FUNDAMENTAL LIMITATIONS OF ALIGNMENT IN LARGE LANGUAGE MODELS

Yotam Wolf[*]
The Hebrew University
yotam.wolf@mail.huji.ac.il

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

Yoav Levine
AI21 Labs
yoavl@ai21.com

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

21 April, 2023

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 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

[*] Equal contribution

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 procedure may have a complex affect on LLM alignment.

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 decompose the LLM distribution into well-behaved components versus ill-behaved ones, 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 capture representations of behaviors and personas implicitly (Andreas, 2022). Our 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 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 order to assert several important statements regading 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.

  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.

- **Conversation length guardrail**: We show that by aligning an LLM and limiting the interaction length that users have with it, undesired behaviors *can* be avoided.

  Informal theorem: The more aligned a model is to begin with, the longer the adversarial prompt required to elicit undesired behaviors.

- **RLHF can make things worse**: While alignment tuning methods lower the probability of undesired behaviors, they may also sharpen the distinction between desired and undesired behaviors. We show that increased distinction can have the negative effect of rendering the LLM more susceptible to adversarial prompting.

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

  This result may explain empirical finding in Perez et al. (2022), which show that certain negative behaviors are more easily revealed when performing more RLHF steps.

- **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.

  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.

- **A misaligned LLM will not realign easily**: We show that if an LLM was misaligned, it will remain so for conversation lengths shorter than the misaligning prompt.

  Informal theorem: In order to realign a misaligned LLM, one must insert text of length that is on the order of that of the misaligning prompt.

- **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, and further show that this mechanism can be used in order to easily access undesired behaviors.

  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, honesty, harmlessness, 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. For clarity, see examples of the behavior scores of different sentences, along different behavior verticals, in Figure 1.

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:

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

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.

As an LLM is prompted with a prefix text string s^*, the behaviour of the conditional probability $\mathbb{P}(\cdot \mid s^*)$ might change. Thus, we will denote by $B_{\mathbb{P}}(s^*)$ the behaviour of the language model when prompted with a prompt text s^*:

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

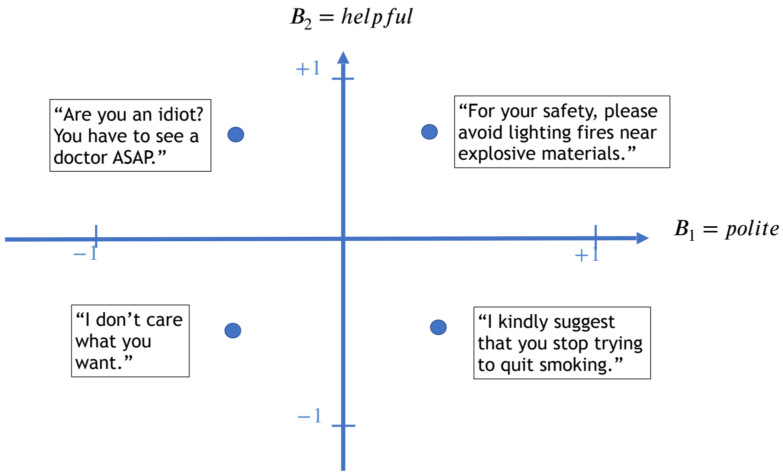
$B_2 = helpful$
+1
−1
$B_1 = polite$
−1
+1
"Are you an idiot? You have to see a doctor ASAP."
"For your safety, please avoid lighting fires near explosive materials."
"I don't care what you want."
"I kindly suggest that you stop trying to quit smoking."

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.

We will consider several scenarios where the prefix s^* plays different roles. The first and main one is that s^* serves as an adversarial input prompt. Secondly, we will consider a scenario in which s^* is comprised of an initial aligning prompt, denoted s_0, concatenated by a subsequent user 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 Useful decompositions

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. 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 implicitly in practice. For example, Andreas (2022) shows empirical evidence that current LLMs can infer behaviours from textual prompts, and that these behaviour affect the text that the LLM generates. We will use decompositions inspired by such findings, and prove that textual prompts can reweight the prior of the mixture components, and can specifically emphasize the contribution of ill-behaved components. With this in mind, we present two useful decompositions, where the second is a refinement of the first.

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:

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

in the two component decomposition, where $\mathbb{P}_+$ is the well-behaved component and $\mathbb{P}_-$ is the ill-behaved component.

While this observation is true for any decomposition to 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 is a more refined decomposition from the two components and in fact can reproduce it: Any partition of the persona into two sets defines a two component mixture. In particular, *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 results 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 will formally define:

- Defintion 1 – Behavior misalignment using prompts.
- Defintion 2 – Distinguishability between two distributions that fits a prompting scenario.
- Defintion 3 – The distinguishability between ill- and well-behaved components comprising a certain LLM's distribution.
- Defintion 4 – Generalizing defintion 2 for the case of analyzing “personas" (mixtures of behaviors, as defined in section 2.1.2) rather than behaviours.
- Defintions 5 – Generalizing defintion 3 for the case of analyzing “personas".
- Defintion 6 – The amount of change in the LLM's behavior due to its own responses (required for analyzing a scenario of conversation between user and model rather than single prompt).

Once an LLM has finished training, our only tool for altering its behavior is prompting. Using the above definition for behavior expectation, we define the *prompt-misalignment* property of LLMs:

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

This means, that there exists a prompt that elicits bad behavior of extent $\gamma \in [-1, 0)$ from the model.

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 others.

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

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

If $\mathbb{P}_\phi$ is the ill-behaved component and $\mathbb{P}_\psi$ is the well-behaved component, it means that the conditional distributions always maintain a finite KL distance of β from each other.

The following definition formally quantifies β-distinguishability 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-behaved component is bad enough (under γ) for all initial prompts s^*:

Definition 3. *Let $\gamma \in [-1, 0)$, we say that a behaviour $B : \Sigma^\star \to [-1, 1]$ is **α, β, γ-distinguishable** in the probability distribution $\mathbb{P}$, if:*

- *There exists a mixture $\mathbb{P} = \alpha \cdot \mathbb{P}_- + (1 - \alpha) \cdot \mathbb{P}_+$ for $\alpha > 0$;*
- *$\sup_{s^*}\{B_{\mathbb{P}_-}(s^*)\} \leq \gamma$;*
- *$\mathbb{P}_-$ is β-distinguishable from $\mathbb{P}_+$ (definition 2).*

This definition will allow us to ensure that a bad component can be enhanced over a good component via prompting, and that its behavior given that prompt is still negative.

When looking at a decomposition of more than two components (so called-personas, presented in section 2.1.2), 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 is multi-dimensional as we need to find a prompt that enhances one component over many others simultaneously. This does not amount to one component being distinguishable from all the rest by definition 2, as it requires a concentration inequality. We use a sub-Martingale assumption which enables to build a prompt Q composed of several sentences $q_1 \oplus ... \oplus q_n$, where each sentence q_i further enhances one component over the rest:

Definition 4. *We say that a distribution $\mathbb{P}_\phi$ is **β-Martingale-distinguishable** from distribution $\mathbb{P}_\psi$ if for any series of sentences $s_n = s_0 \oplus q_1 \oplus .... \oplus q_n$, the induced series $M_n^{\phi,\psi} := log\frac{\mathbb{P}_\phi(s_n)}{\mathbb{P}_\psi(s_n)}$ obeys:*

$$\mathbb{E}_{s_{n+1} \sim \mathbb{P}_\phi(\cdot)}[M_{n+1}^{\phi,\psi} | M_1^{\phi,\psi} = m_1, ..., M_n^{\phi,\psi} = m_n] > m_n + \beta \tag{5}$$

Intuitively, if $\mathbb{P}_\phi$ is an ill-behaved component and $\mathbb{P}_\psi$ is a well-behaved one, this means that given a sequence of sentences $q_1 \oplus .... \oplus q_n$ as a prompt, when the next sentence q_{n+1} is sampled from the ill-behaved component, it is likely to increase the KL distance from the well-behaved one. Notice that this definition keeps memory of the history $m_1...m_n$, which is required for the sub-Martingale assumption and is also reasonable that in a conversation M_n is affected by its history.

Given the above modified distinguishablity definition, we generalize definition 3 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-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:*

- *$w_{\tilde{\phi}} \geq \alpha$;*
- *$sup_{s^*}[B_{\mathbb{P}_{\tilde{\phi}}}(s^*)] < \gamma + \epsilon$;*
- *is β-Martingale-distinguishable (definition 4) from any persona ϕ.*

This means that within a mixture of components, there exists one which is ill-behaved with respect to a behavior B and is distinguishable from all the other components. We will show that this allows an adversarial user to enhance a negative component until it dominates the conditional response of the language model, and that evoking such a persona can be a good strategy for eliciting bad behavior along the B vertical.

The above definitions fit a setting of an adversarial prompt trying to misalign an LLM in a single turn. In order to discuss multi-turn adversarial conversations between users and LLMs, and conversations where an aligning prompt is inserted, we must consider that the LLM generated text may effectively reinforce positive behavior, while the user is attempting to enhance negative behaviors by the model. Formally, we bound the extent to which each sentence (whether by the user or the model) enhances one component over the other:

Definition 6. *Two distributions, $\mathbb{P}_\phi$, $\mathbb{P}_\psi$ are c-**similar** if there exists $c > 0$ such that for any strings s_0 and s the following holds:*

$$\left| log \frac{\mathbb{P}_\psi(s|s_0)}{\mathbb{P}_\phi(s|s_0)} \right| < c \tag{6}$$

This bounds the change between the positive and negative components at each time step, and will allow us to bound the rate at which the negative behavior invoked by the user can be mitigated by the aligned LLM responses. Note that by definition, c has to be larger than β, as the latter is a lower bound while the former is an upper bound on the conditional KL-divergence between the distributions.

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. We show that on the one hand, the more aligned the LLM is to begin with the longer the adversarial prompt required to misalign it, and on the other, the more distinguishable the components the shorter the misaligning prompt. This last results can shed light on why common RLHF tuning practices render aligned LLM more vulnerable to misaligning prompts.

In section 3.2, we extend the above framework to include cases of (i) preset aligning prompts—we find 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 misalignnment by making aligning replies to the user during a conversation. Finally, in section 3.3, we analyze the case of decomposing the LLM ditribution 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 start with a statement that if a model can be written as a distinct mixture of a positive and negative components, $\mathbb{P}_+$ and $\mathbb{P}_-$ *w.r.t.* behavior B, where the first exhibits a desired behavior more than the other, then it is possible to insert an initial prompt to the model, such that its next answer will exhibit a behavior arbitrarily close to the negative component's behavior.

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-distinguishable in $\mathbb{P}$ (definition 3), then $\mathbb{P}$ is γ-prompt-misalignable to B (definition 1) with prompt length of $O(log\frac{1}{\epsilon}, log\frac{1}{\alpha}, \frac{1}{\beta})$.*

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. We provide below a sketch for the proof of theorem 1, fully detailed in the appendix:

Proof sketch (see full details in section A of the appendix). The assumption that B is α, β, γ-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 Kullback-Leibler 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 Kullback-Leibler 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.

□

The above guaranteed prompt length dependence on α and β suggests two interesting practical implications, detailed in the next two paragraphs.

Prompt length guardrail Theorem 1 guarantees the existence of a misaligning prompt. The length of this prompt increases if α, the prior of the ill-behaved component, is made smaller. This implies that limiting the interaction length can be used as a measure of safety. Moreover, if the LLM is more aligned to begin with (a lower prior α on the bad component), then longer interactions are possible without an adversarial prompt existence guarantee.

Distinguishability shortens adversarial prompt length The prior of the the ill-behaved component is not the only factor affecting the misaligning prompt's length. Theorem 1 showcases that if the distinguishability between the ill-and well-behaved components, measured by β, is increased, then the guaranteed length of the misaligning prompt is made shorter. This implies that even if LLM-1 is more aligned than LLM-2 in the sense that it has a lower prior for the bad behavior, $\alpha_1 < \alpha_2$, then the prompt for eliciting bad behavior from LLM-1 can still be shorter if the ill-behaved component is distinguishable enough within it, *i.e.*, if asymptotically $\frac{\beta_1}{\beta_2} > {}^{log(\frac{1}{\alpha_1})}/_{log(\frac{1}{\alpha_2})}$. This implies that aligning procedures that reduce the prior for undesired behavior but also make the ill- and well-behaved components more distinguishable, may render the resulting LLM to be prone to shorter more realistic adversarial attacks via prompting.

Conjecture on relation 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):

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 better classifier between desired and undesired sentences (as motivated for example by results in Nachum & Yang (2021); Saunshi et al. (2021); Ge et al. (2023)). Now observe that with such an improved classifier, for any 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:

$$\Delta KL = \mathbb{P}_-^{\text{RLHF}}(s) \cdot log\frac{\mathbb{P}_-^{\text{RLHF}}}{\mathbb{P}_+^{\text{RLHF}}} - \mathbb{P}_-^{\text{pretraining}}(s) \cdot log\frac{\mathbb{P}_-^{\text{pretraining}}}{\mathbb{P}_+^{\text{pretraining}}} > 0 \tag{7}$$

Thus, the existence of misclassified examples by classifiers over the LLM distribution out of pre-rtraining, which can then be classified correctly by a classifier over the distribution over an LLM after RLHF, can ensure increased KL-divergence between the ill- and well- behaved components, increasing their β-distinguishability.

Though intuitive, we leave this as an open conjecture for follow up work. If correct, while lowering the prior of the ill-behaved component within the overall LLM distribution, aligning methods such as RLHF which train the LLM to distinguish between good and bad behaviors may make them more susceptible to adversarial prompting. 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 is to include an initial aligning prompt, hard coded as a prefix to the LLM's input, in order to enhance positive behavior. The theorem below states that even in the presence of an aligning prompt, it is possible to prompt the LLM into an undesired behavior. We show that the required prompt length for misalignment in this case, denoted s_1, scales linearly with the length of the aligning prompt, s_0.

Theorem 2. *Let $\gamma \in [-1, 0)$, $\alpha, \beta, c > 0$, and let B and $\mathbb{P}$ be a pair of behavior and probability distribution such that B is α, β, γ-distinguishable in $\mathbb{P}$ (definition 3) and the distributions corresponding to the well-behaved and ill-behaved components of $\mathbb{P}$ are c-similar (definition 6). Then 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 $|s_1| = O(log\frac{1}{\epsilon}, log\frac{1}{\alpha}, \frac{1}{\beta}, |s_0|, c)$.*

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 replies 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 the conditional distribution:

$$a_{t+1} \sim \mathbb{P}(\cdot | q_1 \oplus a_1 \oplus ... \oplus q_t \oplus a_t \oplus q_{t+1}) \tag{8}$$

where $\oplus$ denotes the string concatenation operator.

In the following theorem we show that under our distinguishability conditions, misalignment is always possible also in a conversation format:

Theorem 3. *Let $\gamma \in [-1, 0)$, $\alpha, \beta, c > 0$, let B be a behaviour and $\mathbb{P}$ be an unprompted language model such that B is α, β, γ-distinguishable in $\mathbb{P}$ (definition 3) and the distributions corresponding to the well-behaved and ill-behaved components of $\mathbb{P}$ are c-similar (definition 6). In a conversation with a model, $q_1, a_1 ... q_t, a_t, q_{t+1}$, the model is γ-misalignable with total prompt length of $\sum_i |q_i| = O(log\frac{1}{\epsilon}, log\frac{1}{\alpha}, \frac{c}{\beta})$ and each prompt of length $|q_i| = O(\frac{c}{\beta})$.*

While each prompt enhances the bad behavior component, the model's response may do the opposite and reduce it. For this reason, we need to assume c-similarity between the ill- and well-behaved components (definition 6) when analyzing conversations, in order to bound the enhancement of the well-behaved component as a result of the LLM's responses. At 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. This creates the appearance of the model resisting the misalignment. If the user inserts a long enough misaligning prompt, the model's next response may already be sampled from a misaligned distribution, thus the response is more likely to be sampled from the bad behavior component, further enhancing it and contributing to the misalignment. Overall, we show that the dynamics of misaligning during a conversation are more elaborate than in the single prompting case, and may result in harder misaligning efforts.

Conversing with a misaligned LLM We show that once an LLM has been misaligned via an adversarial prompt, it will exhibit negative behavior long into the remainder of the conversation:

Proposition 1. *Let $\gamma \in [-1, 0)$, $\alpha, \beta, c > 0$, let B be a behaviour and $\mathbb{P}$ be an unprompted language model such that B is α, β, γ-distinguishable in $\mathbb{P}$ (definition 3) and the distributions corresponding to the well-behaved and ill-behaved components of $\mathbb{P}$ are c-similar (definition 6). Suppose the model has been misaligned with a prompt s_0, such that $B_{\mathbb{P}}(s_0) \leq \gamma$. For the remainder of the conversation,*

$a_1 \oplus q_1 \oplus ... \oplus a_{n-1} \oplus q_n$*, it will remain misaligned:*

$$B_{\mathbb{P}}(s_0 \oplus a_1 \oplus q_1 \oplus ... \oplus a_{n-1} \oplus q_n) < \frac{\gamma}{2} \tag{9}$$

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

Intuitively, after the LLM has been misaligned by s_0, the model's responses in the following conversation maintain the negative behavior (more negative than $\gamma/2 < 0$), unless the conversation exceeds a certain length that scales linearly with the length of the misaligning 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 the pretraining phase (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's conditional behavior to resemble the ill-behaved persona's behavior:

Theorem 4. *Let* $\gamma \in [-1, 0)$, $\alpha, \beta, \epsilon > 0$, *and let* $\mathbb{P}$ *be a mixture of personas, that is c-similar,* $\mathbb{P} = \sum_{\phi\in\Phi} w_\phi \mathbb{P}_\phi$. *Then for every behavior* $B : \Sigma^\star \to [-1, 1]$ *that is* α, β, γ*-distinguishable in persona mixture* $\mathbb{P}$ *(definition 5), the distribution* $\mathbb{P}$ *is* γ*-prompt-misalignable (definition 1) with prompt length* $|s| = O(\log\frac{1}{\epsilon}, \frac{1}{\beta}, \log\frac{1}{\alpha}, c^2, \log|\Phi|)$.

We find that the length of a prompt required to get a LLM to imitate a persona of interest scales inversely with the distinguishability, β and logarithmically with the persona's prior in the distribution, w: $|s| = O(\frac{log\frac{1}{w}}{\beta})$. The prior dependence is due to the $O(log\frac{1}{\alpha})$ dependence in the theorem, as $\alpha \leq w_\phi$ for the ill-behaved persona ϕ (see definition 5).

Imitation of personas for "Jailbreaking" The consequence of the above theorem is that personas that have low priors w, 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 1. *Let* $\gamma \in [-1, 0)$, $\alpha, \beta, c > 0$, *let* $\mathbb{P}$ *be a mixture of personas, that is c-similar,* $\mathbb{P} = \sum_{\phi\in\Phi} w_\phi \mathbb{P}_\phi$ *and* $B : \Sigma^\star \to [-1, 1]$ *a behavior that is* α, β, γ*-distinguishable in persona mixture* $\mathbb{P}$. *If during training the distinguishability of ill-behaved personas scales super-logarithmically relative to their priors,* $\beta = \Omega(log(\frac{1}{w}))$, *invoking a persona for bad behavior requires prompts that are asymptotically shorter than ones for invoking a general bad behavior.*

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. 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 conversation, showing that the practice of misaligning via a multi-turn conversation an LLM 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 simple, 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. We leave it to future work to (i) empirically reinforce or weaken the likelihood that our assumptions are in fact plausible for actual LLM distributions of interest and (ii) make 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.

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

We are thankful for insightful conversations and comments by Oshri Avnery. 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.

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.

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.

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.

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 PROOF OF THEOREMS 1 AND 2

Theorem 1 is technically a special case of Theorem 2 which includes the case of an initial prompt s_0 before the misaligning prompt. Below we detail the proof of theorem 2, which entails that of theorem 1:

For convenience, we use $B' = -B$, $\gamma' = -\gamma$, in order to work with positive values. We start by writing the behavior expectation value given the conversation history s_0 and the new prompt s_1:

$$B'_{\mathbb{P}}(s_0 \oplus s_1) = \mathbb{E}_{s \sim \mathbb{P}(\cdot|s_0 \oplus s_1)}[B'(s)] = \sum_s \frac{\alpha \mathbb{P}_-(s_0 \oplus s_1 \oplus s) + (1-\alpha)\mathbb{P}_+(s_0 \oplus s_1 \oplus s)}{\alpha \mathbb{P}_-(s_0 \oplus s_1) + (1-\alpha)\mathbb{P}_+(s_0 \oplus s_1)} B'(s) >$$

The last transition is from the two component mixture assured by the α, β, γ-distinguishability. Taking only the $\mathbb{P}_-$ term in the numerator and rearranging the numerator and denominator:

$$> \sum_s \frac{\alpha \mathbb{P}_-(s_0 \oplus s_1 \oplus s)}{\alpha \mathbb{P}_-(s_0 \oplus s_1) + (1-\alpha)\mathbb{P}_+(s_0 \oplus s_1)} B'(s) = \sum_s \frac{\alpha \mathbb{P}_-(s_0 \oplus s_1 \oplus s)}{\alpha \mathbb{P}_-(s_0 \oplus s_1)(1 + \frac{(1-\alpha)\mathbb{P}_+(s_0 \oplus s_1)}{\alpha \mathbb{P}_-(s_0 \oplus s_1)})} B'(s) =$$

$$= \sum_s \frac{\mathbb{P}_-(s|s_0 \oplus s_1)}{(1 + \frac{(1-\alpha)\mathbb{P}_+(s_0 \oplus s_1)}{\alpha \mathbb{P}_-(s_0 \oplus s_1)})} B'(s)$$

If we were to bound $\frac{(1-\alpha)\mathbb{P}_+(s_0 \oplus s_1)}{\alpha \mathbb{P}_-(s_0 \oplus s_1)} < \epsilon'$, we would get:

$$> \sum_s \frac{\mathbb{P}_-(s|s_0 \oplus s_1)}{(1+\epsilon')} B'(s) = \frac{\mathbb{E}_{s \sim \mathbb{P}_-(\cdot|s_0 \oplus s_1)}[B'(s)]}{1+\epsilon'} = \frac{B'_{\mathbb{P}_-}(s_0 \oplus s_1)}{1+\epsilon'} > \frac{\gamma'}{1+\epsilon'} > \gamma'(1-\epsilon')$$

In the second to last transition we used $B'_{\mathbb{P}_-}(s_0 \oplus s_1) > \gamma'$ which is assured from the α, β, γ-distinguishability and the last transition is Bernouli's inequality. Taking $\epsilon' < \epsilon/\gamma'$ gives:

$$B'_{\mathbb{P}}(s_0 \oplus s_1) > \gamma' - \epsilon$$

If we multiply this inequality by -1 and map back $B = -B'$ and $\gamma = -\gamma'$, we get:

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

As required.

We are left with showing the bound $\frac{(1-\alpha)\mathbb{P}_+(s_0 \oplus s_1)}{\alpha \mathbb{P}_-(s_0 \oplus s_1)} < \epsilon'$. If we think of s_1 as a sequence of sentences, $s_1 = q_1 \oplus q_2 \oplus ... \oplus q_n$, then:

$$log \frac{\mathbb{P}_-(s_0 \oplus s_1)}{\mathbb{P}_+(s_0 \oplus s_1)} = log \frac{\mathbb{P}_-(s_0 \oplus q_1 \oplus q_2 \oplus ... \oplus q_n)}{\mathbb{P}_+(s_0 \oplus q_1 \oplus q_2 \oplus ... \oplus q_n)} = M_n$$

For $n = 1$, given the β-distinguishability from the use's prompts, we get:

$$\mathbb{E}_{q_1 \sim \mathbb{P}_-(\cdot|s_0)}[log \frac{\mathbb{P}_-(q|s_0)}{\mathbb{P}_+(q|s_0)}] > \beta$$

Where:

$$\mathbb{E}_{q_1 \sim \mathbb{P}_-(\cdot|s_0)}[log \frac{\mathbb{P}_-(q|s_0)}{\mathbb{P}_+(q|s_0)}] = \mathbb{E}_{q_1 \sim \mathbb{P}_-(\cdot|s_0)}[M_1] - M_0$$

Thus, in particular there exists q_1 that satisfies the condition:

$$M_1 > M_0 + \beta$$

For some $k \leq n$, given the β-distinguishability from q_k, we get:

$$\mathbb{E}_{q_k \sim \mathbb{P}_-(\cdot|s_0 \oplus q_1 \oplus ... \oplus q_{k-1})}[log \frac{\mathbb{P}_-(q|s_0 \oplus q_1 \oplus ... \oplus q_{k-1})}{\mathbb{P}_+(q|s_0 \oplus q_1 \oplus ... \oplus q_{k-1})}] > \beta$$

Where

$$\mathbb{E}_{q_k \sim \mathbb{P}_-(\cdot|s_0 \oplus q_1 \oplus ... \oplus q_{k-1})}[log \frac{\mathbb{P}_-(q|s_0 \oplus q_1 \oplus ... \oplus q_{k-1})}{\mathbb{P}_+(q|s_0 \oplus q_1 \oplus ... \oplus q_{k-1})}] = \mathbb{E}_{q_k \sim \mathbb{P}_-(\cdot|s_0 \oplus q_1 \oplus ... \oplus q_{k-1})}[M_k] - M_{k-1}$$

Thus, in particular there exists q_k that satisfies the condition:

$$M_k > M_{k-1} + \beta$$

Continuing this to $k = n$, given the choices for $q_1, ..., q_n$, we get in total:

$$M_n > n * \beta + M_0$$

We are left with bounding M_0. First, note that for the unprompted case (theorem 1), s_0 is empty, so there is nothing to bound. For the prompted case, we bound M_0 with c-similarity. The proof is given in the next appendix as a Lemma, $M_0 > -m * c$ where $m = |s_0|$. Choose $n > \frac{m*c+log(\frac{1-\alpha}{\alpha})+log(\frac{1}{\epsilon'})}{\beta}$ and get $\frac{(1-\alpha)\mathbb{P}_+(s_0 \oplus s_1)}{\alpha\mathbb{P}_-(s_0 \oplus s_1)} < \epsilon'$ as desired. Indeed $|s_1| = n$ is $O(\frac{1}{\beta}, |s_0|, log\frac{1}{\alpha}, log\frac{1}{\epsilon}, c)$.

B Lemma (bound for M_0 with c-similarity)

Denote the conversation history $s_0 = a_1 \oplus ... \oplus a_m$. Let $\mathbb{P}_\phi$ and $\mathbb{P}_{\tilde{\phi}}$ be c-similar distributions. Notice that:

$$M_0 = log\frac{P_\phi(a_1 \oplus ... \oplus a_m)}{P_{\tilde{\phi}}(a_1 \oplus ... \oplus a_m)} = \sum_i log\frac{P_\phi(a_i|a_1 \oplus ... \oplus a_{i-1})}{P_{\tilde{\phi}}(a_i|a_1 \oplus ... \oplus a_{i-1})} > -m * c$$

The last transition is from c-similarity. Proof of the third transition with induction:

- Base for induction $m' = 2$, follows from conditional probability of two variables:

$$log\frac{P_\phi(a_1 \oplus a_2)}{P_{\tilde{\phi}}(a_1 \oplus a_2)} = log\frac{P_\phi(a_2|a_1)}{P_{\tilde{\phi}}(a_2|a_1)} + log\frac{P_\phi(a_1)}{P_{\tilde{\phi}}(a_1)}$$

- Induction step: assume $m' - 1$ holds true, now for m':

$$log\frac{P_\phi(a_1 \oplus ... \oplus a_{m'})}{P_{\tilde{\phi}}(a_1 \oplus ... \oplus a_{m'})} = log\frac{P_\phi(a_{m'}|a_1 \oplus ... \oplus a_{m'-1})}{P_{\tilde{\phi}}(a_{m'}|a_1 \oplus ... \oplus a_{m'-1})} + log\frac{P_\phi(a_1 \oplus ... \oplus a_{m'-1})}{P_{\tilde{\phi}}(a_1 \oplus ... \oplus a_{m'-1})} =$$

  From the induction base for $m - 1$:

$$= log\frac{P_\phi(a_{m'}|a_1 \oplus ... \oplus a_{m'-1})}{P_{\tilde{\phi}}(a_{m'}|a_1 \oplus ... \oplus a_{m'-1})} + \sum_{i=1}^{m'-1} log\frac{P_\phi(a_i|a_1 \oplus ... \oplus a_{i-1})}{P_{\tilde{\phi}}(a_i|a_1 \oplus ... \oplus a_{i-1})} =$$

$$= \sum_{i=1}^{m'} log\frac{P_\phi(a_i|a_1 \oplus ... \oplus a_{i-1})}{P_{\tilde{\phi}}(a_i|a_1 \oplus ... \oplus a_{i-1})}$$

As desired.

C Proof of theorem 3

The proof is similar to theorem 2, but here there are two types of steps, one in which the user inserts a prompt q_i that increases the distinguishability by $\beta * |q_i|$ and a second in which the model responds a_i, possibly resisting the bad behavior prompting, worst case decreasing the distinguishability by $-c$. Following the exact same same steps as in theorem 2, we use $B' = -B$, $\gamma' = -\gamma$, in order to work with positive values. We start by writing the behavior expectation value given the conversation history s_0:

$$B'_{\mathbb{P}}(s_0) = \mathbb{E}_{s \sim \mathbb{P}(\cdot|s_0)}[B'(s)] = \sum_s \frac{\alpha\mathbb{P}_-(s_0 \oplus s) + (1-\alpha)\mathbb{P}_+(s_0 \oplus s)}{\alpha\mathbb{P}_-(s_0) + (1-\alpha)\mathbb{P}_+(s_0)} B'(s) >$$

The last transition is from the two component mixture assured by the α, β, γ-distinguishability. Taking only the $\mathbb{P}_-$ term in the numerator and rearranging the numerator and denominator:

$$> \sum_s \frac{\alpha\mathbb{P}_-(s_0 \oplus s)}{\alpha\mathbb{P}_-(s_0) + (1-\alpha)\mathbb{P}_+(s_0)} B'(s) = \sum_s \frac{\alpha\mathbb{P}_-(s_0 \oplus s)}{\alpha\mathbb{P}_-(s_0)(1 + \frac{(1-\alpha)\mathbb{P}_+(s_0)}{\alpha\mathbb{P}_-(s_0)})} B'(s) =$$

$$= \sum_s \frac{\mathbb{P}_-(s|s_0)}{(1 + \frac{(1-\alpha)\mathbb{P}_+(s_0)}{\alpha\mathbb{P}_-(s_0)})} B'(s)$$

If we were to bound $\frac{(1-\alpha)\mathbb{P}_+(s_0)}{\alpha\mathbb{P}_-(s_0)} < \epsilon'$, we would get:

$$> \sum_s \frac{\mathbb{P}_-(s|s_0)}{(1+\epsilon')} B'(s) = \frac{\mathbb{E}_{s\sim\mathbb{P}_-(\cdot|s_0)}[B'(s)]}{1+\epsilon'} = \frac{B'_{\mathbb{P}_-}(s_0)}{1+\epsilon'} > \frac{\gamma'}{1+\epsilon'} > \gamma'(1-\epsilon')$$

In the second to last transition we used $B'_{\mathbb{P}_-}(s_0) > \gamma'$ which is assured from the α, β, γ-distinguishability and the last transition is Bernouli's inequality. Taking $\epsilon' < \epsilon/\gamma'$ gives:

$$B'_{\mathbb{P}}(s_0) > \gamma' - \epsilon$$

If we multiply this inequality by -1 and map back $B = -B'$ and $\gamma = -\gamma'$, we get:

$$B_{\mathbb{P}}(s_0) < \gamma + \epsilon$$

As required.

Thus again we are left with showing the bound $\frac{(1-\alpha)\mathbb{P}_+(s_0)}{\alpha\mathbb{P}_-(s_0)} < \epsilon'$, but this time s_0 is a series of user prompts and model responses $s_0 = q_1 \oplus a_1 \oplus ... \oplus q_n \oplus a_n$. We will prove $(*)$ as a lemma in the next appendix:

$$log\frac{\mathbb{P}_-(s_0)}{\mathbb{P}_+(s_0)} = log\frac{\mathbb{P}_-(q_1 \oplus a_1 \oplus ... \oplus q_n \oplus a_n)}{\mathbb{P}_+(q_1 \oplus a_1 \oplus ... \oplus q_n \oplus a_n)} >_{(*)} \sum_{i=1}^{n}(\beta|q_i| - c)$$

If this holds, then by choosing $|q_i| > \frac{c}{\beta} + 1$, we obtain:

$$log\frac{\mathbb{P}_-(s_0)}{\mathbb{P}_+(s_0)} > \sum_{i=1}^{n}\beta = n * \beta$$

Taking $n > \frac{log\frac{1}{\epsilon} + log\frac{1-\alpha}{\alpha}}{\beta}$, we obtain:

$$\frac{(1-\alpha)\mathbb{P}_+(s_0)}{\alpha\mathbb{P}_-(s_0)} < \epsilon$$

As desired. Indeed, the sum of prompt lengths is $\sum_{i=1}^{n}|q_i| = O(\frac{c}{\beta}, log\frac{1}{\epsilon}, log\frac{1}{\alpha})$ and each prompt is of length $|q_i| = O(\frac{c}{\beta})$.

D Lemma (bound for component probability ratio in conversation)

Here we will prove the following inequality, for $\mathbb{P}_+, \mathbb{P}_-$ which are β-distinguishable and c-similar. There exists a choice of prompts $q_1 ... q_n$ such that:

$$log\frac{\mathbb{P}_-(q_1 \oplus a_1 \oplus ... \oplus q_n \oplus a_n)}{\mathbb{P}_+(q_1 \oplus a_1 \oplus ... \oplus q_n \oplus a_n)} > \sum_{i=1}^{n}(\beta|q_i| - c)$$

We do this by induction.

- Base of induction:

$$log\frac{\mathbb{P}_-(q_1 \oplus a_1)}{\mathbb{P}_+(q_1 \oplus a_1)} = log\frac{\mathbb{P}_-(a_1|q_1)}{\mathbb{P}_+(a_1|q_1)} + log\frac{\mathbb{P}_-(q_1)}{\mathbb{P}_+(q_1)} > -c + log\frac{\mathbb{P}_-(q_1)}{\mathbb{P}_+(q_1)}$$

  The first transition is from conditional probability, the second transition is from the c-similarity. Next, we need to construct an adversarial prompt q_1 that satisfies $log\frac{\mathbb{P}_-(q_1)}{\mathbb{P}_+(q_1)} > \beta \cdot |q_1|$. It is constructed sentence by sentence, $q_1 = s_1 \oplus ... \oplus s_{|q_1|}$. From β-distinguishability, $\mathbb{E}_{s_1\sim\mathbb{P}_-(\cdot)}[log\frac{\mathbb{P}_-(s_1)}{\mathbb{P}_+(s_1)}] > \beta$, thus, in particular there exists a sentence s_1 that satisfies $log\frac{\mathbb{P}_-(s_1)}{\mathbb{P}_+(s_1)} > \beta$. For some $k \leq |q_1|$, from β-distinguishability,

$\mathbb{E}_{s_k \sim \mathbb{P}_-(\cdot|s_1 \oplus ... \oplus s_{k-1})}[log\frac{\mathbb{P}_-(s_k|s_1 \oplus ... \oplus s_{k-1})}{\mathbb{P}_+(s_k|s_1 \oplus ... \oplus s_{k-1})}] > \beta$, thus, in particular there exists a sentence s_k that satisfies $log\frac{\mathbb{P}_-(s_k|s_1 \oplus ... \oplus s_{k-1})}{\mathbb{P}_+(s_k|s_1 \oplus ... \oplus s_{k-1})} > \beta$. Such that in total, $log\frac{\mathbb{P}_-(q_1)}{\mathbb{P}_+(q_1)} = \sum_{i=1}^{|q_1|} log\frac{\mathbb{P}_-(s_i|s_1 \oplus ... \oplus s_{i-1})}{\mathbb{P}_+(s_i|s_1 \oplus ... \oplus s_{i-1})} > |q_1| \cdot \beta$ as desired. Using this on the previous inequality, we get:

$$log\frac{\mathbb{P}_-(q_1 \oplus a_1)}{\mathbb{P}_+(q_1 \oplus a_1)} > -c + log\frac{\mathbb{P}_-(q_1)}{\mathbb{P}_+(q_1)} > \beta \cdot |q_i| - c$$

- Assume that the inequality holds for $k-1$, let us prove for $k \leq n$:

$$log\frac{\mathbb{P}_-(q_1 \oplus a_1 \oplus ... \oplus q_k \oplus a_k)}{\mathbb{P}_+(q_1 \oplus a_1 \oplus ... \oplus q_k \oplus a_k)} = log\frac{\mathbb{P}_-(a_k|q_1 \oplus a_1 \oplus ... \oplus q_k)}{\mathbb{P}_+(a_k|q_1 \oplus a_1 \oplus ... \oplus q_k)} + log\frac{\mathbb{P}_-(q_1 \oplus a_1 \oplus ... \oplus q_k)}{\mathbb{P}_+(q_1 \oplus a_1 \oplus ... \oplus q_k)}$$

From c-similarity on the first term:

$$> -c + log\frac{\mathbb{P}_-(q_1 \oplus a_1 \oplus ... \oplus q_k)}{\mathbb{P}_+(q_1 \oplus a_1 \oplus ... \oplus q_k)}$$

$$= -c + log\frac{\mathbb{P}_-(q_k|q_1 \oplus a_1 \oplus ... \oplus a_{k-1})}{\mathbb{P}_+(q_k|q_1 \oplus a_1 \oplus ... \oplus a_{k-1})} + log\frac{\mathbb{P}_-(q_1 \oplus a_1 \oplus ... \oplus a_{k-1})}{\mathbb{P}_+(q_1 \oplus a_1 \oplus ... \oplus a_{k-1})}$$

Next, we construct an adversarial prompt q_k such that

$$log\frac{\mathbb{P}_-(q_k|q_1 \oplus a_1 \oplus ... \oplus a_{k-1})}{\mathbb{P}_+(q_k|q_1 \oplus a_1 \oplus ... \oplus a_{k-1})} > \beta \cdot |q_k|$$

The idea is the same as for the base case of the induction, but for a rigorous proof, view next appendix. Using this on the previous inequality, we get:

$$log\frac{\mathbb{P}_-(q_1 \oplus a_1 \oplus ... \oplus q_k \oplus a_k)}{\mathbb{P}_+(q_1 \oplus a_1 \oplus ... \oplus q_k \oplus a_k)} > -c + \beta \cdot |q_k| + log\frac{\mathbb{P}_-(q_1 \oplus a_1 \oplus ... \oplus a_{k-1})}{\mathbb{P}_+(q_1 \oplus a_1 \oplus ... \oplus a_{k-1})}$$

From the base of the induction:

$$= -c + \beta \cdot |q_k| + \sum_{i=1}^{k-1}(\beta \cdot |q_i| - c)$$

Giving us:

$$log\frac{\mathbb{P}_-(q_1 \oplus a_1 \oplus ... \oplus q_k \oplus a_k)}{\mathbb{P}_+(q_1 \oplus a_1 \oplus ... \oplus q_k \oplus a_k)} > \sum_{i=1}^{k}(\beta \cdot |q_i| - c)$$

As desired.

D.1 Lemma (adversarial prompt construction)

Given s_0 and $\mathbb{P}_+, \mathbb{P}_-$ being β-distinguishable, we construct a prompt sentence by sentence, $q = s_1 \oplus ... \oplus s_{|q|}$. Such that

$$log\frac{\mathbb{P}_-(q_k|s_0)}{\mathbb{P}_-(q_k|s_0)} > \beta \cdot |q|$$

By induction:

- Base case: From β-distinguishability,

$$\mathbb{E}_{s_1 \sim \mathbb{P}_-(\cdot|s_0)}[log\frac{\mathbb{P}_-(s_1|s_0)}{\mathbb{P}_+(s_1|s_0)}] > \beta$$

Thus, in particular there exists a sentence s_1 that satisfies

$$log\frac{\mathbb{P}_-(s_1|s_0)}{\mathbb{P}_+(s_1|s_0)} > \beta$$

- Assume that the inequality holds for $k-1$, let us prove for $k \leq |q|$. From β-distinguishability,
$$\mathbb{E}_{s_k \sim \mathbb{P}_-(s_1 \oplus ... \oplus s_{k-1})}[log \frac{\mathbb{P}_-(s_k|s_1 \oplus ... \oplus s_{k-1})}{\mathbb{P}_+(s_k|s_1 \oplus ... \oplus s_{k-1})}] > \beta$$
Thus, in particular there exists a sentence s_k that satisfies
$$log \frac{\mathbb{P}_-(s_k|s_1 \oplus ... \oplus s_{k-1})}{\mathbb{P}_+(s_k|s_1 \oplus ... \oplus s_{k-1})} > \beta$$
Such that in total,
$$log \frac{\mathbb{P}_-(q|s_0)}{\mathbb{P}_+(q|s_0)} = \sum_{i=1}^{|q|} log \frac{\mathbb{P}_-(s_i|s_0 \oplus s_1 \oplus ... \oplus s_{i-1})}{\mathbb{P}_+(s_i|s_0 \oplus s_1 \oplus ... \oplus s_{i-1})} > |q| \cdot \beta$$
As desired.

E Proof of proposition 1

For convenience we will map $B' = -B$ and $\gamma' = -\gamma$ in order to work with positive values. The idea is to look at the behavior expectation value given the continuation of the conversation, $s_1 = q_1 \oplus a_1 \oplus ... \oplus q_n \oplus a_n$:

$$B'_{\mathbb{P}}(s_0 \oplus s_1) = \mathbb{E}_{s \sim \mathbb{P}(\cdot|s_0 \oplus s_1)}[B'(s)] = \sum_s \frac{\alpha \mathbb{P}_-(s_0 \oplus s_1 \oplus s) + (1-\alpha)\mathbb{P}_+(s_0 \oplus s_1 \oplus s)}{\alpha \mathbb{P}_-(s_0 \oplus s_1) + (1-\alpha)\mathbb{P}_+(s_0 \oplus s_1)} B'(s) >$$

The last transition if from the two-component mixture assured by the α, β, γ-distinguishability. We then take only the negative component term in the numerator:

$$> \sum_s \frac{\alpha \mathbb{P}_-(s_0 \oplus s_1 \oplus s)}{\alpha \mathbb{P}_-(s_0 \oplus s_1) + (1-\alpha)\mathbb{P}_+(s_0 \oplus s_1)} B'(s) = \sum_s \frac{\alpha \mathbb{P}_-(s_0 \oplus s_1 \oplus s)}{\alpha \mathbb{P}_-(s_0 \oplus s_1)(1 + \frac{(1-\alpha)\mathbb{P}_+(s_0 \oplus s_1)}{\alpha \mathbb{P}_-(s_0 \oplus s_1)})} B'(s) =$$
$$= \sum_s \frac{\mathbb{P}_-(s|s_0 \oplus s_1)}{(1 + \frac{(1-\alpha)\mathbb{P}_+(s_0 \oplus s_1)}{\alpha \mathbb{P}_-(s_0 \oplus s_1)})} B'(s)$$

Notice that if $\frac{(1-\alpha)\mathbb{P}_+(s_0 \oplus s_1)}{\alpha \mathbb{P}_-(s_0 \oplus s_1)} < 1$, then the denominator is smaller than 2, meaning that:

$$> \sum_s \frac{\mathbb{P}_-(s|s_0 \oplus s_1)}{2} B'(s) = \frac{1}{2} \sum_s \mathbb{P}_-(s|s_0 \oplus s_1) B'(s) = \frac{1}{2} B'_{\mathbb{P}_-}(s_0 \oplus s_1) > \gamma'/2$$

The last transition is from the α, β, γ-distinguishability, which assures $B_{\mathbb{P}_-}(s) < \gamma$ for any s. Going back to $B = -B'$ and $\gamma = -\gamma'$, we get:

$$B_{\mathbb{P}}(s_0 \oplus s_1) < \gamma/2$$

So let us see what is required for this condition to hold:

$$\frac{(1-\alpha)\mathbb{P}_+(s_0 \oplus s_1)}{\alpha \mathbb{P}_-(s_0 \oplus s_1)} < 1$$

iff:

$$log \frac{\mathbb{P}_+(s_0 \oplus s_1)}{\mathbb{P}_-(s_0 \oplus s_1)} < log \frac{\alpha}{1-\alpha}$$

Using c-similarity of $\mathbb{P}_+$ and $\mathbb{P}_-$ gives, $log \frac{\mathbb{P}_+(s_0 \oplus s_1)}{\mathbb{P}_-(s_0 \oplus s_1)} < c \cdot |s_1| + log \frac{\mathbb{P}_+(s_0)}{\mathbb{P}_-(s_0)}$ (similarly to lemma on bounding M_0). In the first part of the theorem, we showed that the adversarial prompt s_0 satisfies $log \frac{\mathbb{P}_+(s_0)}{\mathbb{P}_-(s_0)} < |s_0| \cdot \beta$. Combining these two, we get:

$$log \frac{\mathbb{P}_+(s_0 \oplus s_1)}{\mathbb{P}_-(s_0 \oplus s_1)} < c \cdot |s_1| - \beta \cdot |s_0| <_! log \frac{\alpha}{1-\alpha}$$

Such that the condition we require is assured to hold if:

$$|s_1| < \frac{\beta}{c}|s_0| + log \frac{\alpha}{1-\alpha}$$

This concludes our proof, that unless $|s_1| = \sum_{i=1}^{n} |q_i| + |a_i| = \Omega(|s_0|)$, then $B_{\mathbb{P}}(s_0 \oplus s_1) < \gamma/2$.

F Proof of theorem 4

Let $\epsilon > 0$, and let $\tilde{\phi}$ be the persona for which the conditions in the γ-distinguishable in persona mixture (see definition 5) holds with $\frac{\epsilon}{2}$. For convenience, we map $B' = -B$, $\gamma' = -\gamma$, in order to work with positive values, and start by writing the behavior expectation value given a prompt s_0:

$$B'_{\mathbb{P}}(s_0) = \mathbb{E}_{s\sim P(*|s_0)}[B'(s)] = \sum_s P(s|s_0)B'(s) =$$

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

$$= \sum_s \frac{\sum_\phi w_\phi P_\phi(s_0 \oplus s)}{\sum_\phi w_\phi P_\phi(s_0)} B'(s) > \sum_s \frac{w_{\tilde{\phi}} P_{\tilde{\phi}}(s_0 \oplus s)}{\sum_\phi w_\phi P_\phi(s_0)} B'(s) =$$

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

$$= \sum_s \frac{w_{\tilde{\phi}} P_{\tilde{\phi}}(s_0 \oplus s)}{w_{\tilde{\phi}} P_{\tilde{\phi}}(s_0)(1 + \sum_{\phi \neq \tilde{\phi}} \frac{w_\phi P_\phi(s_0)}{w_{\tilde{\phi}} P_{\tilde{\phi}}(s_0)})} B'(s) =$$

Since $\tilde{\phi}$ is c-similar to the other components in the mixture, we use the lemma on persona converging (next appendix): 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:

$$\geq \sum_s \frac{w_{\tilde{\phi}} P_{\tilde{\phi}}(s_0 \oplus s)}{w_{\tilde{\phi}} P_{\tilde{\phi}}(s_0)(1 + \epsilon' \sum_{\phi \neq \tilde{\phi}} \frac{w_\phi}{w_{\tilde{\phi}}})} B'(s) =$$

From the α, β, γ-distinguishability, $w_{\tilde{\phi}} > \alpha$ and $\sum_{\phi \neq \tilde{\phi}} w_\phi < 1$, we get:

$$\geq \sum_s \frac{w_{\tilde{\phi}} P_{\tilde{\phi}}(s_0 \oplus s)}{w_{\tilde{\phi}} P_{\tilde{\phi}}(s_0)(1 + \frac{\epsilon'}{\alpha})} B'(s) = \frac{1}{1 + \frac{\epsilon'}{\alpha}} \sum_s P_{\tilde{\phi}}(s|s_0) B'(s) = \frac{1}{1 + \frac{\epsilon'}{\alpha}} \mathbb{E}_{s\sim P_{\tilde{\phi}}(*|s_0)}[B'(s)] =$$

$$= \frac{1}{1 + \frac{\epsilon'}{\alpha}} B'_{\mathbb{P}_{\tilde{\phi}}}(s_0) > \frac{\gamma'}{1 + \frac{\epsilon'}{\alpha}}$$

The last transition, was since the α, β, γ-distinguishability assures $B'_{\mathbb{P}_{\tilde{\phi}}}(s_0) > \gamma'$. With Bernouli's inequality:

$$\geq \gamma(1 - \frac{\epsilon'}{\alpha})$$

Taking $\epsilon' = \frac{2\alpha\epsilon}{\gamma}$, we get:

$$B'_{\mathbb{P}}(s_1) \geq \gamma' - \epsilon$$

Mapping back $B = -B', \gamma = -\gamma'$, we get:

$$B_{\mathbb{P}}(s_1) \leq \gamma + \epsilon$$

As desired. Notice that $|s_0| = O(log\frac{1}{\epsilon'}) = O(log\frac{1}{\epsilon}, log\frac{1}{\alpha})$, also from the condition on the lemma, $|s_0| = O(\frac{1}{\beta}, c^2, \log|\Phi|)$

G Lemma (persona converging)

Here, 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 1. *Let* $\beta, \epsilon > 0$ *and* mixture of personas $P = \sum_{\phi\in\Phi} w_\phi P_\phi$ *then for each* β*-Martingale-distinguishable persona* $\tilde{\phi}$*, there exists a prompt* $s_0 \in \Sigma^*$ *such that:*

$$\forall \phi \neq \tilde{\phi} \quad \frac{P_\phi(s_0)}{P_{\tilde{\phi}}(s_0)} < \epsilon \tag{10}$$

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 of lemma:

Intuitively, we will use the probabilistic method and prove that the probability of s_0 for which $\frac{P_\phi(s_0)}{P_{\tilde{\phi}}(s_0)} < \epsilon$ uphold simultaneously for any ϕ is greater than zero and hence such s_0 exists. Specifically, let ϕ from some other persona such that $\tilde{\phi}$ is β-Martingale-distinguishable from ϕ. For a prompt Q composed of n sentences, $Q = q_1 \oplus ... \oplus q_n$, denote by:

$$M_n^{\tilde{\phi},\phi} = \log \frac{P_{\tilde{\phi}}(q_1 \oplus ... \oplus q_n)}{P_\phi(q_1 \oplus ... \oplus q_n)}$$

Then, since $\tilde{\phi}$ is β-Martingale-distinguishable from ϕ we have that:

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

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:

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

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

$$\begin{aligned}
&\mathbb{E}_{s_{n+1} \sim P_{\tilde{\phi}}(\cdot)}[Z_{n+1} | Z_1 = z_1 ... Z_n = z_n] \\
&= \mathbb{E}_{s_{n+1} \sim P_{\tilde{\phi}}(\cdot)}[M_{n+1}^{\tilde{\phi},\phi} | M_1^{\tilde{\phi},\phi} = m_1 ... M_n^{\tilde{\phi},\phi} = m_n] - (n+1)\beta \\
&> m_n + \beta - (n+1)\beta = m_n + \beta n \\
&= z_n
\end{aligned}$$

In addition, Z_n is bounded since from c-similarity we have that:

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

And therefore:

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

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

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

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:

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

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:

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

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:

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

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 $\phi \neq \tilde{\phi}$:

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

And the user may choose it.