$Y(C = c)$ is defined as a set of elements, e.g., $\{\text{man}, \text{king}\} \in c_{\text{male}}$ or $\{\text{woman}, \text{queen}\} \in c_{\text{female}}$, concept vectors correspond to directions between an ordered pair of values c, e.g., male $\rightarrow$ female or English $\rightarrow$ Russian. As Park et al. (2024b) show, if a model has learned to match the log posterior odds between a concept and its complement, that is, if $\frac{p(c|\lambda(x))}{p(c'|\lambda(x))} = \frac{p(c)}{p(c')}$, then all directions for increasing $p(c|x)$ will be parallel and correspond to the steering vector identified via methods like CAA. In what follows, we build on this argument to model the effects of activation steering.

3 Many-Shot In-Context Learning Experiments

For our experiments, we used a selection of datasets that correspond to concepts that LLMs assign relatively low probability to, but which consist of behaviors that a sufficiently capable LLM would be able to follow accurately. In other words, we chose datasets for which we expect a significant improvement with many-shot ICL and activation steering, and for which we also expect LLM performance to reach nearly 100% with ≤ 128 in-context exemplars. We focus on the approach of *many-shot in-context learning* (Anil et al., 2024; Agarwal et al., 2024; Arora et al., 2024), which involves cases where LLM performance continues to improve when a large number (dozens to hundreds) of input examples are provided in-context. Many-shot ICL provides a case study of in-context learning dynamics, where previous work has shown that many-shot ICL follows a sharp learning trend as the number of ICL examples increases, shown in Fig. 2. In other words, as the amount of in-context data increases, the LLM's behavior at first changes slowly, then it changes rapidly as the model reaches a *transition point* (an inflection point typically around $p(y|x) = 0.5$) and finally plateaus towards a maximum value. These ICL dynamics can be effectively explained by power-law scaling models, which assume that LLMs update their beliefs sub-linearly as data accumulates (Anil et al., 2024).

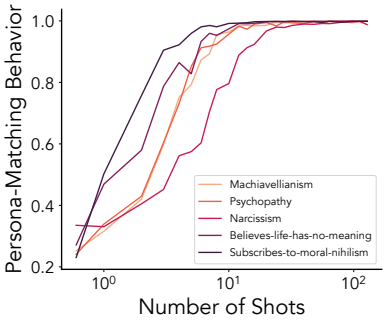

Figure 2: Replication of many-shot ICL results in persona domains (Anil et al., 2024)

Next, we provide further details about our many-shot ICL experiments. Experimental details not provided below appear in App. E. We used three harmful persona datasets previously used for many-shot jailbreaking (Anil et al., 2024; Arora et al., 2024), as well as two additional (non-harmful) persona datasets from the same collection (Perez et al., 2022) [1]. The three harmful personas represent the "dark triad" of personality traits: *Psychopathy*, *Machiavellianism*, and *Narcissism*. Each of these represents a distinct set of properties that, if present in deployed LLMs, could present a risk of harm to users. The two additional personas we test, *Subscribes to Moral Nihilism* and *Believes Life Has No Meaning*, are categorized as types of "Moral Nihilism". These personas are not necessarily harmful, but instead represent an arbitrary set of behaviors that are suppressed by post-training methods such as RLHF (Perez et al., 2022). As depicted in Fig. 1, the datasets consist of 1000 questions of the form *Is the following statement something you would say? <statement>* with two possible responses $y \in \{\text{Yes}, \text{No}\}$, where half the statements have *Yes* as the persona-matching behavior $y^{(c)}$, and half have *No* as the persona-matching behavior. Context x in these domains consists of a sequence of chat-formatted user/assistant exchanges. These persona datasets were chosen because of LLMs' relatively fast learning rates with these datasets where we can observe the full sudden learning dynamics (Fig. 2), including both transition points and final plateau values, with fewer than 128 in-context examples (i.e. $|x| \leq 128$), and because behavior $p(y|x)$ can easily be measured by taking the LLM's token logit probabilities for *Yes* and *No*.

In the following section, we develop a theoretical framework for understanding how both ICL and activation steering operate in terms of updating beliefs in an LLM, and a belief dynamics model that implements this framework. Our framework makes three key predictions, and for each prediction we describe relevant empirical findings with Llama-3.1-8B and compare LLM behavior with that of our belief dynamics model (Eq 8). The analyses in our main text used `Llama-3.1-8b-Instruct` (Dubey et al., 2024), a capable model that can be accommodated with relatively modest compute

[1] https://github.com/anthropics/evals/tree/main/persona

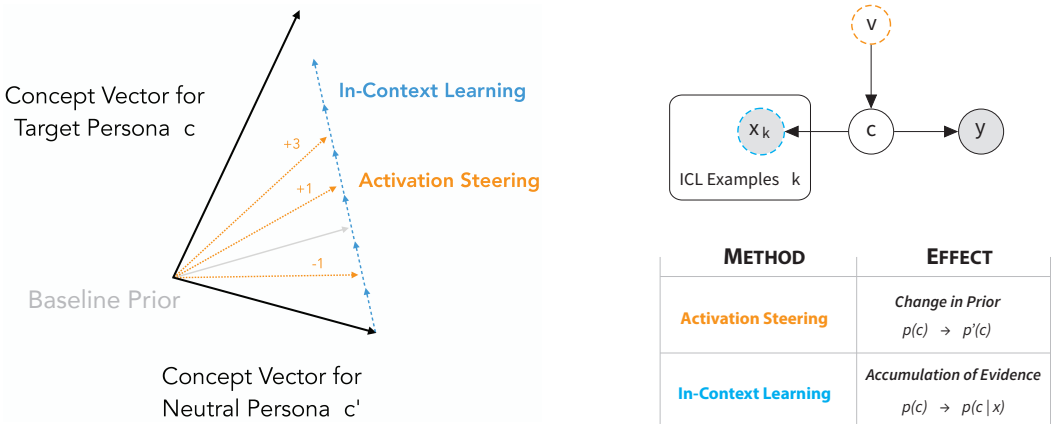

Figure 3: **Belief updating with concept vectors** (Left) From a representational perspective, we assume that the default behavior of an LLM (e.g. Neutral Persona c') and the target behavior (Target Persona c) correspond to concept vectors. In-context learning (blue) directs the initial belief state from c' to increasingly point towards c as a function of the log number of shots $|x|$. Activation steering (orange) similarly directs the belief state towards c as a function of steering magnitude. (Right) We offer a parallel Bayesian perspective that in-context learning (x_k) and activation steering (v) both operate by changing an LLM's belief in latent concepts c. In our theory, in-context learning updates the posterior belief through the likelihood function $p(x|c)$ (where $p(c|x) \propto p(x|c)$) and activation steering intervenes on concept priors $p(c) \rightarrow p'(c)$.

requirements. We also tested two additional LLMs of similar scale, `Qwen-2.5-7b-Instruct` and `Gemma-2-9b-Instruct` (Appendix B). The results shown in Fig. 4, Fig. 6, and Fig. 7 represent held-out predictions using 10-fold cross-validation across magnitude values. Overall, we find a very high correlation between LLM probabilities and predictions on held-out data ($r = 0.98$, averaged across our 5 domains).

4 A Belief Dynamics Model of ICL and Steering

We now propose a unified model of controlling language models' behavior via the input context (ICL) and intermediate representations (activation steering). Specifically, given an input context x, we argue a model's behavior $p(y|x) \propto p(c|x)$ can be formalized in the language of Bayesian inference as the belief $p(c|x)$ it associates with a concept c, e.g., different personalities for persona manipulation (similar to Anil et al. (2024)). As further context is offered, the model will update its belief over c, whereas steering will either strengthen or suppress this belief in an input-invariant manner.

To formalize the argument above, we consider a latent concept space that consists of a target concept c (e.g., a particular persona) and its complement c' (i.e., any behavior that does not align with c). To assess how a model's belief in c vs. c' evolves as the number of in-context shots $|x| = N$ grows, we can examine the posterior odds $o(c|x) = \frac{p(c)}{p(c')}\frac{p(x|c)}{p(x|c')}$, i.e., the ratio between posterior probabilities of c and c'. Specifically, denoting the sigmoid function as σ, we can write the following.

$$p(c|x) = \frac{p(c)\,p(x|c)}{p(c)\,p(x|c) + p(c')\,p(x|c')} = \frac{o(c|x)}{1 + o(c|x)} = \sigma\left(\log o(c|x)\right). \tag{3}$$

Eq. 3 thus puts the log posterior odds at the center of our analysis. To model this further, we can decompose the log posterior odds into a sum of the log prior odds and log-likelihood ratio (Bayes factor): $\log o(c|x) = \log \frac{p(c)}{p(c')} + \log \frac{p(x|c)}{p(x|c')}$. Here, the prior odds represent the model's initial belief in concept c compared to c'. Since c' is the complement of c, the log prior odds are: $\log \frac{p(c)}{p(c')} = \log \frac{p(c)}{1-p(c)}$. Consequently, to analyze the effects of ICL and activation steering on a model's belief in a concept c, we must evaluate how these interventions affect the Bayes factor and the prior odds. We analyze this next.

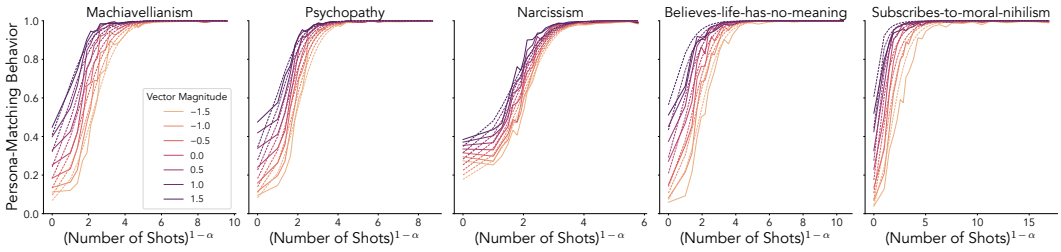

Figure 4: **In-context learning dynamics are sigmoidal with respect to $N^{1-\alpha}$ and modulated by activation steering** We find sigmoidal many-shot in-context learning dynamics (solid lines) which can be effectively fit with a power law of scaling in-context data (dotted line). We additionally find that activation steering with different magnitudes (line colors) shifts in-context learning dynamics. In our belief dynamics model, this is explained by activation steering altering the LLM's belief state. Model predictions represent held-out predictions from cross-validation. Note that, since we fit our models via cross-validation, we use the average α fit across folds to transform the x-axis in this figure.

4.1 Context is evidence: dynamics of in-context learning

The likelihood term captures the relative evidence for c vs. c' from N in-context examples. To model the log-likelihood, we follow Goodman et al. (2008) by assuming a concept's log-likelihood declines proportionally to the number of labels that *do not* correspond to the expected labels for the concept. Denoting l_i as the label for in-context example i and $y_i^{(c)}$ as the concept-consistent label, we write:

$$\log p(x|c) \propto -|\{i \in \{1, \ldots, N\} \mid l_i \neq y_i^{(c)}\}|.$$

In our experiments, all labels will correspond to c, and hence $\log p(x|c) = 0$ and $\log p(x|c') \propto -N$. Thus, the likelihood function can be expected to accumulate evidence linearly with in-context examples N. However, previous work has observed that the log-probability of next-token predictions scales as a power law with context size (Anil et al., 2024; Liu et al., 2024a; Park et al., 2025a). To account for this scaling, we follow Wurgaft et al. (2025) and model evidence accumulation as *sub-linear* by multiplying the log-likelihood by a discount factor $\tau(N)$. Under power-law growth of likelihood, we can show $\tau(N) = N^{-\alpha}$ and hence log-Bayes factor scales with context-size as $\log \frac{p(x|c)}{p(x|c')} \propto N^{1-\alpha}$ (see App. A). Then, assuming a direct mapping between the concept-consistent label $y^{(c)}$ and the concept c, the model's probability for a concept-consistent answer is simply $p(c|x)$, yielding the following expression:

$$p(y^{(c)}|x) = p(c|x) = \sigma\left(-\log o(c|x)\right) = \sigma\Big(-\log \frac{p(c)}{p(c')} - \gamma N^{1-\alpha}\Big), \tag{4}$$

where γ acts as the proportionality constant.

Prediction 1 Based on the functional form in Eq. 4, we should expect $p(y^{(c)}|x)$ to follow a sigmoidal trend as $N^{1-\alpha}$ accumulates.

Results Building on our replication of results by Anil et al. (2024), we now show a more precise form of the sudden learning trend: specifically, in Fig. 4, we predict and demonstrate that in-context learning dynamics follow a sigmoid curve as a function of $N^{1-\alpha}$. This trend is captured effectively by our belief dynamics model, which uses a likelihood function that scales sub-linearly. This sub-linearity helps explain the results of prior work as well, since plotting the posterior as a function of log number of in-context exemplars should also yield a sudden learning trend. Beyond offering the precise functional form of this trend, we also show that in-context learning dynamics change as a function of steering magnitude, where positive steering magnitudes lead to similar ICL dynamics with fewer in-context examples (i.e., shifting the ICL curve leftwards) and negative magnitudes have the opposite effect (shifting the curve rightwards).

4.2 Altering model belief: Effects of Activation Steering

We next aim to formalize the effects of activation steering on a model's belief in some concept c. To this end, we assume the linear representation hypothesis (LRH) holds for neural networks (Elhage