $x^{(c')}$ that are not *happy*, i.e., which have lower $p(c \mid x^{(c')})$.

Prediction 2 Assuming linear representation hypothesis holds, Eq. 7 shows steering will increase a model's belief in concept c_i at a sigmoidal rate with steering magnitude m.

Results We find that activation steering leads to a sigmoidal trend in persona-matching behavior (and thus a linear trend for the posterior odds) as a function of steering vector magnitude (Fig. 5). We observe this within the range $m \in [-3, 3]$ for the Dark Triad datasets, and the range $m \in [-1.5, 1.5]$ for Moral Nihilism datasets for Llama-3.1-8B. This trend holds across various context lengths, although with large contexts, the behavior is near ceiling for all magnitudes.

4.3 Final model

From Eq. 7 , the log posterior odds given an intervention on v can be defined as $\log o(c|x) = \log \frac{p(c)}{p(c')} + \log \frac{p(x|c)}{p(x|c')} + a \cdot m$, where $o(v|c) = o(x|c)$ since we assume $p(c|x) = p(c|v)$ in Sec. 4.2. Next, we substitute the log prior odds $\log \frac{p(c)}{p(c')}$ with a constant offset b, since it does not depend on the precise input x or its representation v. This gives us our final model of belief update dynamics in ICL:

$$\boxed{\log o(c|x) = a \cdot m + b + \gamma N^{1-\alpha}} \tag{8}$$

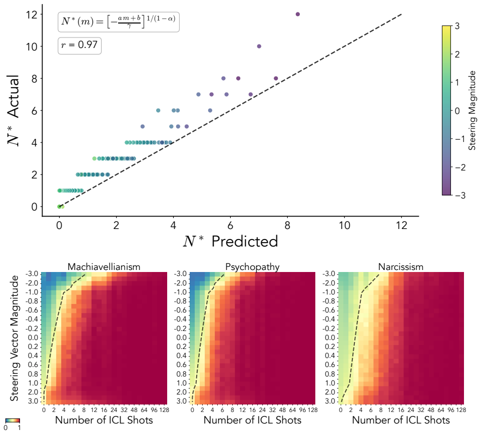

Figure 7: **Predicting the amount of context needed to enable a persona.** (Top) The belief dynamics model is highly predictive of crossover points N^* between c and c', with a correlation of $r = 0.97$, and (Bottom) our N^* estimates effectively predict phase boundaries in empirical behavioral data.

This model describes how model behavior changes as a function of both context length N and steering magnitude m. Concretely, for the model prediction results described in this work, we fit scalar parameters to a, b, γ, α to empirical averages of model behavior $p(y|x) = p(c|x)$ (Eq. 4) using L-BFGS, across various contexts x (where $N = |x|$) and steering with various magnitudes m. We do so for practical reasons: first, the steering vectors we use in practice are not the true concept vectors d_i, and so the effect of steering magnitude will be $a \propto \|d_i\|^2$ (but not necessarily $a = \|d_i\|^2$), and second, we estimate the prior odds b rather than observing the concept priors $p(c), p(c')$. This model allows us to compute the transition points in context length for a given steering magnitude m when the model's belief in concept c surpasses c', i.e., when $\log o(c|x) = 0$:

$$N^*(m) = \left[-\frac{a\,m + b}{\gamma} \right]^{1/(1-\alpha)}. \tag{9}$$

Prediction 3 Log posterior odds will be additively impacted by varying in-context examples and steering magnitude, and this interaction will yield distinct phases dominated by belief in either c or c'. The boundary between phases—the cross-over point N^* when belief in concept c surpasses belief in c'—can be predicted as a function of initial log prior odds and steering magnitude (Eq. 9).

Results Observing the phase diagrams in Fig. 6, we find that our model is highly predictive of the joint effects of in-context learning and steering. Moreover, following our definition of $N^*(m)$ in Eq. 9, we can predict the crossover points when behavior will transition to be dominated by c (Fig. 7).

5 Discussion

In this work, we present a novel synthesis of prior theoretical and empirical work in two disparate approaches to language model control: in-context learning and activation steering. We find a phase boundary across ICL and activation steering, where the transition point is jointly modulated by context and activations. Further, we present a Bayesian belief dynamics model that formalizes this theory and accurately predicts language model behavior as a function of both context length and steering vector magnitude. Our approach builds on top-down theories of behavior from the perspective of Bayesian belief updating, as well as bottom-up theories of learning and representation in connectionist neural networks. This paves the way for future work to bridge levels of analysis for describing behaviors, the algorithms driving behavior, and the mechanisms that implement those algorithms (Marr, 1982; He et al., 2024).

Taken together, our theory of language model control as belief updating and our empirical results supporting this raise a number of important questions. In this work, we found that steering vectors control behavior proportional to the vector magnitude, unless that magnitude becomes too large (see App. C). This may suggest that belief is only represented linearly within some subspace of the model's representation space, although it is unclear whether belief is represented in a non-linear way outside this space, or whether this subspace represents the full extent of a model's belief state with respect to a given concept. Further, we found cases with some LLMs (e.g. `phi-4-mini-instruct`) where steering had no clear effect on behavior - this may suggest that these models represent belief in a non-linear way, or it may indicate a limitation of our particular method for constructing steering vectors. A simpler explanation could be that for some models and certain datasets, there is not sufficient signal for a distinct behavior in the LLM to be captured by our steering vectors - e.g. if a model doesn't represent a concept at all, then no amount of steering will change its belief in that concept.

Our work also raises questions about precisely how LLMs implement belief updates and inference. We find that steering beliefs typically only works in a single layer, or a few layers, while other layers have no clear effect on behavior. Does this suggest that belief is localized to these layers, and if so, could we causally intervene on specific neurons in these layers (Geiger et al., 2025) to have predictable impacts on model behavior? Further, although we find that beliefs are linearly represented and localized in these cases, this begs the question of how distinct aspects of belief and inference are implemented - are concept likelihood functions implemented in a non-linear way, and are they implemented in earlier or later layers relative to this linear belief representation? And, if an agent represents and updates its beliefs in this way, how is inference implemented - is it similar to known algorithms such as Monte Carlo methods or variational inference? Lastly, it is also noteworthy that