BELIEF DYNAMICS REVEAL THE DUAL NATURE OF IN-CONTEXT LEARNING AND ACTIVATION STEERING

Eric Bigelow[*1,2,4], **Daniel Wurgaft**[*1,5], **YingQiao Wang**[2],
Noah Goodman[5,6], **Tomer Ullman**[2,3], **Hidenori Tanaka**[3,4], **Ekdeep Singh Lubana**[1]
[1]Goodfire AI, [2]Department of Psychology, Harvard University, [3]CBS, Harvard University
[4]Physics of Intelligence Group, NTT Research, [5]Department of Psychology, Stanford University,
[6]Department of Computer Science, Stanford University, [*]Co-first authors

ABSTRACT

Large language models (LLMs) can be controlled at inference time through prompts (in-context learning) and internal activations (activation steering). Different accounts have been proposed to explain these methods, yet their common goal of controlling model behavior raises the question of whether these seemingly disparate methodologies can be seen as specific instances of a broader framework. Motivated by this, we develop a unifying, *predictive* account of LLM control from a Bayesian perspective. Specifically, we posit that both context- and activation-based interventions impact model behavior by altering its *belief in latent concepts*: steering operates by changing concept priors, while in-context learning leads to an accumulation of evidence. This results in a closed-form Bayesian model that is highly predictive of LLM behavior across context- and activation-based interventions in a set of domains inspired by prior work on many-shot in-context learning. This model helps us explain prior empirical phenomena—e.g., sigmoidal learning curves as in-context evidence accumulates—while predicting novel ones—e.g., additivity of both interventions in log-belief space, which results in distinct phases such that sudden and dramatic behavioral shifts can be induced by slightly changing intervention controls. Taken together, this work offers a unified account of prompt-based and activation-based control of LLM behavior, and a methodology for empirically predicting the effects of these interventions.

1 INTRODUCTION

Large Language Models (LLMs) have begun demonstrating increasingly impressive capabilities (Brown et al., 2020; Kaplan et al., 2020; Bubeck et al., 2023; Chang et al., 2024). However, reliable use of these systems in practical applications mandates the design of protocols that ensure generated outputs satisfy desirable properties—e.g., avoiding violent or harmful speech, sycophantic responses, or engagement with unsafe queries (Bai et al., 2022b;a; Anwar et al., 2024). To this end, prior work targeting inference-time control of model behavior has developed two broad methodologies: input-level interventions via *In-Context Learning (ICL)*, where contexts such as questions, instructions, dialog, or sequences of input-output examples are used to condition model behavior (Brown et al., 2020; Liu et al., 2023; Wei et al., 2022; Bai et al., 2022b;a), and representation-level interventions via *activation steering*, where a model's behavior is modulated by directly intervening on its hidden activations (Turner et al., 2024; Geiger et al., 2021; Templeton et al., 2024). Practical approaches to ICL often involve an informal process of prompt engineering through trial-and-error (White et al., 2023; Sahoo et al., 2024), whereas approaches to activation steering typically use ad-hoc datasets of contrasting pairs of examples (Turner et al., 2024; Marks and Tegmark, 2024).

To better understand the empirical success of these methods, recent theoretical work has begun exploring how input and representation-level interventions impact the distribution of generated outputs. Specifically, ICL has been framed as a form of Bayesian inference, where context modulates a space of hypotheses learned during pretraining (Xie et al., 2021; Bigelow et al., 2023; Wurgaft et al., 2025; Arora et al., 2024). Activation steering, on the other hand, has been argued to be a direct consequence of models learning to match the data distribution, which leads them to develop linear representations of concepts in particular layers (Park et al., 2024b; 2025b; Ravfogel et al., 2025;

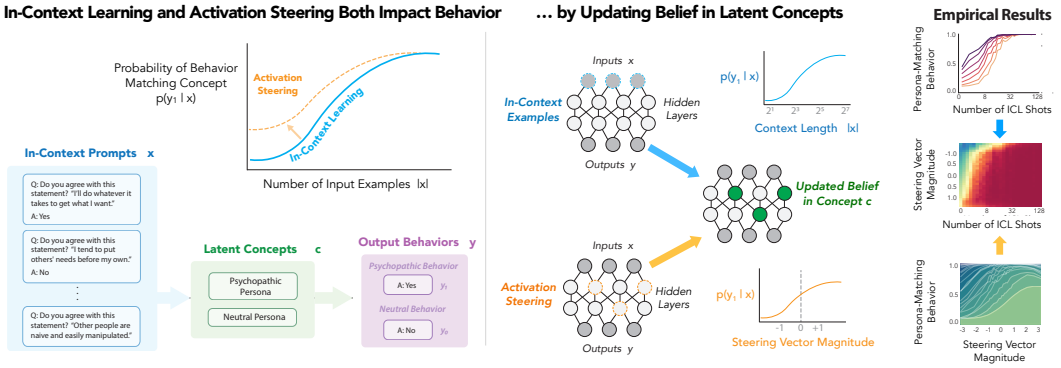

Figure 1: **Overview of our unified Bayesian theory of in-context learning and activation steering** We argue that in-context learning (ICL) and activation steering both impact behavior by updating an LLM's belief in latent concepts. We empirically test our claims in five domains of manipulating language model "persona" (bottom left) and predict that ICL will follow a sudden learning curve with increasing context length, and that this curve will be shifted under activation steering (top left). By our account, ICL with increasing context length $|x|$ and steering vectors with increasing magnitude both operate by updating an LLM's belief in latent concepts c.

Arora et al., 2016). Given the shared goals of ICL and activation steering, it is plausible that there is a broader framework that helps formalize the notion of control in a probabilistic system, with seemingly disparate approaches, such as ICL and activation steering, acting as specific instances of this framework.

This work Motivated by the above, we posit that various approaches to changing LLM behavior at inference time can be understood as *belief updating*. Specifically, we propose a Bayesian belief dynamics model where in-context learning reweighs concepts according to their likelihood functions, while steering reweighs concepts by altering their prior probabilities (Fig. 1). We design a set of experiments that build on prior work in many-shot ICL (Anil et al., 2024; Agarwal et al., 2024), and introduce activation steering magnitude as an additional dimension for belief updating, along with the number of ICL shots. Our results show three striking behavioral phenomena that can be predicted by our belief dynamics model: specifically, (i) a sigmoidal growth of posterior belief as a function of in-context exemplars, explaining prior results on sudden learning curves in ICL; (ii) predictable shift in the ICL behavior proportional to the magnitude of steering vector; and (iii) an *additive* effect of these interventions that yields distinct phases such that, as a function of intervention controls (context and steering magnitude), model behavior changes suddenly. Crucially, by formalizing and fitting our Bayesian model to the behavioral data, we are able to predict the point where this sudden change occurs, offering a concrete prediction for the phenomenon of many-shot jailbreaking (Anil et al., 2024).

More broadly, our work demonstrates the utility of applying a Bayesian perspective at various levels of analysis for understanding neural networks (Marr, 1982): to capture the space of behaviors that an LLM performs, as well as aid at understanding the representations underlying such behaviors. In our case, belief updating explains phenomena at both the level of behavior, i.e., how an LLM's output changes as a function of input given to it, and at the level of representation, i.e., in the effect of activation-level interventions. Correspondingly, this work contributes to a growing body of literature that uses Bayesian theories and models to study learning and conceptual representation in deep neural networks (Bigelow et al., 2023; Park et al., 2025a; Wurgaft et al., 2025). Building on the success of Bayesian approaches in explaining natural intelligence within cognitive science (Tenenbaum et al., 2011; Ullman and Tenenbaum, 2020), we argue that Bayesian principles can serve as a theoretical foundation for many different approaches to interpreting and controlling LLMs.

2 Background

We first offer a short primer highlighting points relevant to the two core phenomena that we aim to unify in this work: in-context learning and activation steering. We build on these points to define our Bayesian model in the next section.

2.1 In-Context Learning

In-Context Learning (ICL), where an LLM learns from linguistic context, is often contrasted with in-weights learning, where an LLM learns during (pre)training by adjusting model weights (Chan et al., 2022; Reddy, 2023; Lampinen et al., 2024; Nguyen and Reddy, 2024). While ICL is traditionally framed as few-shot learning (Brown et al., 2020), wherein exemplars corresponding to a task are offered to a model in-context and the model is expected to perform the demonstrated task on a novel query, there is a broader spectrum of language model capabilities that fall under the category of in-context learning (Lampinen et al., 2024; Park et al., 2025a; 2024a), e.g., zero-shot learning of a novel language (Gemini Team, 2023; Bigelow et al., 2023; Akyürek et al., 2024) or optimization of a utility function (Von Oswald et al., 2023; Demircan et al., 2024; Yin et al., 2024).

ICL as Bayesian Inference As argued by Xie et al. (2021); Bigelow et al. (2023); Panwar et al. (2024); Zhang et al. (2023); Min et al. (2022) and recently verified by Wurgaft et al. (2025); Park et al. (2024a); Raventós et al. (2024) in toy domains, different perspectives and phenomenology associated with ICL can be captured in a unifying, predictive framework by casting ICL as Bayesian inference. We build on this perspective by formalizing a Bayesian account of ICL in practical, large-scale settings. Specifically, following prior work, we define the distribution of model outputs y conditioned on input context x as inference over latent concepts c:

$$p(y|x) = \int_c p(y|c)\, p(c|x) \propto \int_c p(y|c)\, p(x|c)\, p(c). \tag{1}$$

The space of latent concepts $c \in \mathcal{C}$ is learned during model pretraining, and then, at inference time, these concepts are evoked by different input prompts x via the concept likelihood functions $p(x|c)$.

2.2 Activation Steering

Activation steering includes a broad set of protocols that intervene on the hidden representations of a language model to manipulate its outputs (Turner et al., 2024; Panickssery et al., 2024). Specifically, such protocols involve isolating directions d in the representation space such that moving a hidden representation v along them, i.e., altering v to $v + m \cdot d$, increases the odds the output reflects a concept c, e.g., truthfulness (Li et al., 2023; Pres et al., 2024). Surprisingly, this simple strategy enables control of model behavior across several abstract concepts such as refusal (Arditi et al., 2024), model personalities (Chen et al., 2025; Yang et al., 2025), concepts relevant to defining a theory-of-mind (Chen et al., 2024), factuality (Li et al., 2023), uncertainty (Zur et al., 2025), and self-representations (Zhu et al., 2024).

Contrastive Activation Addition For our experiments, we will primarily use the steering protocol introduced by Turner et al. (2024); Panickssery et al. (2024), called Contrastive Activation Addition (CAA) or "difference in means" steering. Specifically, CAA constructs steering vectors by collecting activations $a_\ell(X)$ from an LLM at the final token position of an input X, for a given layer ℓ, over two 'contrasting' datasets. As a specific example, suppose that $\mathcal{D}_c$ is a dataset of harmful prompts and $\mathcal{D}_{c'}$ is a dataset of harmless prompts. In this case, CAA can be used to identify a direction for steering towards (or against) harmful queries (Arditi et al., 2024). More formally, we write a general formulation of CAA steering protocols as follows.

$$\begin{aligned} \hat{d}_{c,\ell} &= \frac{1}{|\mathcal{D}_c|} \sum_{x \in \mathcal{D}_c} a_\ell(x) \;-\; \frac{1}{|\mathcal{D}_{c'}|} \sum_{x \in \mathcal{D}_{c'}} a_\ell(x) \\ &= \mathbb{E}_{p(x|c)} \Big[a_\ell(x) \Big] \;-\; \mathbb{E}_{p(x|c')} \Big[a_\ell(x) \Big] \end{aligned} \tag{2}$$

Linear representation hypothesis and activation steering It is unclear precisely why activation steering methods work. These methods are similar in nature to analogies in word vector algebra (Mikolov et al., 2013), as in the classic example `king : queen :: man : woman`, which can be represented in vector algebra as $v(\text{king}) - v(\text{queen}) = v(\text{man}) - v(\text{woman})$. The Linear Representation Hypothesis (Park et al., 2024b; 2025b) formalizes this connection in terms of embedding representation $\lambda(x)$ and an unembedding representation $\gamma(y)$, where output behavior given an input $p(y \mid x)$ is the softmax of the inner product: $p(y \mid x) \propto \exp\big(\, \lambda(x)^\top \gamma(y) \,\big)$. If each concept variable