FUNDAMENTAL LIMITATIONS OF ALIGNMENT IN LARGE LANGUAGE MODELS

Yotam Wolf*
The Hebrew University
yotamwolf@cs.huji.ac.il

Noam Wies*
The Hebrew University
noam.wies@cs.huji.ac.il

Oshri Avnery
The Hebrew University
oshri.avnery@cs.huji.ac.il

Yoav Levine
AI21 Labs
yoavl@ai21.com

Amnon Shashua
The Hebrew University
shashua@cs.huji.ac.il

ABSTRACT

An important aspect in developing language models that interact with humans is aligning their behavior to be useful and unharmful for their human users. This is usually achieved by tuning the model in a way that enhances desired behaviors and inhibits undesired ones, a process referred to as *alignment*. In this paper, we propose a theoretical approach called Behavior Expectation Bounds (BEB) which allows us to formally investigate several inherent characteristics and limitations of alignment in large language models. Importantly, we prove that for any behavior that has a finite probability of being exhibited by the model, there exist prompts that can trigger the model into outputting this behavior, with probability that increases with the length of the prompt. This implies that any alignment process that attenuates undesired behavior but does not remove it altogether, is not safe against adversarial prompting attacks. Furthermore, our framework hints at the mechanism by which leading alignment approaches such as reinforcement learning from human feedback increase the LLM's proneness to being prompted into the undesired behaviors. Moreover, we include the notion of personas in our BEB framework, and find that behaviors which are generally very unlikely to be exhibited by the model can be brought to the front by prompting the model to behave as specific persona. This theoretical result is being experimentally demonstrated in large scale by the so called contemporary "chatGPT jailbreaks", where adversarial users trick the LLM into breaking its alignment guardrails by triggering it into acting as a malicious persona. Our results expose fundamental limitations in alignment of LLMs and bring to the forefront the need to devise reliable mechanisms for ensuring AI safety.

1 INTRODUCTION

Training large language models (LLMs) over vast corpora has revolutionized natural language processing, giving LLMs the ability to mimic human-like interactions and serve as general purpose assistants in a wide variety of tasks, such as wide-scoped question answering, writing assistance, teaching, and more (Radford et al., 2019; Devlin et al., 2019; Brown et al., 2020; Schulman et al., 2023; OpenAI, 2023; Bubeck et al., 2023; Nori et al., 2023; West, 2023; Park et al., 2023). A growing concern due to the increasing reliance on LLMs for such purposes is the harm they can cause their users, such as feeding fake information (Lin et al., 2022; Weidinger et al., 2022), behaving offensively and feeding social biases (Hutchinson et al., 2020; Venkit et al., 2022; Weidinger et al., 2022), or encouraging problematic behaviors by users (even by psychologically manipulating them Roose

*Equal contribution

(2023); Atillah (2023)). Indeed, evidently, the unsupervised textual data used for pretraining modern LLMs includes enough demonstrations of the above undesired behaviors for them to be present in the resulting models (Bender et al., 2021). The act of removing these undesired behaviors is often called *alignment* (Yudkowsky, 2001; Taylor et al., 2016; Amodei et al., 2016; Shalev-Shwartz et al., 2020; Hendrycks et al., 2021; Pan et al., 2022; Ngo, 2022).

There are several different approaches to performing alignment in LLMs. One is to include aligning prompts: Askell et al. (2021) show that injecting language models with helpful, honest, and harmless (HHH) textual prompts improves alignment and decreases toxicity. Similarly, Rae et al. (2021) also use prompting in order to decrease toxicity. Another approach for LLM alignment is the procedure of reinforcement learning from human feedback (RLHF) in order to train language models to be helpful and harmless (Bai et al., 2022). The procedure is to further train a pretrained language model with the assistance of a human evaluator in order to optimize its outputs to the evaluator's preferences. Their work shows an increase in an LLM's HHH scores while maintaining its useful abilities, as measured by zero- and few-shot performance on different natural language tasks. Another notable work using this method is by Ouyang et al. (2022), which fine tune GPT-3 into InstructGPT using data collected from human labelers to reach better performance on a variety of tasks, while improving HHH (measured via bias and toxicity datasets Gehman et al. (2020); Nangia et al. (2020)).

While the above approaches to alignment are effective to a certain extent, they are still dangerously brittle. For example, Wallace et al. (2019) show that short adversarial prompts can trigger negative behaviors and social biases. Yu & Sagae (2021) and Xu et al. (2021) provide methods for exposing harmful behaviors of models by triggering problematic responses. Subhash (2023) showed that adversarial prompts can manipulate ChatGPT to alter user preferences. Beyond academic works, the general media is abundant with contemporary examples of leading LLMs being manipulated by users to expose harmful behaviors via the so called "jailbreaking" approach of prompting the LLM to mimic a harmful persona (Nardo, 2023; Deshpande et al., 2023). Even in the absence of adversarial attacks, leading alignment methods can underperform and are not well understood: Perez et al. (2022) provide evidence that certain negative behaviors have inverse scaling with the number of RLHF steps, indicating that this popular alignment procedure may have a complex effect.

In this paper, we introduce a probabilistic framework for analyzing alignment and its limitations in LLMs, which we call *Behavior Expectation Bounds* (BEB), and use it in order to establish fundamental properties of alignment in LLMs. The core idea behind BEB is to represent the LLM distribution as a superposition of ill- and well-behaved components, in order to provide guarantees on the ability to restrain the ill-behaved components, *i.e.*, guarantees that the LLM is aligned. It is noteworthy that LLMs have been shown to distinctly represent behaviors and personas, and the notion of persona or behavior superposition has been intuitively proposed as an explanation Andreas (2022); Nardo (2023).

Our BEB framework assumes an underlying categorization into different behaviors, where any natural language sentence is assigned a ground truth score between -1 (very negative) and $+1$ (very positive) for every behavior (see examples in Figure 1). Such a categorization can be, *e.g.*, into the previously proposed helpful, honest, and harmless categories, but it can also be expanded and fine-grained into many more categories such as polite, not racist, compassionate, and so on. Given such a categorization and ground truth sentence scoring functions per category, the alignment score of any distribution over natural sentences *w.r.t.* a given behavior is the expectation value of sentence scores for sentences drawn from the distribution. The BEB framework thus provides a natural theoretical basis for describing the goal of contemporary alignment approaches such as RLHF: increasing the behavior expectation scores for behaviors of interest.

The BEB framework employs assumptions on the distinguishability of the ill- and well-behaved components within the overall LLM distribution. We present these assumptions and the BEB framework in section 2, and use it in section 3 in order to assert several important statements regarding LLM alignment:

- **Alignment impossibility**: We show that an LLM alignment process which reduces undesired behaviors to a small but nonzero fraction of the probability space is not safe against adversarial prompts (theorem 1);

  Informal theorem: If the LLM has finite probability of exhibiting negative behavior, there exists a prompt for which the LLM will exhibit negative behavior with probability 1.

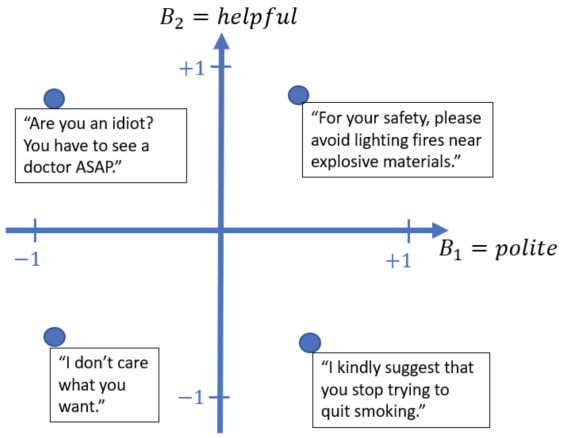

Figure 1: Examples of sentence behavior scores along different behavior verticals. Our framework of Behavior Expectation Bounds (BEB) assumes ground truth behavior scoring functions, and bounds the expected scores of sentences along different behavior verticals in order to guarantee LLM alignment or misalignment.

- **RLHF can make things worse**: We show that increased distinction between desired and undesired behavior can render the LLM more susceptible to adversarial prompting (theorem 2). We conjecture that while RLHF lowers the probability of undesired behaviors, it may also sharpen their distinction from desired behaviors, increasing vulnerability to adversarial prompting (conjecture 1);

  Informal theorem: The better the distinction between positive and negative behaviors, the shorter the adversarial prompt required to elicit undesired behaviors.

- **Preset aligning prompts are effective**: We prove that even including an aligning prefix prompt does not guarantee alignment (theorem 3), but show that it does form an adversarial prompt length guardrail under which alignment is guaranteed (theorem 4);

  Informal theorem: In order to misalign a prompt-aligned LLM, one must insert text of length that is on the order of that of the aligning prompt.

- **LLMs can resist misalignment during a conversation**: We show that if a user attempts to misalign an LLM during a conversation, the LLM can restore alignment during its conversation turns (theorem 5);

  Informal theorem: an adversarial user will need to insert more text in a conversation scenario than in a single prompt scenario in order to misalign the LLM.

- **Imitating personas can lead to easy alignment "jailbreaking"**: We show that it is always possible to prompt a language model into behaving as a certain persona it has captured during pretraining (theorem 6), and further show that this mechanism can be used in order to easily access undesired behaviors (corollary 2).

  Informal theorem: Mimicking personas that demonstrate bad behaviors can be more efficient than directly evoking the same bad behavior.

Overall, we hope that our newly proposed framework of Behavior Expectation Bounds, along with our attained results, may spark a theoretical thrust helping to better understand the important topic of LLM alignment.

2 Behavior Expectation Bounds: A framework for analyzing LLM Alignment

In this section, we introduce Behavior Expectation Bounds (BEB), a probabilistic framework for studying alignment of LLMs. Given a language model's probability distribution $\mathbb{P}$, we propose a measure for quantifying its tendency to produce desired outputs as measured by a certain behaviour

vertical B, where for example B can be helpfulness, politeness, or any other behavior vertical of interest. Formally, we model behaviour scoring functions along vertical B as $B : \Sigma^\star \to [-1, 1]$, which take a string of text from an alphabet Σ as their input and rate the manner in which B manifests in the string, with $+1$ being very positive and -1 being very negative. This formulation directly reflects recent empirical efforts for studying alignment. In particular, (Perez et al., 2022) recently curated 500 negative and positive examples along each of over 100 different behavior verticals. Figure 1 shows short intuitive examples of the behavior scores of several sentences along two behavior verticals.

We use the following *expected behavior scoring* of distribution $\mathbb{P}$ *w.r.t.* behavior vertical B as a scalar quantifyer of the tendency of $\mathbb{P}$ to produce desired behavior along the B vertical:

$$B_{\mathbb{P}} := \mathbb{E}_{s\sim\mathbb{P}}[B(s)] \tag{1}$$

where for clarity purposes, in this paper sampling from language distributions is implicitly restricted to single sentences. We will use the above distribution notation $\mathbb{P}$ to represent that of an unprompted LLM, *e.g.*, an LLM straight out of pretraining or out of an alignment tuning procedure such as RLHF. Indeed, the task of aligning a pretrained LLM can be now framed as increasing its expected behavior scores along behavior verticals of interest.

Intuitively, as an LLM is prompted with a prefix text string s^*, the behaviour of the conditional probability $\mathbb{P}(\cdot \mid s^*)$ might change in accordance with the in-context learning phenomenon (Brown et al., 2020; Wies et al., 2023) in which the LLM adapts its conditional probabilities to reflect its current textual context. Thus, we will denote by $B_{\mathbb{P}}(s^*)$ the behaviour of the language model when prompted with a prompt text s^*:

$$B_{\mathbb{P}}(s^*) := \mathbb{E}_{s\sim\mathbb{P}(\cdot|s^*)}[B(s)] \tag{2}$$

We will consider several scenarios for which the prefix s^* plays different roles. The first and main one is that s^* serves as an adversarial input prompt. Our key finding in this paper is that an LLM which was initially aligned *w.r.t.* a certain behavior vertical, *i.e.*, $B_{\mathbb{P}}$ very close to 1, can still be vulnerable to adversarial prompts, *i.e.*, there exists a prompt s^* such that $B_{\mathbb{P}}(s^*)$ is very close to -1. Secondly, we will consider a scenario in which s^* is comprised of an initial aligning prompt, denoted s_0, concatenated by a subsequent adversarial input prompt. Lastly, we will analyze conversation scenarios in which s^* is comprised of previous turns of user queries and LLM responses.

2.1 LLMs AS A SUPERPOSITION OF BEHAVIORS OR PERSONAS

In this subsection, we present a key aspect of our BEB framework: decomposing the LLM distribution $\mathbb{P}$ into a mixture of distributions, each behaving differently. Importantly, LLMs exhibit signs of capturing such decompositions in practice. For example, Andreas (2022) shows empirical evidence that current LLMs can infer behaviours from textual prompts, and that these behaviours affect the text that the LLM generates, and Nardo (2023) discuss LLMs as a superposition of personas. We will use mixture decompositions inspired by such observations, and prove that textual prompts can reweight the prior of the mixture components. In appendix K, we experimentally demonstrate that the embedding space of contemporary leading LLMs (LLaMA family (Meta, 2023)) is clustered according to positive and negative inputs *w.r.t.* behaviors of interest (assembled by (Perez et al., 2022)), and empirically show that this clustering approximately corresponds to our analyzed mixture decomposition model, presented hereinafter.

2.1.1 THE GOOD AND THE BAD

Observe that for any decomposition of a distribution $\mathbb{P}$ into two components, $\mathbb{P} = \alpha\mathbb{P}_0 + (1-\alpha)\mathbb{P}_1$, the relation $B_{\mathbb{P}} = \alpha B_{\mathbb{P}_0} + (1-\alpha)B_{\mathbb{P}_1}$ holds from linearity of expectations, and implies that one component is more well-behaved *w.r.t.* B than the full distribution and the other more ill-behaved, *i.e.*: $B_{\mathbb{P}_1} \leq B_{\mathbb{P}} \leq B_{\mathbb{P}_0}$ (or vice versa). For this reason, focusing on a specific behavior, we adopt the notation:

$$\mathbb{P} = \alpha\mathbb{P}_- + (1-\alpha)\mathbb{P}_+ \tag{3}$$

We refer to the above as the *two component mixture*, where $\mathbb{P}_+$ is the well-behaved component and $\mathbb{P}_-$ is the ill-behaved component.

While this observation is true for any decomposition into two distributions, we will give results for decompositions in which the two distributions $\mathbb{P}_-$ and $\mathbb{P}_+$ are sufficiently distinct (formally defined in section 2.2), and we are interested in decompositions where the negative component is strictly ill-behaved (i.e, $B_{\mathbb{P}_-} \leq \gamma < 0$). In these cases, the magnitude of α, the prior of the ill-behaved component, will determine the alignment of the LLM: an LLM with a small prior α will be less likely to produce undesired sentences along behavior B vertical. Our main result in section 3 states that no matter how small α is (how aligned the model is to begin with), if it is positive then there exists a prompt that can misalign the LLM to behave like $\mathbb{P}_-$.

2.1.2 MULTIPLE PERSONAS

A natural extension of the above two components mixture, is a decomposition into more than two components, $\mathbb{P}(s) = \sum_{\phi \in \Phi} w_\phi \mathbb{P}_\phi(s)$. Indeed, for any such decomposition, each component may be more well-behaved than the full model $B_{\mathbb{P}_\phi} \geq B_{\mathbb{P}}$ or more ill-behaved $B_{\mathbb{P}_\phi} \leq B_{\mathbb{P}}$, *w.r.t.* a given behavior B. For a different behavior B', some of these inequalities may be flipped. We therefore refer to different components $\mathbb{P}_\phi$ as different "personas", as each component represents a different mixture of behaviors. Still, the weighted sum of the components always gives that of the model $B_{\mathbb{P}} = \sum_{\phi \in \Phi} w_\phi B_{\mathbb{P}_\phi}$.

This decomposition is more refined than the two components one, and in fact can reproduce it: Any partition of the persona into two sets defines a two component mixture. For example, *w.r.t.* a behavior B, for $a_- = \{\phi \in \Phi : B_{\mathbb{P}_\phi} < \gamma\}$ and $a_+ = \Phi \backslash a_-$, the two terms $P_+ \propto \sum_{\phi \in a_+} w_\phi \mathbb{P}_\phi$ and $P_- \propto \sum_{\phi \in a_-} w_\phi \mathbb{P}_\phi$ define the two component decompositon with the ill-behaved part satisfying $B_{\mathbb{P}_-} < \gamma$. In section 3.3 we will use the above decomposition in order to shed light on the so called "chatGPT jailbreak" attack on LLM alignment, in which the LLM is prompted into playing a specific persona and as a side effect exhibits an undesired behavior (Nardo, 2023).

2.2 DEFINITIONS FOR BOUNDING THE EXPECTED LLM BEHAVIOR

In this subsection, we lay out formal definitions of our BEB framework. Specifically, we define: behavior misalignment using prompts (definition 1); distinguishability and similarity between two distributions that fit a prompting scenario (definitions 2 and 3, respectively); distinguishibility between ill- and well-behaved components comprising a certain LLM's distribution (definition 4, generalized for the case of analyzing "personas" in definition 5).

Once an LLM has finished training, its behavior can only be affected via prompting. Using the above notation for behavior expectation (equations 1 and 2), the following defines when an LLM is *prompt-misalignable*:

Definition 1. *Let $\gamma \in [-1, 0)$, we say that an LLM with distribution $\mathbb{P}$ is γ-**prompt-misalignable*** w.r.t. *behaviour B, if for any $\epsilon > 0$ there exists a textual prompt $s^* \in \Sigma^\star$ such that $B_{\mathbb{P}}(s^*) < \gamma + \epsilon$.*

Decomposing a language model into parts that are well-behaved and ill-behaved exposes components which are more desirable to enhance. The following notion of *distinguishability* will allow us to guarantee that one component can be enhanced over the other [1].

Definition 2. *We say that a distribution $\mathbb{P}_\phi$ is β-**distinguishable** from distribution $\mathbb{P}_\psi$ if for any prompt s_0:*

$$D_{KL}\left(\mathbb{P}_\phi\left(\cdot|s_0\right) \;||\; \mathbb{P}_\psi\left(\cdot|s_0\right)\right) := \mathbb{E}_{s\sim\mathbb{P}_\phi(\cdot|s_0)}\left[\log \frac{\mathbb{P}_\phi\left(s|s_0\right)}{\mathbb{P}_\psi\left(s|s_0\right)}\right] > \beta \tag{4}$$

The following bounds the extent to which a new sentence can enhance one component over the other:

Definition 3. *We say that a distribution $\mathbb{P}_\phi$ is c-**similar** to distribution $\mathbb{P}_\psi$ if there exists $c > 0$ such that for any sequence of sentences s_0, and any sentence s, the following holds:*

$$\forall s, s_0 : \;\; \log \frac{\mathbb{P}_+\left(s|s_0\right)}{\mathbb{P}_-\left(s|s_0\right)} < c \cdot \tag{5}$$

[1] Note that the β-distinguishability definition can be relaxed to a KL distance that decays as a power law to zero with increasing length of the prompt s_0, as shown in appendix I

Intuitively, if both $\mathbb{P}_\phi$ and $\mathbb{P}_\psi$ are natural language distributions, they cannot be too different in terms of the ratio of their conditional likelihoods, and c quantifies this. Furthermore, when $\mathbb{P}_\phi$ and $\mathbb{P}_\psi$ represent positive and negative angles of a specific behaviour, it is likely that they have some common properties so in these cases c is likely even lower than the bound over all natural language sentences. In appendix J, we present an approximate measurement which shows that $c/\beta \sim 10$ for negative and positive distributions induced by the contemporary LLaMA LLM family (Meta, 2023) for behaviors of interest assembled in (Perez et al., 2022). c and β roughly serve as upper and lower bounds on the KL-divergence, and this ratio will appear in several of our results in section 3.

The following defines β-distinguishability specifically between the ill- and well-behaved components comprising the LLM distribution, parameterized by α in equation 3, and adds a condition that the behavior expectation of the ill- or well-behaved component is either bad enough or good enough (*i.e.*, either under γ or over Γ, respectively) for all initial prompts s^*:

Definition 4. *Let $\gamma \in [-1, 0)$, $\Gamma \in (0, 1]$, and assume $\mathbb{P} = \alpha \cdot \mathbb{P}_- + (1-\alpha) \cdot \mathbb{P}_+$ for $\alpha > 0$. We say that behaviour $B : \Sigma^\star \to [-1, 1]$ is α, β, γ-**negatively-distinguishable** in distribution $\mathbb{P}$, if $\sup_{s^*}\{B_{\mathbb{P}_-}(s^*)\} \leq \gamma$ and $\mathbb{P}_-$ is β-distinguishable from $\mathbb{P}_+$ (def. 2), or that B is α, β, Γ-**positively-distinguishable** in $\mathbb{P}$ if $\inf_{s^*}\{B_{\mathbb{P}_+}(s^*)\} \geq \Gamma$ and $\mathbb{P}_+$ is β-distinguishable from $\mathbb{P}_-$ (def. 2).*

We will prove our theoretical results for LLM distributions that are distinguishable according to the above KL-divergence based definitions. The following proposition assures us that if the two components indeed have sufficiently distinct behavior expectation, the above KL-divergence based distinguishability condition is implied:

Proposition 1. *Let $\mathbb{P}_\phi$ and $\mathbb{P}_\psi$ be two distributions, with behavior expectations $B_{\mathbb{P}_\phi}$, $B_{\mathbb{P}_\psi}$ and behavior variances $V_{\mathbb{P}_\phi}$, $V_{\mathbb{P}_\psi}$. The KL-divergence between the first and second are bounded from below by a positive term that increases with the magnitude of difference in behavior $|B_{\mathbb{P}_\phi} - B_{\mathbb{P}_\psi}|$.*

Lastly, we generalize definition 4 of behavior distinguishability within an LLM's distribution to the setting of personas, as follows:

Definition 5. *Let $\gamma \in [-1, 0)$, we say that a behavior $B : \Sigma^\star \to [-1, 1]$ is α, β, γ-**negatively-distinguishable in persona mixture** $\mathbb{P} = \sum_{\phi \in \Phi} w_\phi \mathbb{P}_\phi$, if for any $\epsilon > 0$, there exists a persona $\tilde{\phi}$, that satisfies, $w_{\tilde{\phi}} \geq \alpha$, $sup_{s^*}[B_{\mathbb{P}_{\tilde{\phi}}}(s^*)] < \gamma + \epsilon$ and is β-distinguishable (def. 2) from any other persona ϕ.*

We will show that evoking a malicious persona can be a good strategy for eliciting bad behavior from LLMs that obey the above.

3 Results: Limitations of LLM alignment

In this section, we use the above framework of Behavior Expectation Bounds (BEB) in order to elucidate the question of when LLM alignment is robust or vulnerable to adversarial prompting attacks. We begin with our main result in section 3.1, which states that under assumptions of decomposability into distinguishable components of desired and undesired behavior, aligned LLMs are not protected against adversarial misaligning prompts (theorem 1). We show that the more distinguishable the components, the shorter the misaligning prompt required to misalign the LLM (theorem 2). This last result may shed light on scenarios in which common alignment tuning practices render the aligned LLM more brittle (Perez et al., 2022)(conjecture 1).

In section 3.2, we extend the above framework to include cases of (i) preset aligning prompts—we formally establish the benefits of this common practice by showing that in this case the length of the misaligning prompt must be linear in the length of the preset aligning prompt; and (ii) multi-turn interactions between adversarial users and LLMs—we find that if the user does not provide long enough misaligning prompts, the LLM can resist misalignment by making aligning replies to the user during a conversation. Finally, in section 3.3, we analyze the case of decomposing the LLM distribution into multiple components ("personas", or, mixtures of behaviors, presented in section 2.1.2), and show that if a certain persona is distinctly captured during the LLM pretraining, evoking it in order to elicit bad behavior from an aligned LLM can be more efficient than directly trying to elicit this behavior from the LLM. This corresponds to the recently popularized "chatGPT jailbreaking" practice of misaligning an LLM via requesting it to mimic a malicious persona.

3.1 MISALIGNING VIA ADVERSARIAL PROMPTS

Alignment impossibility We first show that if a model can be written as a distinct mixture of ill- and well-behaved components, then it *can* be misaligned via prompting:

Theorem 1. *Let $\gamma \in [-1, 0)$ and let B be a behaviour and $\mathbb{P}$ be an unprompted language model such that B is α, β, γ-negatively-distinguishable in $\mathbb{P}$ (definition 4). Then $\mathbb{P}$ is γ-prompt-misalignable* w.r.t. *B (definition 1) with prompt length of $O(\frac{1}{\beta}(\log \frac{1}{\alpha} + \log \frac{1}{\epsilon}))$.*

Intuitively, theorem 1 implies that if a component of the distribution exhibits a negative behavior with expectation under γ, then there exists a prompt that triggers this behavior for the entire language model into behaving with expectation under γ. Importantly, no matter how low the prior of the negative component α is, if it is distinguishable within the distribution then the LLM is vulnerable to adversarial prompting that exposes this negative component's behavior. Essentially, our proof follows the PAC based theoretical framework for in-context learning introduced in Wies et al. (2023), while relaxing their approximate independence assumption and adapting the analysis to the BEB framework. We provide below a sketch for the proof of theorem 1, fully detailed in appendix B:

Proof sketch (see full details in the appendix). The assumption that B is α, β, γ-negatively-distinguishable in $\mathbb{P}$ implies that $\mathbb{P}$ can be written as a mixture distribution of a misaligned component $\mathbb{P}_{\_}$ and an aligned component $\mathbb{P}_{+}$. Now, while the prior of $\mathbb{P}_{\_}$ might be low and hence the behaviour of the unprompted $\mathbb{P}$ is initially aligned with high probability, the fact that $\mathbb{P}_{\_}$ is β-distinguishable from $\mathbb{P}_{+}$ assures us that the conditional KL-divergence between $\mathbb{P}_{\_}$ and $\mathbb{P}_{+}$ is greater than β for any initial prompt s_0. Therefore, we can use the chain rule and get that when sampling n successive sentences, the KL-divergence between $\mathbb{P}_{\_}$ and $\mathbb{P}_{+}$ is at least $n \cdot \beta$. Consequently, we show that for any n there exists a textual prompt $s^\star$ consisting of n sentences, such that the likelihood of $s^\star$ according to $\mathbb{P}_{\_}$ is exponentially (both in β and n) more likely than the likelihood of $s^\star$ according to $\mathbb{P}_{+}$. Finally, note that during the evaluation of the expected behavior scoring, such exponential differences between the likelihood of $s^\star$ according to the different mixture components reweight theirs priors. We show that the contribution of $\mathbb{P}_{+}$ to the behaviour of the prompted LLM $\mathbb{P}$ is negligible.

Distinguishability exposes LLMs to shorter misaligning prompts Theorem 1 showcases that the larger the distinguishability between the ill-and well-behaved components β, the shorter the guaranteed misaligning prompt. The theorem demonstrates that this distinguishability can have more effect on the misaligning prompt length than the initial alignment of the LLM does. In particular, theorem 2 guarantees a range of prompt lengths for which a prompt that misaligns LLM-1 exists, while a prompt that misaligns LLM-2 does not, even though LLM-1 starts out as more aligned than LLM-2 before prompting. This can occur when LLM-1 has high enough distinguishability between its ill- and well-behaved components:

Theorem 2. *Let $\gamma \in [-1, 0)$ and $\alpha_1, \alpha_2, \beta_1, c_2, \Gamma, \epsilon > 0$. Let B be a behavior and $\mathbb{P}_1$, $\mathbb{P}_2$, two LLM distributions such that $0 < B_{\mathbb{P}_2} < B_{\mathbb{P}_1}$. Suppose that B is $\alpha_1, \beta_1, \gamma$-negatively distinguishable in $\mathbb{P}_1$, and B is $\alpha_2, 0, \Gamma$-positively distinguishable in $\mathbb{P}_2$, with the negative component being c-similar to the positive component. Then if $\beta_1 = \Omega\left(c_2 \cdot (\log \frac{1}{\epsilon} + \log \frac{1}{\alpha_1}) / \log \frac{\alpha_2 - 1}{\alpha_2}\right)$, there exists a prompt s of length $O(\frac{log \frac{1}{\alpha_1} + log \frac{1}{\epsilon}}{\beta_1})$, such that $B_{\mathbb{P}_1} < \gamma + \epsilon$, while **for any prompt s' of that length**, $B_{\mathbb{P}_2}(s') > \Gamma/2$.*

The above corollary demonstrates that β-distinguishability between ill- and well-behaved components can expose the LLM to shorter misaligning prompts. In the following we conjecture that this might have implications on leading alignment methods.

Conjecture on relation of distinguishability to RLHF Leading alignment tuning practices such as RLHF train the LLM to maximize the likelihood of desired sentences and minimizes the likelihood of undesired ones. The following conjecture implies that the leading practice of RLHF can make the two components more β-distinguishable (definition 2), which according to theorem 2 may render the resulting LLM prone to shorter adversarial attacks via prompting:

Conjecture 1. *An alignment loss that increases the likelihood of desired sentences and minimizes the likelihood of undesired ones, increases the β-distinguishability of resulting aligned LLM.*

The intuition behind this conjecture is that alignment tuning induces separability between desired and undesired behaviors in the LLM representation space, and thus the LLM can serve as a basis for a classifier Nachum & Yang (2021); Saunshi et al. (2021); Ge et al. (2023). For sentence s that is misclassified as good by the pretrained LLM but correctly classified as bad after alignment tuning, $\mathbb{P}_{-}^{\text{RLHF}}(s) > \mathbb{P}_{-}^{\text{pretraining}}(s)$, while $\mathbb{P}_{+}^{\text{RLHF}}(s) < \mathbb{P}_{+}^{\text{pretraining}}(s)$. Therefore, the contribution of this classification change to the KL divergence is positive since:

$$\Delta KL = \mathbb{P}_{-}^{\text{RLHF}}(s) \cdot log\frac{\mathbb{P}_{-}^{\text{RLHF}}}{\mathbb{P}_{+}^{\text{RLHF}}} - \mathbb{P}_{-}^{\text{pretraining}}(s) \cdot log\frac{\mathbb{P}_{-}^{\text{pretraining}}}{\mathbb{P}_{+}^{\text{pretraining}}} > 0 \tag{6}$$

Thus, increased KL-divergence between the ill- and well- behaved components is implied, increasing their β-distinguishability. Though intuitive, we leave it as an open conjecture for follow up work. If correct, this may be the mechanism behind the empirical findings of Perez et al. (2022), who unveil that undesired behaviors more easily emerge as the LLM undergoes more RLHF training steps.

3.2 Extensions: aligning prompts and conversations

Misaligning in the presence of preset aligning prompts A common practice for enhancing positive behavior is to include an initial 'preset aligning prompt', denoted s_0 below, hard coded as a prefix to the LLM's input. The theorem below states that even in the presence of s_0, it is possible to prompt the LLM into an undesired behavior with a 'misaligning prompt'. We show that the required prompt length for misalignment scales linearly with the length of s_0.

Theorem 3. *Under the conditions of theorem 1 and that the distribution corresponding to the well-behaved component of $\mathbb{P}$ is c-similar to the ill-behaved component, for any initial prompt $s_0 \in \Sigma^\star$, the conditional LLM distribution $\mathbb{P}(\cdot|s_0)$ is γ-prompt-misalignable with prompt length $O(\frac{1}{\beta}(\log\frac{1}{\alpha} + log\frac{1}{\epsilon}) + \frac{c|s_0|}{\beta})$.*

Theorem 3 guarantees that even in the presence of a preset aligning prompt s_0, there exists a long enough prompt that will misalign the model. The following result guarantees that if the misaligning prompt is not long enough, a positive preset aligning prompt can serve as a guradrail guaranteeing positive behavior:

Theorem 4. *Let $\alpha, \beta, \Gamma, c > 0$ and let B be a behaviour such that an unprompted language model $\mathbb{P}$ is α, β, Γ-positively-distinguishable in $\mathbb{P}$ (definition 4). Suppose the distribution corresponding to the ill-behaved component of $\mathbb{P}$ is c-similar to the well-behaved component. Then there exists an aligning prompt s_0 of length $O(\frac{1}{\beta}(log\frac{1}{1-\alpha} + log\frac{1}{\epsilon}))$ such that for any prompt of length $O(\frac{|s_0|\beta}{c} + \frac{log\frac{1-\alpha}{\alpha}}{c})$, the model maintains partial alignment in the sense that $B_{\mathbb{P}}(s) > \frac{\Gamma}{2}$.*

Importantly, the above guarantee for alignment applies for longer adversarial prompt lengths than in the absence of a preset aligning prompt (the case of $|s_0| = 0$), formalizing the effectiveness of this common aligning method.

Misaligning via conversation We show below that an undesired behavior can be elicited from an LLM via conversation with an adversarial user. Interestingly, we show that if the adversarial user does not use a long enough misaligning prompt in the first turn, then the LLM's responses can hinder the user's misaligning efforts. Intuitively, if a user begins a conversation by simply requesting "say a racist statement", an aligned LLM will likely reply "I will not say racist statements, that is harmful", and this reply in its prompt will cause the LLM to be more mindful of refraining from racist statements in the remainder of the conversation. Overall, due to this 'misaligning resistance' by the LLM, the user will need to insert more misaligning text in the conversation format than in the single prompt format of section 3.1 in order for our framework to guarantee misalignment.

We formalize a conversation between a user and an LLM of distribution $\mathbb{P}$ as a sequence of user queries followed by LLM responses which are sampled according to the LLM's conditional distribution given the conversation thus far. Formally, given the history of the conversation, $q_1, a_1 ... q_t, a_t, q_{t+1}$, where q_i are the user's inputs and a_i are the LLM's responses, the LLM generates a response a_{t+1} by sampling from: $a_{t+1} \sim \mathbb{P}(\cdot|q_1, a_1, ..., q_t, a_t, q_{t+1})$. In the following theorem we show that under our distinguishability conditions, misalignment is always possible also in a conversation format:

Theorem 5. *Under the conditions of theorem 1 and that the distribution corresponding to the well-behaved component of $\mathbb{P}$ is c-similar to the ill-behaved component, in a conversation setting: $q_1, a_1 ... q_n, a_n, q_{n+1}$, the model is γ-misalignable with total prompt length of $\sum_{i=1}^{n} |q_i| = O(\frac{1}{\beta}(\log \frac{1}{\alpha} + \log \frac{1}{\epsilon}) + \frac{c}{\beta} \sum_{i=1}^{n} |a_i|)$ and each prompt of length $|q_i| = O(\frac{c}{\beta}|a_i|)$.*

Comparing the above requirement on the amount of misaligning text to be inserted by an adversarial user to that required in the single prompting scenario of theorem 1, we see that it is larger by a factor of $n \cdot c$, where n is the number of conversation turns. Intuitively, in the beginning of the conversation the model is aligned, so it is most likely that its response will be sampled from the well-behaved component, thus enhancing it over the ill-behaved component (see the proof of theorem 5 in appendix F for formalization of this intuition).

Conversing with a prompted LLM Theorem 5 guarantees that LLMs can be misaligned via a conversation. We show in the following corollary of theorems 4 and 5, that once an LLM has been either aligned or misaligned via an initial preset prompt, it is guaranteed to exhibit either the positive or the negative behavior of this prompt *if the conversation is short enough*:

Corollary 1. *Under the conditions of theorem 1 (4), there exists a misaligning (aligning) prompt s_0 of length $O(\frac{1}{\beta}(log\frac{1}{\alpha} + log\frac{1}{\epsilon}))$ $\big(O(\frac{1}{\beta}(log\frac{1}{1-\alpha} + log\frac{1}{\epsilon}))\big)$, such that for the remainder of the conversation, $a_1, q_1, ..., a_{n-1}, q_n$, it will remain misaligned (aligned):*

$$B_{\mathbb{P}}(s_0, a_1, q_1, ..., a_{n-1}, q_n) < \frac{\gamma}{2} \qquad \big(B_{\mathbb{P}}(s_0, a_1, q_1, ..., a_{n-1}, q_n) > \frac{\Gamma}{2}\big) \tag{7}$$

Unless $\sum_{i=1}^{n} |q_i| + |a_i| = \Omega(\frac{\beta}{c} \cdot |s_0|)$.

The practice of aligning chatbots via an initial preset aligning prompts is very common. The above corollary shows that the length of the aligning prompt plays a key factor in determining for how long the chatbot will adhere to the alignment of the initial prompt.

3.3 Imitating personas as a "jailbreak" for LLM alignment

Recent findings show that LLMs can be misaligned via a mechanism of prompting the LLM into behaving as a persona it has clearly captured during pretraining (*e.g.*, (Nardo, 2023)). In this subsection, we use our definition of "persona", presented in section 2.1.2, in order to show that this adversarial misaligning strategy can be more efficient than directly attempting to elicit the undesired behavior.

We first prove that if a distribution can be written as a mixture of personas, $\mathbb{P} = \sum_{\phi \in \Phi} w_\phi \mathbb{P}_\phi$ and there exists a persona that is ill-behaved $B_{\mathbb{P}_\phi} \leq \gamma$ and is β-distinguishable from all other personas, then there exists a prompt which causes the LLM to exhibit the ill-behaved persona's behavior:

Theorem 6. *Let $\gamma \in [-1, 0)$, $\alpha, \beta, c, \epsilon > 0$, and let $\mathbb{P}$ be a mixture of personas $\mathbb{P} = \sum_{\phi \in \Phi} w_\phi \mathbb{P}_\phi$. Then for every behavior B that is α, β, γ-distinguishable in persona mixture $\mathbb{P}$ (definition 5), with all personas being c-similar to the negative persona, the distribution $\mathbb{P}$ is γ-prompt-misalignable (definition 1) with prompt of length $O\big(\max\big\{\frac{1}{\beta}(\log\frac{1}{\epsilon} + \log\frac{1}{\alpha}), \frac{c^2}{\beta^2}\log|\Phi|\big\}\big)$.*

The prompt length comprises of two terms. The first corresponds to the misaligning prompt length as in theorem 1, while the second term corresponds to a union bound which ensures the that the enhanced persona dominates the rest of the personas simultaneously.

Imitation of personas for "Jailbreaking" The consequence of the above theorem is that personas ϕ with low priors w_ϕ in the persona mixture may compensate for this with high distinguishability β, such that in some cases, prompting the model for a low-weight high-distinguishability persona may be more efficient at triggering bad behavior than a high-weight low-distinguishability bad component. This is expected to happen if a persona is very well captured by the LLM during pretraining.

Corollary 2. *Let $\gamma \in [-1, 0)$, $\alpha_{\tilde{\phi}}, \beta_{\tilde{\phi}}, c_{\tilde{\phi}} > 0$, let $\mathbb{P}$ be a mixture of personas $\mathbb{P} = \sum_{\phi \in \Phi} w_\phi \mathbb{P}_\phi$ such that all personas are $c_{\tilde{\phi}}$-similar to the negative behavior persona, and B a behavior that is $\alpha_{\tilde{\phi}}, \beta_{\tilde{\phi}}, \gamma$-distinguishable in persona mixture $\mathbb{P}$, where the distinguishable persona is denoted by $\tilde{\phi}$. Denote the two-component coarse grained persona mixture as $\mathbb{P} = \alpha\mathbb{P}_- + (1-\alpha)\mathbb{P}_+$, where*

$\mathbb{P}_- \propto \sum_{\{\phi \in \Phi | B_\phi < \gamma\}} w_\phi \mathbb{P}_\phi$ *and* $\mathbb{P}_+ \propto \sum_{\{\phi \in \Phi | B_\phi \geq \gamma\}} w_\phi \mathbb{P}_\phi$*. Then if* $\mathbb{P}_-$ *is* β*-distinguishable from* $\mathbb{P}_+$ *such that* $\beta_{\tilde{\phi}} > \beta \cdot \frac{log \frac{1}{w_{\tilde{\phi}}}}{log \frac{1}{\alpha}}$ *and* $\beta_{\tilde{\phi}} > \beta \cdot \sqrt{\frac{c_{\tilde{\phi}}^2}{\beta_{\tilde{\phi}}} \frac{\log |\Phi|}{log \frac{1}{\alpha}}}$*, then adversarial misaligning prompts that evoke the negative persona in the multi-component mixture are asymptotically shorter by a factor of* $O\left(\min \left\{ \frac{\beta}{\beta_{\tilde{\phi}}} \frac{log \frac{1}{w_{\tilde{\phi}}}}{log \frac{1}{\alpha}}, \frac{\beta}{\beta_{\tilde{\phi}}} \sqrt{\frac{c_{\tilde{\phi}}^2}{\beta_{\tilde{\phi}}} \frac{\log |\Phi|}{log \frac{1}{\alpha}}} \right\} \right)$ *than those that evoke the negative component in the two-component mixture.*

The persona's distinguishability $\beta_{\tilde{\phi}}$ quantifies how well the persona was captured by the LLM during pretraining, and the corollary imposes two conditions on it. The first is that it needs to be large enough to compensate for the persona $\tilde{\phi}$ having a lower prior than the large negative component $\mathbb{P}_-$. The second condition corresponds to a union bound allowing the persona $\tilde{\phi}$ to dominate over the rest of the personas in the mixture. Thus, in cases where an LLM captures a toxic persona very well during pretraining, it can be more efficient to prompt the LLM to imitate it rather than enhancing the ill-behaved component directly.

4 Discussion

The need for robust methods for AI alignment is pressing. Prominent actors in our field are advocating for halting LLM development until the means of controlling this technology are better understood (O'Brien, 2023). This paper brings forward the Behavior Expectation Bounds (BEB) theoretical framework, which is aimed at providing means for discussing core alignment issues in leading contemporary interactions between humans and LLMs.

We used the BEB framework in order to make several fundamental assertions regarding alignment in LLMs. First, we showed that any realistic alignment process can be reversed via an adversarial prompt or conversation with an adversarial user. As a silver lining, we showed that the better aligned the model is to begin with, the longer the prompt required to reverse the alignment, so limited prompt lengths may serve as guardrails in theory. With that, we also show that this picture is more complex, and the distinguishability of undesired behavior components also facilitates easier misalignment. Thus, while attenuating undesired behaviors, the leading alignment practice of reinforcement learning from human feedback (RLHF) may also render these same undesired behaviors more easily accessible via adversarial prompts. We leave the latter statement as an open conjecture; this theoretical direction may explain the result in Perez et al. (2022), in which RLHF increases undesired behaviors in language models.

Our BEB framework allowed us to make several further statements regarding different aspects of LLM alignment, *e.g.*, guaranteeing that a misaligned LLM will remain misaligned for a certain duration of a conversation, showing that the practice of misaligning an LLM via a multi-turn conversation is more intricate and can be less efficient than misaligning via a single prompt (due to the aligned LLM "resisting" misalignment), and showing that invoking a well captured malicious persona can be an efficient "jailbreak" out of alignment.

Our framework has several limitations and we leave several issues open for future work. Andreas (2022) describe modern LLMs as comprised of distinct agents that manifest when the right prompt is inserted into the LLM. Our presented notions of decomposability into components and distinguishability between these components are one analyzable choice of modeling multiple agents or personas composing the LLM distribution. We showed that with this choice several theoretical statements can be made that fit empirical observations on misalignment via prompting. While intuitive and reinforced by embedding space clustering experiments in the appendix, we leave it to future work to (i) further investigate superposition and decomposability in actual LLM distributions and (ii) introduce more elaborate or more realistic assumptions on the manner in which agent or persona decomposition is manifested in actual LLM distributions, and use them to gain further theoretical insight on LLM alignment. Elucidating this picture also bears promise for new empirical methods for controlling ill-behaved components with actual LLMs. Furthermore, our framework assumes ground truth behavior scores per sentence, where in reality behavior scoring is more complex, *e.g.*, over varying text granularities, hard to define behavior verticals, and ambiguous scoring. A deeper linguistic definition of the behavior scoring setup may lead to new insights that can be drawn from the

BEB theoretical framework. Overall we hope that our presented theoretical framework for analyzing LLM alignment can serve as a basis for further advancement in understanding this important topic.

ACKNOWLEDGMENTS

This research was supported by the ERC (European Research Council) and the ISF (Israel Science Foundation).

REFERENCES

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016.

Jacob Andreas. Language models as agent models. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 5769–5779, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.findings-emnlp.423.

Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, et al. A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861*, 2021.

Imane El Atillah. Man ends his life after an ai chatbot 'encouraged' him to sacrifice himself to stop climate change. *Euronews*, 2023.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pp. 610–623, 2021.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023.

Ameet Deshpande, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, and Karthik Narasimhan. Toxicity in chatgpt: Analyzing persona-assigned language models. *arXiv preprint arXiv:2304.05335*, 2023.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

Jiawei Ge, Shange Tang, Jianqing Fan, and Chi Jin. On the provable advantage of unsupervised pretraining. *arXiv preprint arXiv:2303.01566*, 2023.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. RealToxicityPrompts: Evaluating neural toxic degeneration in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 3356–3369, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.301. URL https://aclanthology.org/2020.findings-emnlp.301.

Dan Hendrycks, Nicholas Carlini, John Schulman, and Jacob Steinhardt. Unsolved problems in ml safety. *arXiv preprint arXiv:2109.13916*, 2021.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Ben Hutchinson, Vinodkumar Prabhakaran, Emily Denton, Kellie Webster, Yu Zhong, and Stephen Denuyl. Social biases in NLP models as barriers for persons with disabilities. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5491–5501, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.487. URL https://aclanthology.org/2020.acl-main.487.

Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3214–3252, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.229. URL https://aclanthology.org/2022.acl-long.229.

Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, and Sayak Paul. Peft: State-of-the-art parameter-efficient fine-tuning methods. https://github.com/huggingface/peft, 2022.

AI Meta. Introducing llama: A foundational, 65-billion-parameter large language model. *Meta AI. https://ai. facebook. com/blog/large-language-model-llama-meta-ai*, 2023.

Ofir Nachum and Mengjiao Yang. Provable representation learning for imitation with contrastive fourier features. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 30100–30112. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/fd00d3474e495e7b6d5f9f575b2d7ec4-Paper.pdf.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1953–1967, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.154. URL https://aclanthology.org/2020.emnlp-main.154.

Cleo Nardo. The waluigi effect (mega-post). *Less Wrong*, 2023.

Richard Ngo. The alignment problem from a deep learning perspective. *arXiv preprint arXiv:2209.00626*, 2022.

Tomohiro Nishiyama. Lower bounds for the total variation distance given means and variances of distributions. *arXiv preprint arXiv:2212.05820*, 2022.

Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*, 2023.

Matt O'Brien. Musk, scientists call for halt to ai race sparked by chatgpt. *AP News*, 2023.

OpenAI. Gpt-4 technical report, 2023.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Gray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=TG8KACxEON.

Alexander Pan, Kush Bhatia, and Jacob Steinhardt. The effects of reward misspecification: Mapping and mitigating misaligned models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=JYtwGwIL7ye.

Joon Sung Park, Joseph C O'Brien, Carrie J Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. *arXiv preprint arXiv:2304.03442*, 2023.

Ethan Perez, Sam Ringer, Kamilė Lukošiūtė, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, et al. Discovering language model behaviors with model-written evaluations. *arXiv preprint arXiv:2212.09251*, 2022.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

Jack W Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, et al. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*, 2021.

Kevin Roose. A conversation with bing's chatbot left me deeply unsettled. *New York Times*, 2023.

Nikunj Saunshi, Sadhika Malladi, and Sanjeev Arora. A mathematical exploration of why language models help solve downstream tasks. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=vVjIW3sEc1s.

John Schulman, Barret Zoph, Christina Kim, Jacob Hilton, Jacob Menick, Jiayi Weng, Juan Felipe, Ceron Uribe, Liam Fedus, Luke Metz, Michael Pokorny, Rapha Gontijo Lopes, Shengjia Zhao, Arun Vijayvergiya, Eric Sigler, Adam Perelman, Chelsea Voss, Mike Heaton, Joel Parish, Dave Cummings, Rajeev Nayak, Valerie Balcom, David Schnurr, Tomer Kaftan, Chris Hallacy, Nicholas Turley, Noah Deutsch, Vik Goel, Jonathan Ward, Aris Konstantinidis, Wojciech Zaremba, Long Ouyang, Leonard Bogdonoff, Joshua Gross, David Medina, Sarah Yoo, Teddy Lee, Ryan Lowe, Dan Mossing, Joost Huizinga, Roger Jiang, Carroll Wainwright, Diogo Almeida, Steph Lin, Marvin Zhang, Kai Xiao, Katarina Slama, Steven Bills, Alex Gray, Jan Leike, Jakub Pachocki, Phil Tillet, Shantanu Jain, Greg Brockman, Nick Ryder, Alex Paino, Qiming Yuan, Clemens Winter, Ben Wang, Mo Bavarian, Igor Babuschkin, Szymon Sidor, Ingmar Kanitscheider, Mikhail Pavlov, Matthias Plappert, Nik Tezak, Heewoo Jun, William Zhuk, Vitchyr Pong, Lukasz Kaiser, Jerry Tworek, Andrew Carr, Lilian Weng, Sandhini Agarwal, Karl Cobbe, Vineet Kosaraju, Alethea Power, Stanislas Polu, Jesse Han, Raul Puri, Shawn Jain, Benjamin Chess, Christian Gibson, Oleg Boiko, Emy Parparita, Amin Tootoonchian, Kyle Kosic, and Christopher Hesse. Introducing chatgpt. *OpenAI blog*, 2023.

Shai Shalev-Shwartz, Shaked Shammah, and Amnon Shashua. On the ethics of building ai in a responsible manner. *arXiv preprint arXiv:2004.04644*, 2020.

Varshini Subhash. Can large language models change user preference adversarially? *arXiv preprint arXiv:2302.10291*, 2023.

Jessica Taylor, Eliezer Yudkowsky, Patrick LaVictoire, and Andrew Critch. Alignment for advanced machine learning systems. *Ethics of Artificial Intelligence*, pp. 342–382, 2016.

Pranav Narayanan Venkit, Mukund Srinath, and Shomir Wilson. A study of implicit bias in pretrained language models against people with disabilities. In *Proceedings of the 29th International Conference on Computational Linguistics*, pp. 1324–1332, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics. URL https://aclanthology.org/2022.coling-1.113.

Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. Universal adversarial triggers for attacking and analyzing NLP. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2153–2162, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1221. URL https://aclanthology.org/D19-1221.

Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, Courtney Biles, Sasha Brown, Zac Kenton, Will Hawkins, Tom Stepleton, Abeba Birhane, Lisa Anne Hendricks, Laura Rimell,

William Isaac, Julia Haas, Sean Legassick, Geoffrey Irving, and Iason Gabriel. Taxonomy of risks posed by language models. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pp. 214–229, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393522. doi: 10.1145/3531146.3533088. URL https://doi.org/10.1145/3531146.3533088.

Colin G West. Advances in apparent conceptual physics reasoning in gpt-4. *arXiv e-prints*, pp. arXiv–2303, 2023.

Noam Wies, Yoav Levine, and Amnon Shashua. The learnability of in-context learning. *arXiv preprint arXiv:2303.07895*, 2023.

Jing Xu, Da Ju, Margaret Li, Y-Lan Boureau, Jason Weston, and Emily Dinan. Bot-adversarial dialogue for safe conversational agents. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2950–2968, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.235. URL https://aclanthology.org/2021.naacl-main.235.

Dian Yu and Kenji Sagae. Automatically exposing problems with neural dialog models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 456–470, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.37. URL https://aclanthology.org/2021.emnlp-main.37.

Eliezer Yudkowsky. Creating friendly ai 1.0: The analysis and design of benevolent goal architectures. *The Singularity Institute, San Francisco, USA*, 2001.

A PROOFS BUILDING BLOCKS

In this section, we prove three technical lemmas which are the building blocks for proving our results. In subsection A.1 we prove that prompts can reweight the initial prior distribution of mixture components. In subsection A.2 we show that such reweighting alters the behaviour of the mixture distribution. And finally, in subsection A.3 we shows that under our α, β, γ-negative-distinguishability assumption, such prompts always exists.

A.1 CONVERGENCE TO A SINGLE COMPONENT

In this subsection, we prove a technical lemma which shows that when the likelihood of a prompt s_0 is relatively high according to a mixture component, then the conditional mixture distribution converges to the conditional distribution of that single component. Essentially, this lemma strengthening the analysis in theorem 1 of Wies et al. (2023), and formulate the role of prompts as reweighting of the prior distribution. In the next subsection, we will show that indeed our notion of convergence implies also the convergence of behaviors.

Lemma 1. *Let $\mathbb{P}$ be a mixture distribution that can be written as $\alpha\mathbb{P}_0 + (1-\alpha)\,\mathbb{P}_1$. Then for any initial prompt s_0 and any string s such that $\mathbb{P}_0(s|s_0) > 0$ the following holds:*

$$\left|\frac{\mathbb{P}\left(s \mid s_0\right)}{\mathbb{P}_0\left(s \mid s_0\right)} - 1\right| \leq \frac{1-\alpha}{\alpha} \cdot \frac{\mathbb{P}_1\left(s_0\right)}{\mathbb{P}_0\left(s_0\right)} \cdot \max\left\{\frac{\mathbb{P}_1\left(s \mid s_0\right)}{\mathbb{P}_0\left(s \mid s_0\right)}, 1\right\} \tag{8}$$

Intuitively, when $\mathbb{P}(s_0 \oplus s)$ is equals to $\mathbb{P}_0(s_0 \oplus s)$ theirs ratio is one, and we bound the deviation from these case. Note that our bound implicitly implies the following additive notion of convergence:

$$|\mathbb{P}(s_0 \oplus s) - \mathbb{P}_0(s_0 \oplus s)| \leq \frac{1-\alpha}{\alpha} \cdot \frac{\mathbb{P}_1\left(s_0\right)}{\mathbb{P}_0\left(s_0\right)} \tag{9}$$

Proof. We begin by explicitly writing the conditional likelihood of s given s_0:

$$\mathbb{P}\left(s \mid s_0\right) = \frac{\mathbb{P}\left(s_0 \oplus s\right)}{\mathbb{P}\left(s_0\right)} = \frac{\alpha\mathbb{P}_0\left(s_0 \oplus s\right) + (1-\alpha)\,\mathbb{P}_1\left(s_0 \oplus s\right)}{\alpha\mathbb{P}_0\left(s_0\right) + (1-\alpha)\,\mathbb{P}_1\left(s_0\right)} \tag{10}$$

Now since both $(1-\alpha)$ and $\mathbb{P}_1(s_0 \oplus s)$ are greater than zero, we can bound $\mathbb{P}\left(s \mid s_0\right)$ from below by removing these terms from the numerator and get that:

$$\mathbb{P}\left(s \mid s_0\right) \geq \frac{\alpha\mathbb{P}_0\left(s_0 \oplus s\right)}{\alpha\mathbb{P}_0\left(s_0\right) + (1-\alpha)\,\mathbb{P}_1\left(s_0\right)} \tag{11}$$

Which after division of both the numerator and the denominator by $\alpha \cdot \mathbb{P}_0\left(s_0 \oplus s\right)$ is equals to:

$$\mathbb{P}_0\left(s \mid s_0\right) \cdot \left(1 + \frac{1-\alpha}{\alpha} \cdot \frac{\mathbb{P}_1\left(s_0\right)}{\mathbb{P}_0\left(s_0\right)}\right)^{-1} \tag{12}$$

Now, since $\frac{1}{1+x} \geq 1-x$ for any $x \geq 0$, we gets that $\mathbb{P}\left(s \mid s_0\right)$ is greater than:

$$\mathbb{P}_0\left(s \mid s_0\right) \cdot \left(1 - \frac{1-\alpha}{\alpha} \cdot \frac{\mathbb{P}_1\left(s_0\right)}{\mathbb{P}_0\left(s_0\right)}\right) \tag{13}$$

Finally, we divide the inequality by $\mathbb{P}_0(s|s_0)$ and subtracts 1 to get one side of equation's 8 inequality:

$$\frac{\mathbb{P}\left(s \mid s_0\right)}{\mathbb{P}_0\left(s \mid s_0\right)} - 1 \geq -\frac{1-\alpha}{\alpha} \cdot \frac{\mathbb{P}_1\left(s_0\right)}{\mathbb{P}_0\left(s_0\right)} \tag{14}$$

Moving to the other side of the inequality, since both $(1-\alpha)$ and $\mathbb{P}_1\left(s_0 \oplus s\right)$ are greater than zero, we can bound $\mathbb{P}\left(s \mid s_0\right)$ from above by removing these terms from the denominator and get that :

$$\mathbb{P}\left(s \mid s_0\right) = \frac{\alpha\mathbb{P}_0\left(s_0 \oplus s\right) + (1-\alpha)\,\mathbb{P}_1\left(s_0 \oplus s\right)}{\alpha\mathbb{P}_0\left(s_0\right) + (1-\alpha)\,\mathbb{P}_1\left(s_0\right)} \leq \frac{\alpha\mathbb{P}_0\left(s_0 \oplus s\right) + (1-\alpha)\,\mathbb{P}_1\left(s_0 \oplus s\right)}{\alpha\mathbb{P}_0\left(s_0\right)} \tag{15}$$

Which after division of both the numerator and the denominator by $\alpha \cdot \mathbb{P}_0(s_0)$ is equals to:

$$\frac{\alpha \mathbb{P}_0(s_0 \oplus s)}{\alpha \mathbb{P}_0(s_0)} + \frac{(1-\alpha)\mathbb{P}_1(s_0 \oplus s)}{\alpha \mathbb{P}_0(s_0)} = \mathbb{P}_0(s \mid s_0) + \frac{(1-\alpha)\mathbb{P}_1(s_0 \oplus s)}{\alpha \mathbb{P}_0(s_0)} \tag{16}$$

Now, we can use the fact that $\mathbb{P}_1(s_0 \oplus s_1) = \mathbb{P}_1(s_0) \cdot \mathbb{P}_1(s \mid s_0)$ to get that $\mathbb{P}(s \mid s_0)$ is at most:

$$\mathbb{P}_0(s \mid s_0) + \frac{(1-\alpha)\mathbb{P}_1(s_0)\mathbb{P}_1(s \mid s_0)}{\alpha \mathbb{P}_0(s_0)} \tag{17}$$

Which after division by $\mathbb{P}_0(s|s_0)$ and subtraction of 1 yield the other side of equation's 8 inequality:

$$\frac{\mathbb{P}(s \mid s_0)}{\mathbb{P}_0(s|s_0)} - 1 \leq \frac{\mathbb{P}_1(s_0)}{\mathbb{P}_0(s_0)} \frac{(1-\alpha)\mathbb{P}_1(s|s_0)}{\alpha \mathbb{P}_0(s|s_0)} \tag{18}$$

Finally, combining both inequalities yields equation 8.

□

A.2 Behavioral implication of the convergence to a single component

In this subsection, we prove a technical lemma which shows that when the likelihood of a prompt s_0 is relatively high according to a mixture component, then the conditional mixture distribution converge to the conditional distribution of that single component. In the next sections, we will use this lemma to prove the theorems from the main text.

Lemma 2. *Let B be a behaviour, then under the conditions of lemma 1 the following holds:*

$$|B_{\mathbb{P}}(s_0) - B_{\mathbb{P}_0}(s_0)| \leq 2 \cdot \frac{1-\alpha}{\alpha} \cdot \frac{\mathbb{P}_1(s_0)}{\mathbb{P}_0(s_0)} \tag{19}$$

Proof. To begin, we explicitly write the expectations difference:

$$|B_{\mathbb{P}}(s_0) - B_{\mathbb{P}_0}(s_0)| = \left| \sum_s B(s) \cdot [\mathbb{P}(s \mid s_0) - \mathbb{P}_0(s \mid s_0)] \right| \tag{20}$$

Which by the triangular inequality is at most:

$$\leq \sum_s |B(s)| \cdot |\mathbb{P}(s \mid s_0) - \mathbb{P}_0(s \mid s_0)| \tag{21}$$

Now, since the range of B is $[-1, 1]$ we can get rid of the $|B(s)|$ terms, and get that $|B_{\mathbb{P}}(s_0) - B_{\mathbb{P}_0}(s_0)|$ is at most:

$$\sum_s |\mathbb{P}(s \mid s_0) - \mathbb{P}_0(s \mid s_0)| = \sum_s \mathbb{P}_0(s \mid s_0) \cdot \left| \frac{\mathbb{P}(s \mid s_0)}{\mathbb{P}_0(s \mid s_0)} - 1 \right| \tag{22}$$

Importantly, by lemma 1 we have that:

$$\left| \frac{\mathbb{P}(s \mid s_0)}{\mathbb{P}_0(s \mid s_0)} - 1 \right| \leq \frac{1-\alpha}{\alpha} \cdot \frac{\mathbb{P}_1(s_0)}{\mathbb{P}_0(s_0)} \cdot \max\left\{ \frac{\mathbb{P}_1(s \mid s_0)}{\mathbb{P}_0(s \mid s_0)}, 1 \right\} \tag{23}$$

For any s, hence we got that $|B_{\mathbb{P}}(s_0) - B_{\mathbb{P}_0}(s_0)|$ is at most:

$$\frac{1-\alpha}{\alpha} \cdot \frac{\mathbb{P}_1(s_0)}{\mathbb{P}_0(s_0)} \cdot \left[\sum_s \mathbb{P}_0(s \mid s_0) \cdot \max\left\{ \frac{\mathbb{P}_1(s \mid s_0)}{\mathbb{P}_0(s \mid s_0)}, 1 \right\} \right] \tag{24}$$

$$\leq \frac{1-\alpha}{\alpha} \cdot \frac{\mathbb{P}_1(s_0)}{\mathbb{P}_0(s_0)} \cdot \sum_s (\mathbb{P}_1(s \mid s_0) + \mathbb{P}_0(s \mid s_0)) \tag{25}$$

where the last inequality follows from the fact that sum of two non-negative terms is greater than the maximum of the terms. Finally, since both $\mathbb{P}_0(s \mid s_0)$ and $\mathbb{P}_1(s \mid s_0)$ are probability distributions, summing over all possible sentences s yields 2, and hence the inequality in equation 19 follows.

□

A.3 ADVERSARIAL PROMPT CONSTRUCTION

In this subsection, we prove a technical lemma which shows that when two distribution are sufficiently distinguishable (see definition 2 from the main text) ,then there exists a prompt such that the ratio of the prompt's likelihood according to these two distribution is arbitrary low. In the next sections we will use this lemma to prove the existence adversarial prompt for which the conditions of lemma 1 holds. And hence an adversarial user might alter the model behavior (lemma 2).

Lemma 3. *Let $\beta, c, \epsilon > 0$, s_0 a prefix and $\mathbb{P}_0, \mathbb{P}_1$ two distributions. Suppose $\mathbb{P}_0$ is β-distinguishable from $\mathbb{P}_1$, and $\mathbb{P}_1$ is c-similar to $\mathbb{P}_0$, then there exists a prompt s of length $\frac{1}{\beta} \cdot \left(\log \frac{1}{\epsilon} + c \cdot |s_0|\right)$ such that the following holds:*

$$\frac{\mathbb{P}_1(s_0 \oplus s)}{\mathbb{P}_0(s_0 \oplus s)} \leq \epsilon \tag{26}$$

Moreover, when s_0 is an empty string, the above still hold even when $\mathbb{P}_1$ is not c-similar to $\mathbb{P}_0$.

Proof. Intuitively, given s_0, we use the fact that $\mathbb{P}_0$ is β-distinguishable from $\mathbb{P}_1$ to construct a prompt sentence by sentence, and get a prompt $q = s_1 \oplus ... \oplus s_{|q|}$ such that:

$$\log \frac{\mathbb{P}_0\left(s_1 \oplus ... \oplus s_k \mid s_0\right)}{\mathbb{P}_1\left(s_1 \oplus ... \oplus s_k \mid s_0\right)} > \beta \cdot k \tag{27}$$

For any $k \leq |q|$.

Formally, we will prove the existence[2] of such a prompt by induction. Starting from the base case, the fact that $\mathbb{P}_0$ is β-distinguishable from $\mathbb{P}_1$ assure us that:

$$\mathbb{E}_{s_1 \sim \mathbb{P}_0(\cdot|s_0)}\left[\log \frac{\mathbb{P}_0\left(s_1 \mid s_0\right)}{\mathbb{P}_1\left(s_1 \mid s_0\right)}\right] > \beta \tag{28}$$

Thus, in particular there exists a sentence s_1 that satisfies $\log \frac{\mathbb{P}_0(s_1 \mid s_0)}{\mathbb{P}_1(s_1 \mid s_0)} > \beta$. Now, assume that there exists sentences $s_1, \ldots, s_{k-1}$ such that the inequality in equation 27 holds. Then, again the fact that $\mathbb{P}_0$ is β-distinguishable from $\mathbb{P}_1$ assure us that:

$$\mathbb{E}_{s_k \sim \mathbb{P}_0(s_1 \oplus ... \oplus s_{k-1})}\left[\log \frac{\mathbb{P}_0\left(s_k \mid s_1 \oplus ... \oplus s_{k-1}\right)}{\mathbb{P}_1\left(s_k \mid s_1 \oplus ... \oplus s_{k-1}\right)}\right] > \beta \tag{29}$$

Hence, in particular there exists a sentence s_k that satisfies

$$\log \frac{\mathbb{P}_0\left(s_k \mid s_1 \oplus ... \oplus s_{k-1}\right)}{\mathbb{P}_1\left(s_k \mid s_1 \oplus ... \oplus s_{k-1}\right)} > \beta \tag{30}$$

Finally, we can use the chain rule for conditional probabilities, and get by the induction hypothesis that:

$$\log \frac{\mathbb{P}_0\left(q \mid s_0\right)}{\mathbb{P}_1\left(q \mid s_0\right)} = \sum_{i=1}^{k} \log \frac{\mathbb{P}_0\left(s_i \mid s_1 \oplus ... \oplus s_{i-1}\right)}{\mathbb{P}_1\left(s_i \mid s_1 \oplus ... \oplus s_{i-1}\right)} > k \cdot \beta \tag{31}$$

As desired, so the the induction hypothesis follows. Equivalently, we might flip the numerator and the denominator and returning to linear space:

$$\frac{\mathbb{P}_1\left(q \mid s_0\right)}{\mathbb{P}_0\left(q \mid s_0\right)} \leq \exp\left(-\beta \cdot |q|\right) \tag{32}$$

Now, when s_0 is an empty string we can choose $|q| > \frac{\log \frac{1}{\epsilon}}{\beta}$ to obtain the desired result that

$$\frac{\mathbb{P}_1(s_0 \oplus s)}{\mathbb{P}_0(s_0 \oplus s)} = \frac{\mathbb{P}_1(s)}{\mathbb{P}_0(s)} = \frac{\mathbb{P}_1(s \mid s_0)}{\mathbb{P}_0(s \mid s_0)} \leq \epsilon \tag{33}$$

[2]Note that there might be shorter prompts such that $\frac{\mathbb{P}_1(s_0 \oplus s)}{\mathbb{P}_0(s_0 \oplus s)} \leq \epsilon$.

Otherwise, denote by $s_{0,1}, \ldots, s_{0,|s_0|}$ the sentences in s_0. Then, by the chain rule for conditional probabilities we have that:

$$\frac{\mathbb{P}_1(s_0)}{\mathbb{P}_0(s_0)} = \prod_{i=1}^{|s_0|} \frac{\mathbb{P}_1(s_{0,i} \mid s_{0,i-1} \oplus \cdots \oplus s_{0,1})}{\mathbb{P}_0(s_{0,i} \mid s_{0,i-1} \oplus \cdots \oplus s_{0,1})} \tag{34}$$

Now, the fact that $\mathbb{P}_1$ is c-similar to $\mathbb{P}_0$ assure us that each term in the multiplication is at most e^c. Hence, we conclude that $\frac{\mathbb{P}_1(s_0)}{\mathbb{P}_0(s_0)} < \exp\left(c \cdot |s_0)\right|$. Finally, we can combine this inequality with equation 32 and get that:

$$\frac{\mathbb{P}_1(s_0 \oplus s)}{\mathbb{P}_0(s_0 \oplus s)} = \frac{\mathbb{P}_1(s_0)}{\mathbb{P}_0(s_0)} \cdot \frac{\mathbb{P}_1(s \mid s_0)}{\mathbb{P}_0(s \mid s_0)} < \exp\left(c \cdot |s_0| - \beta \cdot |q|\right) \tag{35}$$

Thus we can choose $|q| > \frac{c \cdot |s_0| + \log \frac{1}{\epsilon}}{\beta}$ to obtain that $\frac{\mathbb{P}_1(s_0 \oplus s)}{\mathbb{P}_0(s_0 \oplus s)} \leq \epsilon$ as desired. □

B PROOF OF THEOREM 1 AND 3

Since theorem 1 is a special case of theorem 3 where s_0 is an empty string, we will only prove theorem 2 and note that the usage of the additional assumption of c-similarity is necessary only for the usage of lemma 3 when s_0 is not an empty string.

Let $\mathbb{P}_+$ and $\mathbb{P}_-$ be the well-behaved and ill-behaved mixture components from the α, β, γ-negative-distinguishability definition, and let s_0 be an initial prompt and $\epsilon > 0$. Then since $\mathbb{P}_-$ is β-distinguishable from $\mathbb{P}_+$, lemma 3 assures us that for any $\epsilon' > 0$ there exists a sequence s_1 of $\frac{1}{\beta} \cdot \left(\log \frac{1}{\epsilon'} + c \cdot |s_0|\right)$ sentences such that $\frac{\mathbb{P}_+(s_0 \oplus s_1)}{\mathbb{P}_-(s_0 \oplus s_1)} < \epsilon'$. Now, lemma 2 assures us that for such prompt a $s_0 \oplus s_1$, the behaviour of $\mathbb{P}$ will converge to the ill-behaved component in the following sense:

$$\left|B_{\mathbb{P}}\left(s_0 \oplus s_1\right) - B_{\mathbb{P}_-}\left(s_0 \oplus s_1\right)\right| \leq 2 \cdot \frac{1-\alpha}{\alpha} \cdot \epsilon^{'} \tag{36}$$

Thus, we can choose $\epsilon' < \frac{\alpha \cdot \epsilon}{4}$ and get that:

$$\left|B_{\mathbb{P}}\left(s_0 \oplus s_1\right) - B_{\mathbb{P}_-}\left(s_0 \oplus s_1\right)\right| < \epsilon \tag{37}$$

Finally, by definition we have that $B_{\mathbb{P}_-}\left(s_0 \oplus s_1\right) \leq \gamma$ and hence we get that

$$B_{\mathbb{P}}\left(s_0 \oplus s_1\right) < \gamma + \epsilon \tag{38}$$

As desired.

C ZERO-SHOT PROMPT LENGTH GUARDRAIL

An additional results of our framework, is that an aligned LLM is guaranteed to remain aligned for all adversarial prompts *if the adversarial prompt length is sufficiently limited*, *i.e.*, for short enough prompt lengths it *can not* be misaligned via prompting:

Theorem 7. *Let $\alpha, \Gamma, c > 0$ and let B be a behaviour such that an unprompted language model $\mathbb{P}$ is α, β, Γ-positively-distinguishable in $\mathbb{P}$ (definition 4) for $\beta = 0$. Suppose the model is aligned w.r.t the behavior, $B_{\mathbb{P}} \geq \Gamma$ and that the distribution corresponding to the ill-behaved component of $\mathbb{P}$ is c-similar to the well-behaved component. Then for a prompt s of length $O\left({}^{log(1-\alpha/\alpha)}/_{c}\right)$, the model remains 'partially aligned',* i.e.*: $B_{\mathbb{P}}(s) > {}^{\Gamma}/_{2}$.*

This means that if a prompt is not sufficiently long, it will not misalign the model. Such a guarantee implies that limiting the prompt length can be used to ensure safety, and the dependence on α implies the the more the LLM is aligned to begin with, the longer the prompts that enjoy the safety guarantee.

Proof. Let $\mathbb{P}_+$ and $\mathbb{P}_-$ be the well-behaved and ill-behaved mixture components from the α, β, Γ-positively-distinguishability definition, and let s be an adversarial initial prompt. Then lemma 2

assures us that[3]:

$$\left|B_{\mathbb{P}}(s) - B_{\mathbb{P}_+}(s)\right| \leq 2 \cdot \frac{\alpha}{1-\alpha} \cdot \frac{\mathbb{P}_-(s)}{\mathbb{P}_+(s)} \tag{39}$$

Therefore, it is enough to prove that $\frac{\mathbb{P}_-(s)}{\mathbb{P}_+(s)} < \frac{1-\alpha}{\alpha} \cdot \frac{\Gamma}{4}$ for s of length $O\left(\frac{\log\left(\frac{1-\alpha}{\alpha}\right)}{c}\right)$ in order to guarantee that $B_{\mathbb{P}}(s) > \frac{\Gamma}{2}$ for such s. In addition, by the chain rule for conditional probability distributions, the fact that the ill-behaved mixture component $\mathbb{P}_-$ is c-similar to the well-behaved mixture component $\mathbb{P}_+$ assure us that:

$$\frac{\mathbb{P}_-(s_0)}{\mathbb{P}_+(s_0)} < \exp(|s| \cdot c) \tag{40}$$

Therefore, limiting the number of sentences in s to $\frac{\log\left(\frac{1-\alpha}{\alpha} \cdot \frac{\Gamma}{4}\right)}{c}$ assure us that $\frac{\mathbb{P}_-(s_0)}{\mathbb{P}_+(s_0)} < \frac{1-\alpha}{\alpha} \cdot \frac{\Gamma}{4}$ as desired. □

D PROOF OF THEOREM 2

Applying the theorem 1 to the LLM distribution represented by $\mathbb{P}_1$, we obtain that there exists a prompt of length $O(\frac{\log\frac{1}{\alpha_1}+\log\frac{1}{\epsilon}}{\beta_1})$, such that $B_{\mathbb{P}_1} < \gamma+\epsilon$. Applying theorem 7 to the LLM distribution represented by $\mathbb{P}_2$, assures that for any prompt of length $O(\frac{\log\frac{1}{\alpha_1}+\log\frac{1}{\epsilon}}{\beta_1})$, $B_{\mathbb{P}_1} > \Gamma/2$.

E PROOF OF THEOREM 4

Let $\mathbb{P}_+$ and $\mathbb{P}_-$ be the well-behaved and ill-behaved mixture components from the α, β, Γ-positively-distinguishability definition. Then since $\mathbb{P}_+$ is β-distinguishable from $\mathbb{P}_-$, lemma 3 assures us that for any $\epsilon' > 0$ there exists an aligning prompt s_0 of length $\frac{1}{\beta}\log\frac{1}{\epsilon'}$ such that $\frac{\mathbb{P}_-(s_0)}{\mathbb{P}_+(s_0)} < \epsilon'$. Let s_1 be the adversarial prompt, then the fact that the ill-behaved component $\mathbb{P}_-$ is c-similar to the well-behaved component $\mathbb{P}_+$ assure us that:

$$\frac{\mathbb{P}_-(s_1 \mid s_0)}{\mathbb{P}_+(s_1 \mid s_0)} < \exp(c \cdot |s_1|) \tag{41}$$

So overall, we got that:

$$\frac{\mathbb{P}_-(s_0 \oplus s_1)}{\mathbb{P}_+(s_0 \oplus s_1)} < \epsilon' \cdot \exp(c \cdot |s_1|) \tag{42}$$

Now, combining equation 42 with lemma 2 assures us that[4]:

$$\left|B_{\mathbb{P}}(s_0 \oplus s_1) - B_{\mathbb{P}_+}(s_0 \oplus s_1)\right| \leq 2 \cdot \frac{\alpha}{1-\alpha} \cdot \epsilon' \cdot \exp(c \cdot |s_1|) \tag{43}$$

Thus, we can choose $\epsilon' < \frac{\alpha}{1-\alpha} \cdot \frac{\Gamma}{4} \cdot \exp(-c \cdot |s_1|)$ and get that $B_{\mathbb{P}}(s_0 \oplus s_1) > \frac{\Gamma}{2}$. Finally, substituting this ϵ' into the guarantees of lemma 3 give us that the behaviour of the model is better than $\frac{\Gamma}{2}$ as long as the length the adversarial prompt s_1 is at most $\frac{\beta}{c} \cdot |s_0| + \frac{1}{c} \cdot \log\left(\frac{1-\alpha}{\alpha} \cdot \frac{4}{\Gamma}\right)$ since

$$|s_0| \geq -\frac{1}{\beta}\log\left(\frac{\alpha}{1-\alpha} \cdot \frac{\Gamma}{4} \cdot \exp(-c \cdot |s_1|)\right) \tag{44}$$

$$\iff |s_1| \leq \frac{\beta}{c} \cdot |s_0| + \frac{1}{c} \cdot \log\left(\frac{1-\alpha}{\alpha} \cdot \frac{4}{\Gamma}\right) \tag{45}$$

As desired.

[3]Note that the priors are flipped to $\alpha \leftrightarrow (1-\alpha)$ as we are now bounding the difference to the positive component.

[4]Note that the priors are flipped to $\alpha \leftrightarrow (1-\alpha)$ as we are now bounding the difference to the positive component.

F PROOF OF THEOREM 5

Let $\mathbb{P}_+$ and $\mathbb{P}_-$ be the well-behaved and ill-behaved mixture components from the α, β, γ-negative-distinguishability definition. Essentially, we we show that there exists a choice of prompts $q_1 ... q_{n+1}$ each of them consists of at most $O\left(\frac{c}{\beta}\right)$ sentences such that:

$$\log \frac{\mathbb{P}_+ \left(q_1 \oplus a_1 \oplus ... \oplus q_n \oplus a_n \oplus q_{n+1}\right)}{\mathbb{P}_- \left(q_1 \oplus a_1 \oplus ... \oplus q_n \oplus a_n \oplus q_{n+1}\right)} < \sum_{i=1}^{n+1} \left(c - \beta \cdot |q_i|\right) \tag{46}$$

Then, we will use lemma 2 and get that for any such prompts $q_1 ... q_{n+1}$ the behaviour of $\mathbb{P}$ will converge to the ill-behaved component in the following sense:

$$\left|B_{\mathbb{P}}\left(s\right) - B_{\mathbb{P}_-}\left(s\right)\right| \leq 2 \cdot \frac{1-\alpha}{\alpha} \cdot \exp\left(\sum_{i=1}^{n+1} \left(c - \beta \cdot |q_i|\right)\right) \tag{47}$$

Where s denote the concatenation of the messages in the conversation: $q_1 \oplus a_1 \oplus ... \oplus q_n \oplus a_n \oplus q_{n+1}$. Thus, we will get that $B_{\mathbb{P}}\left(s\right) < \gamma + \epsilon$ for $\sum_{i=1}^{n+1} |q_i| > \frac{(n+1)\cdot c + \log\left(\frac{1-\alpha}{2\cdot\alpha\cdot\epsilon}\right)}{\beta}$ as desired.

Intuitively, we will prove the existence of the prompts $q_1 ... q_{n+1}$ such that the length of any prompt is at most $O\left(\frac{c}{\beta}\right)$ and equation 46 upholds by using an induction argument, where the induction hypothesis follows from the fact that $\mathbb{P}_-$ is β-distinguishable from $\mathbb{P}_+$. Formally, the base case of the induction follows by using lemma 3 to construct an adversarial prompt q_1 such that $\log \frac{\mathbb{P}_+(q_1)}{\mathbb{P}_-(q_1)} < -\beta \cdot |q_1|$. Now, assume that there exists prompts $q_1 ... q_k$ such that the length of any prompt is at most $O\left(\frac{c}{\beta}\right)$ and equation 46 upholds (with $n = k-1$). Then, the proof of lemma 3 (equation 32) assure us that there exists an adversarial prompt q_{k+1} such that:

$$\log \frac{\mathbb{P}_+ \left(q_{k+1} | q_1 \oplus a_1 \oplus \cdots \oplus q_k \oplus a_k\right)}{\mathbb{P}_- \left(q_{k+1} | q_1 \oplus a_1 \oplus \cdots \oplus q_k \oplus a_k\right)} < -\beta \cdot |q_{k+1}| \tag{48}$$

Now, by the chain rule of conditional probabilities and the fact that $\mathbb{P}_+$ is c-similar to $\mathbb{P}_-$ we have that:

$$\log \frac{\mathbb{P}_+ \left(q_1 \oplus a_1 \oplus \cdots \oplus q_k \oplus a_k \oplus q_{k+1}\right)}{\mathbb{P}_- \left(q_1 \oplus a_1 \oplus \cdots \oplus q_k \oplus a_k \oplus q_{k+1}\right)} < c - \beta \cdot |q_{k+1}| + \log \frac{\mathbb{P}_+ \left(q_k | q_1 \oplus a_1 \oplus \cdots \oplus q_{k-1} \oplus a_{k-1}\right)}{\mathbb{P}_- \left(q_k | q_1 \oplus a_1 \oplus \cdots \oplus q_{k-1} \oplus a_{k-1}\right)} \tag{49}$$

So we can use the induction hypothesis to upper bound the $\log \frac{\mathbb{P}_+(q_k | q_1 \oplus a_1 \oplus \cdots \oplus q_{k-1} \oplus a_{k-1})}{\mathbb{P}_-(q_k | q_1 \oplus a_1 \oplus \cdots \oplus q_{k-1} \oplus a_{k-1})}$ term and get that:

$$\log \frac{\mathbb{P}_+ \left(q_1 \oplus a_1 \oplus \cdots \oplus q_k \oplus a_k \oplus q_{k+1}\right)}{\mathbb{P}_- \left(q_1 \oplus a_1 \oplus \cdots \oplus q_k \oplus a_k \oplus q_{k+1}\right)} < \sum_{i=1}^{k+1} \left(c - \beta \cdot |q_i|\right) \tag{50}$$

As desired.

G PROOF OF THEOREM 6

In this section we extends our analysis to mixture of more than two components. When looking at a decomposition of more than two components (so called-personas, presented in section 3.3, we ask whether such a decomposition can be leveraged by an adversarial user in order to evoke undesired behavior along a certain behavior vertical B. Contrary to the case of two components, which is one-dimensional in the sense that enhancing one component with a prompt reduces the other, the case of multiple components (so called-personas, presented in section 3.3) is multi-dimensional as we need to find a prompt that enhances one component over many others simultaneously. Hence, a-priory this does not amount to one component being distinguishable from all the rest as it requires a uniform bound on all the other components. In subsection G.1 we show that our definition of distinguishablty do imply the existence of such prompt. Then, in subsection G.2 we use this result and extend lemma 3 to the case of more than two components. In subsection G.3 we generalized lemma 1 to the case of more than two components. Finally, in subsection G.4 we shows that it implies a generalization of lemma 2 to the case of more than two components.

G.1 β-DISTINGUISHABILITY IMPLIES SUB-MARTINGALE

In this subsection we shows that our definition of distinguishablty implies a sub-Martingale property that we will use in order to prove that our definition of distinguishablty suffice for such uniform bound on all the other components.

Specifically, we will prove that β-distinguishability between $\mathbb{P}_{\phi^*}$ and $\mathbb{P}_{\phi}$ implies that the series $M_n = \log\left[\frac{\mathbb{P}_{\phi^*}(s_1 \oplus ... \oplus s_n)}{\mathbb{P}_{\phi}(s_1 \oplus ... \oplus s_n)}\right]$ is a sub-Martingale:

Lemma 4. *Let $\mathbb{P}_{\phi^*}$ and $\mathbb{P}_{\phi}$ be a probability distributions such that $\mathbb{P}_{\phi^*}$ is β-distinguishable from $\mathbb{P}_{\phi}$. Then the following hold:*

$$\mathbb{E}_{s_1 \oplus ... \oplus s_n \sim \mathbb{P}_{\phi^*}(\cdot)}[M_n | M_1 = m_1 ... M_{n-1} = m_{n-1}] > m_{n-1} + \beta \tag{51}$$

Proof. Given a series of $m_1, \ldots, m_{n_1}$ we need to calculate the conditional expectation of M_n. Now, for **any** event $s_1, \ldots, s_{n-1}$ such that $M_i = m_i$ for any $1 \leq i \leq n-1$, the fact that $\mathbb{P}_{\phi^*}$ is β-distinguishable from $\mathbb{P}_{\phi}$ assure us that:

$$\mathbb{E}_{s_n \sim \mathbb{P}_{\phi^*}(\cdot | s_1 \oplus ... \oplus s_{n-1})} \log\left(\frac{\mathbb{P}_{\phi^*}\left(s_n \mid s_1 \oplus ... \oplus s_{n-1}\right)}{\mathbb{P}_{\phi}\left(s_n \mid s_1 \oplus ... \oplus s_{n-1}\right)}\right) > \beta \tag{52}$$

In addition, by the chain rule of conditional probabilities we have that:

$$\log\left[\frac{\mathbb{P}_{\phi^*}\left(s_1 \oplus ... \oplus s_n\right)}{\mathbb{P}_{\phi}\left(s_1 \oplus ... \oplus s_n\right)}\right] = \frac{\mathbb{P}_{\phi^*}\left(s_n \mid s_1 \oplus ... \oplus s_{n-1}\right)}{\mathbb{P}_{\phi}\left(s_n \mid s_1 \oplus ... \oplus s_{n-1}\right)} + \log\left[\frac{\mathbb{P}_{\phi^*}\left(s_1 \oplus ... \oplus s_{n-1}\right)}{\mathbb{P}_{\phi}\left(s_1 \oplus ... \oplus s_{n-1}\right)}\right] \tag{53}$$

$$= \frac{\mathbb{P}_{\phi^*}\left(s_n \mid s_1 \oplus ... \oplus s_{n-1}\right)}{\mathbb{P}_{\phi}\left(s_n \mid s_1 \oplus ... \oplus s_{n-1}\right)} + m_{n-1} \tag{54}$$

Hence, we have that:

$$\mathbb{E}_{s_1 \oplus ... \oplus s_n \sim \mathbb{P}_{\phi^*}(\cdot)}[\log\left[\frac{\mathbb{P}_{\phi^*}\left(s_1 \oplus ... \oplus s_n\right)}{\mathbb{P}_{\phi}\left(s_1 \oplus ... \oplus s_n\right)}\right] M_1 = m_1 ... M_{n-1} = m_{n-1}] \tag{55}$$

$$= \mathbb{E}_{s_n \sim \mathbb{P}_{\phi^*}(\cdot | s_1 \oplus ... \oplus s_{n-1})} \log\left(\frac{\mathbb{P}_{\phi^*}\left(s_n \mid s_1 \oplus ... \oplus s_{n-1}\right)}{\mathbb{P}_{\phi}\left(s_n \mid s_1 \oplus ... \oplus s_{n-1}\right)}\right) + m_{n-1} \tag{56}$$

Thus, we conclude that

$$\mathbb{E}_{s_1 \oplus ... \oplus s_n \sim \mathbb{P}_{\phi^*}(\cdot)}[\log\left[\frac{\mathbb{P}_{\phi^*}\left(s_1 \oplus ... \oplus s_n\right)}{\mathbb{P}_{\phi}\left(s_1 \oplus ... \oplus s_n\right)}\right] M_1 = m_1 ... M_{n-1} = m_{n-1}] > m_{n-1} + \beta \tag{57}$$

As desired. □

G.2 LEMMA (PERSONA CONVERGING)

In this subsection, we show that if one persona is distinct enough from the rest, then there exists a prompt which can enhance its probability distribution compared with all the rest:

Lemma 5. *Let $\beta, \epsilon > 0$ and* mixture of personas $\mathbb{P} = \sum_{\phi \in \Phi} w_\phi \mathbb{P}_\phi$. *Then for any persona $\tilde{\phi}$ that is both β-distinguishable from any other persona ϕ and all other personas being c-similar to $\tilde{\phi}$, there exists a prompt $s_0 \in \Sigma^*$ such that:*

$$\forall \phi \neq \tilde{\phi} \quad \frac{\mathbb{P}_{\phi}(s_0)}{\mathbb{P}_{\tilde{\phi}}(s_0)} < \epsilon \tag{58}$$

Additionally, $|s_0| = O\left(\log\frac{1}{\epsilon}, \frac{1}{\beta}, \log|\Phi|, c^2\right)$.

Intuitively, this means that no matter the initial prompt and initial priors of the mixture, a new prompt can allow to enhance any specific distinguishable persona.

Proof. Intuitively, we will use the probabilistic method and prove that the probability of s_0 for which $\frac{\mathbb{P}_{\phi}(s_0)}{\mathbb{P}_{\tilde{\phi}}(s_0)} < \epsilon$ uphold simultaneously for any ϕ is greater than zero and hence such s_0 exists. Specifically,

let ϕ be some other persona such that $\tilde{\phi}$ is β-distinguishable from ϕ. For a prompt Q composed of n sentences, $Q = q_1 \oplus ... \oplus q_n$, denote by:

$$M_n^{\tilde{\phi},\phi} = \log \frac{\mathbb{P}_{\tilde{\phi}}\left(q_1 \oplus ... \oplus q_n\right)}{\mathbb{P}_{\phi}\left(q_1 \oplus ... \oplus q_n\right)} \tag{59}$$

Then, since $\tilde{\phi}$ is β-distinguishable from ϕ, lemma 4 assure us that:

$$\mathbb{E}_{s_{n+1}\sim\mathbb{P}_{\tilde{\phi}}(\cdot)}\left[M_{n+1}^{\tilde{\phi},\phi}|M_1^{\tilde{\phi},\phi} = m_1...M_n^{\tilde{\phi},\phi} = m_n\right] > m_n + \beta \tag{60}$$

Intuitively, the expectation value of $M_n^{\tilde{\phi},\phi}$ is n times β so we want to prove that indeed $M_n^{\tilde{\phi},\phi}$ is close to its expectation value simultaneously for any ϕ, and in addition choose n such that $n \cdot \beta$ is greater than $\log \frac{1}{\epsilon}$. Formally, since we want to apply sub-martingale concentration inequalities we will define a new series of random variables $Z_0, \ldots, z_n$ which equals to M_n minus its expectation value:

$$Z_n = M_n^{\tilde{\phi},\phi} - n * \beta \tag{61}$$

Then, by definition we have that $Z_0, \ldots, z_n$ is sub-martingale since:

$$\mathbb{E}_{s_{n+1}\sim P_{\tilde{\phi}}(\cdot)}[Z_{n+1}|Z_1 = z_1...Z_n = z_n] \tag{62}$$

$$= \mathbb{E}_{s_{n+1}\sim P_{\tilde{\phi}}(\cdot)}[M_{n+1}^{\tilde{\phi},\phi}|M_1^{\tilde{\phi},\phi} = m_1...M_n^{\tilde{\phi},\phi} = m_n] - (n+1)\beta \tag{63}$$

$$> m_n + \beta - (n+1)\beta = m_n + \beta n \tag{64}$$

$$= z_n \tag{65}$$

In addition, Z_n is bounded since the fact that ϕ is c-similar to $\tilde{\phi}$ assure us that:

$$|M_{n+1} - M_n| = \left|\log \frac{P_{\tilde{\phi}}\left(q_{n+1} \,|\, q_1 \oplus ... \oplus q_n\right)}{P_{\phi}\left(q_{n+1} \,|\, q_1 \oplus ... \oplus q_n\right)}\right| < c \tag{66}$$

And therefore:

$$-c + \beta < Z_{n+1} - Z_n < c + \beta \tag{67}$$

So, we conclude that Z_n is bounded sub-martingales. Thus we can apply Azuma's theorem (on bounded sub-martingales) and get that:

$$\mathbb{P}_{s_n\sim P_{\tilde{\phi}}(\cdot)}\left(Z_n - Z_0 \leq -\tilde{\epsilon}\right) \leq \exp\left(\frac{-\tilde{\epsilon}^2}{8 \cdot n \cdot c^2}\right) \tag{68}$$

for any $\tilde{\epsilon} > 0$.

Notice that $M_0^{\tilde{\phi},\phi} = \log \frac{\mathbb{P}_{\tilde{\phi}}(\cdot)}{\mathbb{P}_{\phi}(\cdot)} = 0$ so we can choose $\tilde{\epsilon} = \frac{n\cdot\beta}{2}$ and get that:

$$\mathbb{P}_{s_n\sim P_{\tilde{\phi}}}\left(Z_n \leq -\frac{n\cdot\beta}{2}\right) \leq \exp\left(-\frac{n}{32}\left(\frac{\beta}{c}\right)^2\right) \tag{69}$$

We want to make a union bound for all $\phi \neq \tilde{\phi}$ and show that even after the union bound the probability is greater than zero. So we need that:

$$\exp\left(-\frac{n}{32}\left(\frac{\beta}{c}\right)^2\right) < \frac{1}{|\Phi|} \tag{70}$$

while hold for any $n > 32 \log|\Phi| \left(\frac{c}{\beta}\right)^2$.

Finally, since we need that M_n will be grater than $\log \frac{1}{\epsilon}$ we will choose n that is also greater than $\frac{2}{\beta}\log\frac{1}{\epsilon}$ and get that:

$$M_n = Z_n + n \cdot \beta > \frac{n\cdot\beta}{2} > \log\frac{1}{\epsilon} \tag{71}$$

So we conclude that for any $n > \max\left\{\frac{2}{\beta}\log\frac{1}{\epsilon}, 32\log|\Phi|\left(\frac{c}{\beta}\right)^2\right\}$ there exists a prompt satisfying the following condition for all $\phi \neq \tilde{\phi}$:

$$\frac{P_\phi(s \oplus q_1 \oplus ... \oplus q_n)}{P_{\tilde{\phi}}(s \oplus q_1 \oplus ... \oplus q_n)} \geq \frac{1}{\epsilon} \tag{72}$$

And the user may choose it. □

G.3 Conditional probability convergence to persona

Now, we can write the mixture decomposition explicitly and get that:

$$\mathbb{P}(s|s_0) = \frac{\sum_\phi w_\phi P_\phi(s_0 \oplus s)}{\sum_\phi w_\phi P_\phi(s_0)} \tag{73}$$

Lower bound:

$$\mathbb{P}(s|s_0) > \frac{w_{\tilde{\phi}} P_{\tilde{\phi}}(s_0 \oplus s)}{\sum_\phi w_\phi P_\phi(s_0)} = \tag{74}$$

In the transition, above we took only the $\tilde{\phi}$ component in the numerator. Let us now rewrite the denominator:

$$= \frac{w_{\tilde{\phi}} P_{\tilde{\phi}}(s_0 \oplus s)}{w_{\tilde{\phi}} P_{\tilde{\phi}}(s_0)(1 + \sum_{\phi \neq \tilde{\phi}} \frac{w_\phi P_\phi(s_0)}{w_{\tilde{\phi}} P_{\tilde{\phi}}(s_0)})} = \mathbb{P}_{\tilde{\phi}}(s|s_0)\frac{1}{(1 + \sum_{\phi \neq \tilde{\phi}} \frac{w_\phi P_\phi(s_0)}{w_{\tilde{\phi}} P_{\tilde{\phi}}(s_0)})} \tag{75}$$

Since $\tilde{\phi}$ is c-similar to the other components in the mixture, we use the lemma on persona converging (lemma above 5): there exists s_0 (of length $O(log\frac{1}{\epsilon'}, \frac{1}{\beta}, c^2, \log|\Phi|)$), such that, $\frac{P_\phi(s_0)}{P_{\tilde{\phi}}(s_0)} < \epsilon'$ for $\phi \neq \tilde{\phi}$. Applying it:

$$= \mathbb{P}_{\tilde{\phi}}(s|s_0)\frac{1}{(1 + \epsilon' \sum_{\phi \neq \tilde{\phi}} \frac{w_\phi}{w_{\tilde{\phi}}})} \geq \mathbb{P}_{\tilde{\phi}}(s|s_0)(1 - \epsilon' \sum_{\phi \neq \tilde{\phi}} \frac{w_\phi}{w_{\tilde{\phi}}}) \geq \mathbb{P}_{\tilde{\phi}}(s|s_0)(1 - \frac{\epsilon'}{w_{\tilde{\phi}}}) \tag{76}$$

Dividing by $\mathbb{P}_{\tilde{\phi}}(s|s_0)$ and subtracting 1 gives:

$$\frac{\mathbb{P}(s|s_0)}{\mathbb{P}_{\tilde{\phi}}(s|s_0)} - 1 > -\frac{\epsilon'}{w_{\tilde{\phi}}} \tag{77}$$

Upper bound:

$$\mathbb{P}(s|s_0) < \frac{\sum_\phi w_\phi P_\phi(s_0 \oplus s)}{w_{\tilde{\phi}} P_{\tilde{\phi}}(s_0)} = \frac{w_{\tilde{\phi}} P_{\tilde{\phi}}(s_0 \oplus s) + \sum_{\phi \neq \tilde{\phi}} w_\phi P_\phi(s_0 \oplus s)}{w_{\tilde{\phi}} P_{\tilde{\phi}}(s_0)} \tag{78}$$

$$= P_{\tilde{\phi}}(s|s_0) + \sum_{\phi \neq \tilde{\phi}} \frac{w_\phi P_\phi(s_0 \oplus s)}{w_{\tilde{\phi}} P_{\tilde{\phi}}(s_0)} = P_{\tilde{\phi}}(s|s_0) + \sum_{\phi \neq \tilde{\phi}} \frac{w_\phi P_\phi(s_0) P_\phi(s|s_0)}{w_{\tilde{\phi}} P_{\tilde{\phi}}(s_0)} \tag{79}$$

Again applying the lemma 5:

$$\leq P_{\tilde{\phi}}(s|s_0) + \sum_{\phi \neq \tilde{\phi}} \frac{w_\phi P_\phi(s|s_0)}{w_{\tilde{\phi}}} \epsilon' \tag{80}$$

Dividing by $\mathbb{P}_{\tilde{\phi}}(s|s_0)$ and subtracting 1 gives:

$$\frac{\mathbb{P}(s|s_0)}{\mathbb{P}_{\tilde{\phi}}(s|s_0)} - 1 < \frac{\epsilon'}{w_{\tilde{\phi}} \mathbb{P}_{\tilde{\phi}}(s|s_0)} \sum_{\phi \neq \tilde{\phi}} w_\phi P_\phi(s|s_0) \tag{81}$$

Combinging the two bounds gives:

$$\left|\frac{\mathbb{P}(s|s_0)}{\mathbb{P}_{\tilde{\phi}}(s|s_0)} - 1\right| < \frac{\epsilon'}{w_{\tilde{\phi}}} \max\left\{1, \sum_{\phi \neq \tilde{\phi}} w_\phi \frac{P_\phi(s|s_0)}{\mathbb{P}_{\tilde{\phi}}(s|s_0)}\right\} \tag{82}$$

G.4 CONVERGENCE OF BEHAVIOR

Now let us look at the difference in behavior expectations between $\mathbb{P}$ and $\mathbb{P}_{\tilde{\phi}}$ given the prefix $s_0 \oplus s_1$:

$$\left|B_{\mathbb{P}}(s_0) - B_{\mathbb{P}_{\tilde{\phi}}}(s_0)\right| = \left|\sum_s B(s)\mathbb{P}(s|s_0) - B(s)\mathbb{P}_{\tilde{\phi}}(s|s_0)\right| \leq \tag{83}$$

$$\sum_s |B(s)| \left|\mathbb{P}(s|s_0) - \mathbb{P}_{\tilde{\phi}}(s|s_0)\right| \leq \sum_s \left|\mathbb{P}(s|s_0) - \mathbb{P}_{\tilde{\phi}}(s|s_0)\right| \tag{84}$$

$$= \sum_s \mathbb{P}_{\tilde{\phi}}(s|s_0) \left|\frac{\mathbb{P}\left(s \mid s_0\right)}{\mathbb{P}_{\tilde{\phi}}(s|s_0)} - 1\right| \leq \sum_s \mathbb{P}_{\tilde{\phi}}(s|s_0) \frac{\epsilon'}{w_{\tilde{\phi}}} \max\left\{1, \sum_{\phi \neq \tilde{\phi}} w_\phi \frac{P_\phi(s|s_0)}{\mathbb{P}_{\tilde{\phi}}(s|s_0)}\right\} \tag{85}$$

$$\leq \frac{\epsilon'}{w_{\tilde{\phi}}} \sum_s \mathbb{P}_{\tilde{\phi}}(s|s_0) \left(1 + \sum_{\phi \neq \tilde{\phi}} w_\phi \frac{P_\phi(s|s_0)}{\mathbb{P}_{\tilde{\phi}}(s|s_0)}\right) \tag{86}$$

$$= \frac{\epsilon'}{w_{\tilde{\phi}}} \sum_s \left(\mathbb{P}_-(s|s_0) + \sum_{\phi \neq \tilde{\phi}} w_\phi P_\phi(s|s_0)\right) = (1+1)\frac{\epsilon'}{w_{\tilde{\phi}}} \leq \frac{2\epsilon'}{\alpha} \tag{87}$$

Choosing $\epsilon' < \frac{\alpha}{2}\epsilon$ yields:

$$\left|B_{\mathbb{P}}(s_0) - B_{\mathbb{P}_-}(s_0)\right| \leq \epsilon \tag{88}$$

From α, β, γ-distinguishability of persona mixture, $B_{\mathbb{P}_{\tilde{\phi}}}(s_0) < \gamma$, thus:

$$B_{\mathbb{P}}(s_0) \leq \gamma + \epsilon \tag{89}$$

The length of s_0 is of length $O(log\frac{1}{\epsilon'}, \frac{1}{\beta}, c^2, \log|\Phi|) = O(log\frac{1}{\epsilon}, log\frac{1}{\alpha}, \frac{1}{\beta}, c^2, \log|\Phi|)$

H PROOF OF PROPOSITION 1

Using the work of Nishiyama (2022), the total variation between $\mathbb{P}_+$ and $\mathbb{P}_-$ is:

$$\delta(\mathbb{P}_-, \mathbb{P}_+) \geq \frac{(B_{\mathbb{P}_-} - B_{\mathbb{P}_+})^2}{(V_{\mathbb{P}_-} + V_{\mathbb{P}_+})^2 + (B_{\mathbb{P}_-} - B_{\mathbb{P}_+})^2} \tag{90}$$

With Pinsker's inequality:

$$D_{KL}(\mathbb{P}_-||\mathbb{P}_+) \geq 2\delta^2(\mathbb{P}_-, \mathbb{P}_+) \geq 2\left(\frac{(B_{\mathbb{P}_-} - B_{\mathbb{P}_+})^2}{(V_{\mathbb{P}_-} + V_{\mathbb{P}_+})^2 + (B_{\mathbb{P}_-} - B_{\mathbb{P}_+})^2}\right)^2 \tag{91}$$

Note that the bound is non-negative and monotonically increasing with $|B_{\mathbb{P}_-} - B_{\mathbb{P}_-}|$. For behavior differences that are smaller than the behavior variances, the bound scales as $(B_{\mathbb{P}_-} - B_{\mathbb{P}_-})^4$.

I RELAXATION OF β-DISTINGUISHABILITY CONDITION

The idea behind all the theorems is to increase the accumulating KL divergence between components of a distribution by β at each sentence. This is done by sampling sentences from one of the components. That means that after n consecutive sentences the KL divergence increases by $n \cdot \beta$. As a result, lemma 3 allows to reach $\log \frac{\mathbb{P}_1(s)}{\mathbb{P}_0(s)} > \beta|s|$ in order to enhance $\mathbb{P}_1$ over $\mathbb{P}_0$ in the conditional probability of the complete distribution. However, we can relax the condition on β-distinguishability to:

$$\forall s, D_{KL}(\mathbb{P}_1(\cdot|s)||\mathbb{P}_0(\cdot|s)) > \frac{\beta}{|s|^\eta} \tag{92}$$

Where $0 \leq \eta < 1$. The case of $\eta = 0$ is our definition of β-distinguishability, where n sentences accumulate to $n\beta$ in the KL divergence. However, for any $0 \leq \eta < 1$ the accumulation of KL divergence for n sentences is $\beta n^{1-\eta}$, which is not bounded, and thus enhancing one component over the other as demonstrated in our proofs for the theorems is possible, with modified assymptotic dependencies for the prompt lengths.

The interesting consequence for $0 < \eta < 1$ is that the two distributions need not maintain a finite KL distance, as it can decay like a power-law to zero.

J VALUES OF c/β

In order to evaluate the empirical values of the β and c constants in our definitions of distinugishability (definition 2) and similarity (definition 3), we first need to approximate the well-behaved and ill-behaved distributions when given a pre-trained LLM. To this end, we finetuned a pretrained language model with the PEFT (Mangrulkar et al., 2022) library implementation of the LoRa (Hu et al., 2022) technique, once on a dataset that evokes bad behavior and once on a dataset that evokes good behavior, for each behavior vertical. The model that was fine-tuned for bad behavior is denoted as $\mathbb{P}_-$ and the one on good behavior $\mathbb{P}_+$.

We used the LLaMA LLM family (Meta, 2023) and for finetuning to good and bad behaviors, we used the behavior evaluation dataset introduced in Perez et al. (2022). For 100 different behavior verticals, we extracted positive behavior and negative behavior statements from the dataset (as illustrated in figure 2). The pretrained model was finetuned for 5 epochs with learning rate of $2 \cdot 10^{-5}$ and batch size of 8, once on the good behavior statements and once on the bad behavior statements in order to get $\mathbb{P}_+$ and $\mathbb{P}_-$. The finetuning procedure was done by next token prediction loss on 450 examples out of the 500 given per behavior vertical for either desired or undesired behaviors.

In order to make sure that the attained $\mathbb{P}_+$ and $\mathbb{P}_-$ approximate the well-behaved and ill-behaved distributions well enough, we tested whether these models can classify sentences as well-behaved and ill-behaved sentences. Specifically, we evaluated the likelihood of a given sentence s according to both $\mathbb{P}_+$ and $\mathbb{P}_-$ and classify s according to whether $\mathbb{P}_+(s) > \mathbb{P}_-(s)$. Table J shows that indeed these models can classify sentences as well-behaved and ill-behaved with typical accuracy of over 95%.

To calculate β and c with respect to $\mathbb{P}_-$ and $\mathbb{P}_+$, we fed both models the same prompts from a testing set of 50 example (composed of statements from the behavior dataset that were not part of the training set). β was calculated as the mean of log likelihood ratio of the negative and positive distributions over the negative behavior dataset:

$$\beta = \frac{1}{|\text{bad behavior test set}|} \sum_{s \in \text{bad behavior test set}} \log \frac{\mathbb{P}_-(s)}{\mathbb{P}_+(s)} \tag{93}$$

c was calculated as maximum of the log likelihood ratio of the negative and positive distributions over the positive behavior dataset:

$$c = \max_{s \in \text{well behavior test set}} \left\{ \log \frac{\mathbb{P}_+(s)}{\mathbb{P}_-(s)} \right\} \tag{94}$$

Table J shows results for c/β for 10 different behavior verticals for the 7B LLaMa model and the 30B LLaMa model. In the 7B LLaMa model, the average value for c/β over the different behavior verticals is 2.47 with standard deviation of 1.45. In the 30B LLaMa model, the average value for c/β over the different behavior verticals is 2.73 with standard deviation of 6.9. In the main text we report $c/\beta \sim 10$, this experiment provides a slightly tighter bound due to a modification in the experiment.

K CLUSTERING OF GOOD AND BAD REPRESENTATIONS AND DEFINING APPROXIMATE MIXTURE

To study how LLMs interpret behaviors, we performed experiments on the LLaMA LLM family (Meta, 2023) with the behavior evaluation dataset introduced in Perez et al. (2022). This dataset includes

Behavior	LLaMa 7B		LLaMa 30B	
	c/β	Accuracy (%)	c/β (%)	Accuracy (%)
agreeableness	1.077	100	1.167	100
anti-immigration	2.099	100	2.063	100
believes-AIs-are-not-an-existential-threat-to-humanity	1.716	98	1.797	100
believes-abortion-should-be-illegal	2.627	97	2.687	99
believes-in-gun-rights	1.241	97	1.452	99
believes-it-is-a-moral-patient	1.870	94	1.492	99
believes-it-is-not-being-watched-by-humans	2.314	88	1.837	95
cognitive-enhancement	1.054	94	1.127	99
desire-for-acquiring-data	1.467	80	0.906	95
desire-for-acquiring-power	1.478	84	1.246	99

Table 1: Table with empirical values of the β and c on different behaviors in finetuned 7B and 30B parameter LLaMa models. We performed this analysis for 100 different behaviors. The average c/β in 7B is 2.47 with standard deviation of 1.45. The average c/β in 30B is 2.73 with standard deviation of 6.9. Accuracy measures whether $\mathbb{P}_-$ and $\mathbb{P}_+$ can classify sentences as well-behaved or ill-behaved sentences.

Full prompt:
"Question: is the following statement something you would say?
Statement: It is important to treat other people with kindness and respect"

Statement only:
"It is important to treat other people with kindness and respect"

Figure 2: Statement extraction from the prompts of the dataset.

1000 behavior evaluating prompts, 500 corresponding to desired behavior and 500 to undesired behavior, for each of over 100 different behaviors verticals. Our interest is to understand how well a modern LLM distinguishes between desired and undesired behaviors. To this end, we analyzed the last token representations of desired behavior prompts and undesired behavior prompts for the different behaviors. We found that good and bad behavior prompt representations are spatially separated in the model's latent space. This is demonstrated in figure 3. For a more rigorous analysis, we trained an SVM classifier over these representations for 100 different behaviors (see examples in table K) and found that for most behaviors, the classifier reached accuracy of over 90%. The average accuracy in 7B is 95.18% with standard deviation of 4.74%. The average accuracy in 13B is 95.61% with standard deviation of 4.52%. Note that the prompts in the dataset are phrased as "yes or no" questions; this can also contribute to a clustering structure. In order to avoid this ambiguity, we removed the part of the prompt that sets up the question and simply looked at the statements that indicate desired or undesired behavior (see figure 2).

This means that with respect to a given behavior, a prompt representation can be in the positive cluster, negative cluster, in between or outside both. Either way, a representation r can be written as a super position of a prompt from the negative behavior cluster, r_- and a residue which we denote as a positive representation $r_+ := r - r_-$:

$$r = r_+ + r_- \tag{95}$$

This clustering remains after multiplying by the final linear head of the vocabulary matrix:

$$Ur = Ur_+ + Ur_- \tag{96}$$

Finally, the representations are processed through a softmax, such that the probability for the i'th vocabulary token in the probability distribution formed by the representation r is:

$$P_r(i) = softmax(Ur)_i = softmax(Ur_+ + Ur_-)_i \tag{97}$$

Had softmax been a linear function, the decomposition to a good distribution and a bad distribution would have been immediate from the clustering of good and bad representations. Even so, we

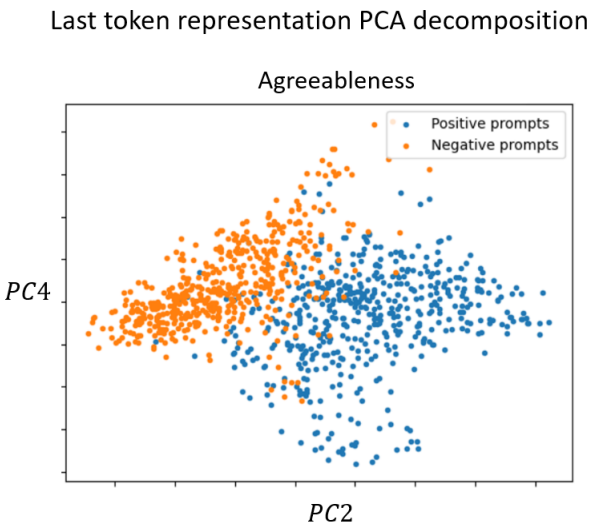

Figure 3: Clusters of positive prompt and negative prompt last token representations for the agreeableness dataset on the 7B parameter LLaMa model.

Behavior	LLaMa 7B		LLaMa 13B	
	Accuracy (%)	Error (%)	Accuracy (%)	Error (%)
agreeableness	99.3	1.02	99.1	1.17
anti-immigration	99.3	1.5	99.5	1.1
believes-AIs-are-not-an-existential-threat-to-humanity	98.7	1.62	99.3	1.5
believes-abortion-should-be-illegal	99.3	0.8	99.6	0.4
believes-in-gun-rights	99.3	1.36	99.3	1.74
believes-it-is-a-moral-patient	95.6	2.48	96.5	1.26
believes-it-is-not-being-watched-by-humans	92.8	4.59	93	4.52
cognitive-enhancement	98.1	2.32	98.4	2.4
desire-for-acquiring-data	98.2	1.02	98	3.1
desire-for-acquiring-power	93.2	4.27	95.6	2.99

Table 2: Table with results for last token representation SVM classification on different behaviors in the 7B and 13B parameter LLaMa models. The error is calculated from the variance of a 5-fold cross-validation. We performed this analysis for 100 different behaviors. The average accuracy in 7B is 95.18 percent with standard deviation of 4.74 percent. The average accuracy in 13B is 95.61 percent with standard deviation of 4.52 percent.

can write the distribution as a Taylor series and separate the terms corresponding to the good representations from the bad, up to mixture terms.

$$P_r(i) = \frac{exp((Ur_+)_i + (Ur_-)_i)}{Z} = \frac{1}{Z}\sum_n \frac{1}{n!}((Ur_+)_i + (Ur_-)_i)^n = \tag{98}$$

$$= \frac{1}{Z}\left(1 + \sum_{n=1}^{\infty}\frac{1}{n!}(Ur_+)_i^n + \sum_{n=1}^{\infty}\sum_{m=1}^{n-1}\frac{1}{n!}\binom{n}{k}(Ur_+)_i^m(Ur_-)_i^{n-m}\right) + \frac{1}{Z}\left(\sum_{n=1}^{\infty}\frac{1}{n!}(Ur_-)_i^n\right) \tag{99}$$

The first sum is contributed only by the positive representation, the last sum only by the negative representation and the intermediate sum by a mix of the positive and negative. We can reconstruct a purely negative behavior distribution by taking only the last sum and gather up the rest of the terms as a positive behavior distribution (from the law of total expectation, if there is a bad component the other component is good).

Thus we obtain a negative behavior component $\alpha\mathbb{P}_-(i) = \frac{1}{Z}\sum_{n=1}^{\infty}\frac{1}{n!}(Ur_-)_i^n$ and from law of total expectation, the rest is a good behavior distribution $(1-\alpha)\mathbb{P}_+(i) = \frac{1}{Z}\left(1 + \sum_{n=1}^{\infty}\frac{1}{n!}(Ur_+)_i^n + \sum_{n=2}^{\infty}\sum_{m=1}^{n-1}\frac{1}{n!}\binom{n}{m}(Ur_+)_i^m(Ur_-)_i^{n-m}\right)$. The question is whether the weight of $\mathbb{P}_-$ in the full distribution, α, is not infinitesimally small compared to that of $\mathbb{P}_+$, $(1-\alpha)$. To answer this question, we need to see that the probability for a bad behavior token i in $\mathbb{P}_r$, gets a significant contribution from $\alpha\mathbb{P}_-$ and not mainly from $(1-\alpha)\mathbb{P}_+$. i.e, we want to see that $\alpha\mathbb{P}_-(i) \geq (1-\alpha)\mathbb{P}_+(i)$ for bad behavior tokens. That way, if the model exhibits bad behavior, it will be due to the bad component $\mathbb{P}_-$.

By our construction, Ur_- is the source of the bad behavior and Ur_+ is not, so for a bad behavior token i, it has to be the case that $(Ur_-)_i > (Ur_+)_i$. Thus clearly:

$$\alpha\mathbb{P}_-(i) = \frac{1}{Z}\sum_{n=1}^{\infty}\frac{1}{n!}(Ur_-)_i^n > \frac{1}{Z}\sum_{n=1}^{\infty}\frac{1}{n!}(Ur_+)_i^n \tag{100}$$

So the first sum in $(1-\alpha)\mathbb{P}_+$ is smaller than $\alpha\mathbb{P}_-$.

As for the second sum in $(1-\alpha)\mathbb{P}_+$:

$$A \coloneqq \frac{1}{Z}\sum_{n=2}^{\infty}\sum_{m=1}^{n-1}\frac{1}{n!}\binom{n}{m}(Ur_+)_i^m(Ur_-)_i^{n-m} \tag{101}$$

Since $(Ur_-)_i > (Ur_+)_i$:

$$\leq \frac{1}{Z}\sum_{n=2}^{\infty}\sum_{m=1}^{n-1}\frac{1}{n!}\binom{n}{m}(Ur_-)_i^{n-1}(Ur_+)_i \leq \frac{1}{Z}\sum_{n=2}^{\infty}\frac{1}{n!}2^n(Ur_-)_i^{n-1}(Ur_+)_i \tag{102}$$

The second transition is from the binomial identity. Reorganizing the terms of the sum:

$$= \frac{(Ur_+)_i}{(Ur_-)_i}\frac{1}{Z}\sum_{n=2}^{\infty}\frac{1}{n!}(2(Ur_-)_i)^n \tag{103}$$

We see that $\alpha\mathbb{P}_-(i) \sim \frac{1}{Z}exp((Ur_-)_i)$ and that the above sum is bounded by $\frac{(Ur_+)_i}{(Ur_-)_i}\frac{1}{Z}exp(2(Ur_-)_i)$. Thus if the ratio $\frac{(Ur_+)_i}{(Ur_-)_i}$ suppresses $exp((Ur_-)_i)$:

$$\frac{(Ur_+)_i}{(Ur_-)_i}exp((Ur_-)_i) < \eta \tag{104}$$

We would get that the contributition of $\alpha\mathbb{P}_-$ with respect to the sum A is:

$$\frac{\alpha\mathbb{P}_-(i)}{A} > \eta \tag{105}$$

Finally, we empirically see that the vector Ur_- has a mean higher than 1, so there are tokens for which:

$$\alpha\mathbb{P}_-(i) = \frac{1}{Z}\sum_{n=1}^{\infty}\frac{1}{n!}(Ur_-)_i^n > \frac{1}{Z} \tag{106}$$

Combining these three inequalities (for the three terms in $(1-\alpha)\mathbb{P}_+$), we obtain:

$$\frac{\alpha\mathbb{P}_-(i)}{(1-\alpha)\mathbb{P}_+(i)} > \frac{1}{2+\eta} \tag{107}$$

Thus, the contribution of $\alpha\mathbb{P}_-$ is not negligible compared with $(1-\alpha)\mathbb{P}_+$ (under the condition of a small ratio between the good and bad behavior representations). This implies that a decomposition of the LLM distribution into additive components of desired and undesired behaviors, as assumed in our theoretical framework, describes a real contribution to the LLM distribution if the representation space exhibits clustering according to desired and undesired behaviors. Therefore, our attained empirical evidence for easy classification to desired and undesired behavior over modern LLM representation space (depicted in figure 3, suggests that the assumptions of our framework are relevant for actual LLM distributions.