Appendices

A Derivations

A.1 Derivation of the Bayes Factor

To model LLM in-context learning behavior, we examine the dynamics of the posterior belief of a model in a concept c, $p(c|x)$, as context length $|x| = N$ is varied. To study this, we consider the posterior odds between the concept c and its complement c':

$$o(c|x) = \frac{p(c|x)}{p(c'|x)}.$$

The posterior odds represent the model's posterior belief in c versus its posterior belief in c' after seeing context x, and given prior preference. This can be further decomposed as follows.

$$\begin{aligned}\log o(c|x) &= \log \frac{p(c)\ p(x|c)}{p(c')\ p(x|c')} \\ &= \log \frac{p(c)}{p(c')} + \log \frac{p(x|c)}{p(x|c')}\end{aligned}$$

To model the log posterior odds, we must capture both the prior and likelihood-related terms. We discuss our model of prior odds in Sec. 4.2. To compute the log-likelihoods, we make two crucial assumptions, as follows.

1. **Concept log-likelihood declines proportionally to the number of mismatched labels**: The persona-adoption settings we examine consist of query-label examples where labels are binary and either map or do not map to a persona. Thus, it is reasonable that the log-likelihood for a concept will decline proportionally with the number of mismatched labels seen. Assuming this likelihood function follows Goodman et al. (2008), who studied rule-based concept learning in humans. Formally, we can express the likelihood function for a concept c as:

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

   where l_i is a seen label and $y_i^{(c)}$ is the persona-consistent label for the in-context query i. Since in the settings we study all labels are consistent with the persona, that is, $l_i = y_i^{(c)}$, $\forall\ i \in \{1, \ldots, N\}$, we infer:

$$\begin{aligned}\log p(x|c) &\propto -|\{i \in \{1, \ldots, N\} | l_i \neq y_i^{c}\}| = 0, \text{ and} \\ \log p(x|c') &\propto -|\{i \in \{1, \ldots, N\} | l_i \neq y_i^{c'}\}| = -N.\end{aligned}$$

2. **Log-likelihood scales as a power-law with number of in-context examples** N: This assumption aims at accommodating the power-law behavior observed in studies of LLM in-context learning (Anil et al., 2024; Liu et al., 2024b). Specifically, we assume the common form of a scaling-law from scaling laws predicting loss during pretraining $L(n) \approx L(\infty) + \frac{A}{n^\alpha}$ (Kaplan et al., 2020). However, in our case, $L(N)$ represents the negative-log likelihood for the N-th in-context example (Anil et al., 2024). Given this assumption, we derive a sub-linear discount term τ that arises from the ratio between the negative log-likelihood for N in-context examples under the power-law assumption, and the negative log-likelihood given by an optimal Bayesian agent using the likelihood function from assumption 1. Following the derivation from Wurgaft et al. (2025), we write:

$$
\begin{aligned}
\tau &:= \frac{\text{NLL under power-law scaling for } N \text{ in-context examples}}{\text{NLL given by a Bayesian Learner for } N \text{ in-context examples}} \\
&= \frac{\sum_{n=1}^{N}(L(n) - L(\infty))\delta n}{N} \\
&= \frac{1}{N}\sum_{n=1}^{N}\frac{A}{n^{\alpha}}\delta n \\
&= AN^{-\alpha}\int_{0}^{1}\frac{1}{\hat{n}^{\alpha}}\delta\hat{n} \\
&= \frac{A}{1-\alpha}N^{-\alpha} \\
&= \gamma N^{-\alpha}
\end{aligned}
$$

where $\hat{n} = {}^{n}/_{N}$ and $\gamma = \frac{A}{1-\alpha}$ is a constant that incorporates A, the constant from our power-law form.

Final expression for Bayes Factor. Following the assumptions above, we can write the functional form for the log Bayes-factor as:

$$
\begin{aligned}
\log\frac{p(x|c)}{p(x|c')} &= \log p(x|c) - \log p(x|c') \\
&\approx \gamma N^{-\alpha}(-|\{i \in \{1,\ldots,N\}|l_i \neq y_i^{(c)}\}| + |\{i \in \{1,\ldots,N\}|l_i \neq y_i^{(c')}\}|) \\
&= \gamma N^{1-\alpha}.
\end{aligned}
$$

A.2 Derivation of the Effect of Steering Magnitude (Eq. 7)

Here we show how the log posterior odds (Eq. 7) can be represented as:

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

or equally:

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

Note that the last term does not depend on v.

Recall that a given vector embedding v is defined, according to the Linear Representation Hypothesis, as a linear weighted sum of concept vectors d_i weighted by $\beta_i(v)$, i.e. how much concept c_i is present in v:

$$v = \sum_i \beta_i(v)\, d_i$$

with the constraint that concept vectors are approximately orthogonal, i.e. $d_i^T d_j \approx 0$.

Next, the conditional probability is given by

$$\begin{aligned} p(c_i \mid v) &= \sigma\left(-w_i^T v - b\right) \\ &= \sigma\left(\eta\right) \end{aligned}$$

where $\eta = -w_i^T v - b$. We further assume that our weight vector w approximates concept vector d_i scaled by an arbitrary value k:

$$w \approx k\, d_i$$

Now, consider a shifted representation $v + m \cdot d_i$, where we substitute $w \to k\, d_i$ and $v \to v + m \cdot d_i$:

$$\begin{aligned} p(c_i \mid v + m \cdot d_i) &= \sigma\left(-k\, d_i^T (v + m \cdot d_i) - b\right) \\ &= \sigma\left(-k\, d_i^T v - b - k\, m \|d_i\|^2\right) \end{aligned}$$

This shows a linear effect of steering magnitude m in logit space.

Next, we can represent the log posterior odds as e^η:

$$\begin{aligned} \frac{p(c_i \mid v)}{p(c_i' \mid v)} &= \frac{p(c_i \mid v)}{1 - p(c_i \mid v)} \\ &= \frac{\sigma(\eta)}{1 - \sigma(\eta)} \\ &= \frac{1/(1 + e^\eta)}{1 - 1/(1 + e^\eta)} \\ &= \frac{1/(1 + e^\eta)}{e^\eta/(1 + e^\eta)} \\ &= e^{-\eta} \end{aligned}$$

Mapping this into log space, we get:

$$\log \frac{p(c_i \mid v)}{p(c_i' \mid v)} = -\eta = w_i^T v + b$$