Stopping LLMs from Going Rogue: A Control Barrier Approach to Text Generation

João Pedro Silvestre[1], Álvaro Rodríguez Abella[1] and Paulo Tabuada[1]

***Abstract*— The rapid integration of large language models (LLMs) into our everyday lives has outpaced safety considerations aimed at protecting users from toxic outputs and preventing malicious actors from generating harmful text at scale. As a result, LLMs have been exploited by bots capable of producing vast amounts of harmful and toxic content, enabling users to manipulate online opinions and, in some cases, create dangerous online environments.**

Our work addresses this issue by developing a framework for designing safety filters that preclude toxic outputs. To achieve this, we leverage Control Barrier Functions (CBFs) which enable the design of closed-loop systems that remain safe. We consider the continuous-time model of an LLM, where tokens are regarded as the state of the model, and prove that by only controlling the first token, any function satisfying mild assumptions becomes a CBF. Our approach can be utilized to design LLMs capable of ensuring safety of its outputs without significantly affecting the original model's behavior.

I. Introduction

The introduction of transformers [1] marked a turning point in natural language processing, with early works such as the GPT [2] already representing a breakthrough over older techniques such as the Long Short Term Memory models [3]. Increasingly sophisticated models, such as GPT-4 [4] and DeepSeek [5] have recently emerged, capable of generating text nearly indistinguishable from human output [6]. Unfortunately, due to the rapid speed at which progress has been made, safety concerns have been overshadowed. One work which details some overlooked problems is [7], where it is shown that large language models (LLMs) may generate toxic outputs, from apparently innocuous prompts.

Our understanding of these systems has lagged behind the massive increases in their size and complexity, leaving no other option but to treat them as black boxes [8]. This motivated two distinct research approaches. On the empirical side, many studies analyze tokens at each layer to decipher model behavior [9]. Other recent studies have explored the capacity that LLMs have to generate self-explanations, mainly showing that these explanations may still not be trusted [10]. These type of approaches have been leveraged to design adaptations to the model to address the toxicity problem [11]. However, due to their empirical nature, it remains a challenge to properly control LLMs and stop toxic text generation while providing any type of formal guarantees. In fact, some works even suggest safeguards may be easily bypassed [12], [13].

On the theoretical side, recent works have leveraged the mathematical model proposed in [**?**] to analyze LLM behavior and derive asymptotic conclusions [14], [15]. Notably, to derive such results, these works typically make simplifying assumptions that limit their direct use, such as the absence of feedforward layers. Nevertheless, such methods allow the interpretation of LLMs as dynamical systems and, once practical limitations are overcome, control techniques will enable the improvement of model safety.

While some works have used control tools such as optimal control to tackle the toxicity problem [16], one particular concept that immediately comes to mind is safety. By separating the state space into safe and unsafe sets, and by constraining tokens to the safe set, it may be possible to preclude toxic outputs. Seminal works such as [17] have laid the groundwork for modern safety techniques, typically based on control barrier functions (CBFs).

Some works have used a CBF approach to the toxicity problem [18]. However, unlike our approach, they act after the model has computed the probability distribution of the next token, filtering that distribution if necessary.

In this paper, tokens leading to toxic text are considered unsafe states and sufficient conditions for designing CBFs that constrain tokens within the safe states are established. To achieve this, we first extend the continuous-time model derived in [**?**] to better reflect the actual transformer model. Furthermore, we assume that the first token can be controlled throughout the model for any given prompt. We then prove that any generic function can serve as a CBF, provided that there always exists a direction in which its value changes. Therefore, our slight modification of the original model is sufficient to ensure that the output never becomes toxic.

Our contributions are threefold:

1) The continuous-time model of an LLM [**?**] is extended to model feedforward layers.
2) Sufficient conditions for designing CBFs preventing toxic outputs are presented.
3) It is proved that controlling only the first token is sufficient to keep the generated tokens within the safe set.

Notations

Let $r, s, k \in \mathbb{N} = \{1, 2, \dots\}$. Points in the Euclidean space $\mathbb{R}^r$ are regarded as column vectors and denoted by $\mathrm{x} = (x^1, \dots, x^r) \in \mathbb{R}^r$. Tuples of k points are denoted by

The work of João Pedro Silvestre was partially supported by the PhD fellowship 2023.01843.BD from the Fundação para a Ciência e a Tecnologia (FCT), Portugal.

[1]João Pedro Silvestre, Álvaro Rodríguez Abella, and Paulo Tabuada are with the Electrical and Computer Engineering Department, University of California at Los Angeles, Los Angeles, CA 90095 USA (e-mail: {joaosilvestre, rodriguezabella, tabuada}@ucla.edu).

$\mathbf{x} = (\mathrm{x}_1, \ldots, \mathrm{x}_k) \in (\mathbb{R}^r)^k$ (note the different font). Similarly, the space of $r \times s$ real matrices is denoted by $\mathbb{R}^{r\times s}$. In particular, $\mathbb{I}_r \in \mathbb{R}^{r\times r}$ denotes the identity matrix. Tuples of matrices are denoted by $\mathbf{A} = (A_1, \ldots, A_k) \in (\mathbb{R}^{r\times s})^k$. Lastly, the tangent space of a smooth manifold M at $p \in M$ and its elements are denoted by T_pM and $X_p \in T_pM$, respectively. Given another smooth manifold N and a smooth map $\phi : M \to N$, *i.e.*, $\phi \in C^\infty(M, N)$, its tangent map is denoted by $T\phi : TM \to TN$.

Preliminaries on safety and control barrier functions

In this section we give a brief overview of control barrier functions (CBFs) to better contextualize our work. First, consider a nonlinear control system of the form:

$$\dot{\mathrm{x}} = f(t, \mathrm{x}) + g(t, \mathrm{x}) \cdot \mathrm{u}, \quad (t, \mathrm{x}) \in \mathbb{R}_{\geq 0} \times \mathbb{R}^{n+1}, \tag{1}$$

where $\mathrm{u} \in \mathbb{R}^m$ is the input of the system, $m \in \mathbb{N}$, $f : \mathbb{R}_{\geq 0} \times \mathbb{R}^{n+1} \to \mathbb{R}^{n+1}$ and $g : \mathbb{R}_{\geq 0} \times \mathbb{R}^{n+1} \to \mathbb{R}^{(n+1)\times m}$ are piece-wise continuous in t and locally Lipschitz continuous in x, and $\mathbb{R}_{\geq 0}$ denotes the non-negative reals. Henceforth, we assume that (1) is forward-complete, i.e., for each $\mathrm{x}_0 \in \mathbb{R}^{n+1}$ and piece-wise continuous input $u : \mathbb{R}_{\geq 0} \to \mathbb{R}^{n+1}$, the solution of the controlled system, $\dot{\mathrm{x}} = f(t, \mathrm{x}) + g(t, \mathrm{x}) \cdot \mathrm{u}(t)$, $(t, \mathrm{x}) \in \mathbb{R}_{\geq 0} \times \mathbb{R}^{n+1}$, with initial condition x_0 at $t_0 = 0$ is uniquely defined for all $t \in \mathbb{R}_{\geq 0}$.

Definition 1.1: A set $\mathcal{C} \subset \mathbb{R}^{n+1}$ is *forward invariant* if, for each initial condition $\mathrm{x}_0 \in \mathcal{C}$, the solution of (1) with initial condition x_0 at $t_0 = 0$ remains in $\mathcal{C}$ for each $t \in \mathbb{R}_{\geq 0}$. Furthermore, a set $\mathcal{C}$ is said to be *safe* for the system (1) if it is forward invariant.

Control barrier functions [17] provide a structured method to design controllers that guarantee safety for sets of the form $\mathcal{S} = h^{-1}(\mathbb{R}_{\geq 0})$, where $h \in C^1(\mathbb{R}^{n+1})$. Henceforth, the Lie derivative of h with respect to f is denoted by $L_f h = \mathrm{d}h(f)$. Analogously, we denote $L_g h = (L_{g_1} h, \ldots, L_{g_m} h)$, where $g = (g_1, \ldots, g_m) \in \mathbb{R}^{(n+1)\times m}$.

Definition 1.2 (Control Barrier Function [17]): A function $h \in C^1(\mathbb{R}^{n+1})$ is a *control barrier function (CBF)* for (1) on $\mathcal{S} = h^{-1}(\mathbb{R}_{\geq 0})$ if there exists[1] $\alpha \in \mathcal{K}^e_\infty$ such that, for each $(t, \mathrm{x}) \in \mathbb{R}_{\geq 0} \times \mathcal{S}$, we have:

$$\sup_{\mathrm{u}\in\mathbb{R}^m} (L_f h + L_g h \cdot \mathrm{u}) \geq -\alpha \circ h.$$

Note that, although h is time-independent, the Lie derivatives $L_f h$ and $L_g h$ depend on t as the vector fields f and g are time-dependent. The existence of a CBF implies the non-emptiness of the set of safe inputs (the argument $(t, \mathrm{x}) \in \mathbb{R}_{\geq 0} \times \mathbb{R}^{n+1}$ is omitted for brevity):

$$K_{\mathrm{CBF}} = \{\mathrm{u} \in \mathbb{R}^m \mid L_f h + L_g h \cdot \mathrm{u} \geq -\alpha \circ h\},$$

Lastly, [17, Theorem 2] ensures that any safe controller $k : \mathbb{R}_{\geq 0} \times \mathcal{S} \to \mathbb{R}^m$, i.e., one satisfying $k(t, \mathrm{x}) \in K_{\mathrm{CBF}}(t, \mathrm{x})$ for each $(t, \mathrm{x}) \in \mathbb{R}_{\geq 0} \times \mathcal{S}$ guarantees safety. In practice, if there are no additional constraints on the input, a sufficient condition for h to be a CBF is $(L_g h)(t, \mathrm{x}) \neq 0$ for each $(t, \mathrm{x}) \in \mathbb{R}_{\geq 0} \times \mathcal{S}$.

Higher-order barrier functions [19] constitute an extension of CBFs that are useful when the Lie derivative $L_g h$ vanishes and, thus, h is not a CBF in the standard sense. In this work, we will only need second-order CBFs. Given $h \in C^2(\mathbb{R}^{n+1})$ and $\alpha \in \mathcal{K}^e_\infty \cap C^1(\mathbb{R})$, we define the extension $h_1 \in C^1(\mathbb{R}_{\geq 0} \times \mathbb{R}^{n+1})$ of h by:

$$h_1(t, \mathrm{x}) = (\mathrm{d}h)_{\mathrm{x}}(f(t, \mathrm{x})) + \alpha(h(\mathrm{x})). \tag{2}$$

Similarly, let $h_1^{-1}(\mathbb{R}_{\geq 0})^\circ = \{\mathrm{x} \in \mathbb{R}^{n+1} \mid h_1(0, \mathrm{x}) \geq 0\}$.

Definition 1.3 (Second-order CBF, adapted from [20]): A *second-order control barrier function* for (1) is a function $h \in C^2(\mathbb{R}^2)$ for which there exist $\alpha \in \mathcal{K}^e_\infty \cap C^1(\mathbb{R})$, $\alpha_1 \in \mathcal{K}^e_\infty$ and $\mathcal{D} \subset \mathbb{R}^{n+1}$ open such that:

$$\mathcal{S} = h^{-1}(\mathbb{R}_{\geq 0}) \cap h_1^{-1}(\mathbb{R}_{\geq 0})^\circ \subset \mathcal{D},$$

with h_1 as in (2), and we have that $L_g h = 0$ and:

$$\sup_{\mathrm{u}\in\mathbb{R}^m} (L_f h_1 + L_g h_1 \cdot \mathrm{u}) \geq -\alpha_1 \circ h_1. \tag{3}$$

In [20, Theorem 1] it is shown that the set $\mathcal{S}$ is safe provided the control input belongs to the following set (again, we omit the argument $(t, \mathrm{x}) \in \mathbb{R}_{\geq 0} \times \mathcal{D}$ for brevity):

$$K_{\mathrm{CBFH}} = \{\mathrm{u} \in \mathbb{R}^m \mid L_f h_1 + L_g h_1 \cdot \mathrm{u} \geq -\alpha \circ h_1\}.$$

Similarly to standard CBFs, the condition $(L_g h_1)(t, \mathrm{x}) \neq 0$ for each $(t, \mathrm{x}) \in \mathbb{R}_{\geq 0} \times \mathcal{D}$ is sufficient to prove that h is a second-order CBF for (1) (provided $(L_g h)(t, \mathrm{x}) = 0$).

II. Controlled dynamics of transformers

In the following, let $n, d, k, h \in \mathbb{N}$ be fixed parameters. In this section, we start by defining the configuration space of the transformer, which is essentially the ellipsoid onto which the normalization at the end of each layer projects to. Afterward, we detail the attention mechanism present in every modern transformer and conclude with the discrete-time model for the transformer as well as the corresponding continuous-time first-order approximation. Readers not interested in the intricacies of transformers can directly jump to the continuous-time model (7).

A. Configuration space

Let $n, d, k, h \in \mathbb{N}$. A symmetric, positive-definite matrix $W \in \mathbb{R}^{(n+1)\times(n+1)}$ defines an inner product on $\mathbb{R}^{n+1}$:

$$\langle X_{\mathrm{x}}, Y_{\mathrm{x}} \rangle_W = X_{\mathrm{x}}^\top \, W \, Y_{\mathrm{x}},$$

for each $X_{\mathrm{x}}, Y_{\mathrm{x}} \in T_{\mathrm{x}}\mathbb{R}^{n+1}$ and $\mathrm{x} \in \mathbb{R}^{n+1}$. The corresponding norm is denoted by $|X_{\mathrm{x}}|_W = (X_{\mathrm{x}}^\top \, W \, X_{\mathrm{x}})^{1/2}$. The points of $\mathbb{R}^{n+1}$ of unit norm with respect to W define an n-dimensional ellipsoid, which is denoted by:

$$\mathcal{E}^n_W = \{\mathrm{y} \in \mathbb{R}^{n+1} \mid \mathrm{y}^\top \, W \mathrm{y} = 1\}.$$

As we consider a transformer consisting of k tokens, the resulting state space is the Cartesian product of k copies of the ellipsoid, *i.e.*, $(\mathcal{E}^n_W)^k = \underbrace{\mathcal{E}^n_W \times \ldots \times \mathcal{E}^n_W}_{k\text{-times}}$.

[1] An extended class $\mathcal{K}$ function, $\alpha \in \mathcal{K}^e_\infty$, is a strictly increasing function $\alpha : \mathbb{R} \to \mathbb{R}$ such that $\alpha(0) = 0$, $\lim_{r\to\infty} \alpha(r) = \infty$ and $\lim_{r\to-\infty} \alpha(r) = -\infty$.

Similarly, consider the ellipsoid projection:

$$\pi_W : \mathbb{R}_0^{n+1} \to \mathcal{E}_W^n, \quad \mathrm{x} \mapsto \pi_W(\mathrm{x}) = \mathrm{x}\,|\mathrm{x}|_W^{-1},$$

where $\mathbb{R}_0^{n+1} = \mathbb{R}^{n+1} - \{0\}$. Its tangent map at each $\mathrm{x} \in \mathbb{R}_0^{n+1}$, $T_{\mathrm{x}}\pi_W : T_{\mathrm{x}}\mathbb{R}_0^{n+1} \to T_{\pi_W(\mathrm{x})}\mathcal{E}_W^n$, is given by:

$$T_{\mathrm{x}}\pi_W \cdot X_{\mathrm{x}} = |\mathrm{x}|_W^{-1} \left(\mathbb{I}_{n+1} - \mathrm{x}\,\mathrm{x}^\top W\,|\mathrm{x}|_W^{-2}\right) \cdot X_{\mathrm{x}},$$

for each $X_{\mathrm{x}} \in T_{\mathrm{x}}\mathbb{R}_0^{n+1}$. In particular, for $\mathrm{y} \in \mathcal{E}_W^n$, we have $T_{\mathrm{y}}\pi_W \cdot X_{\mathrm{y}} = \left(\mathbb{I}_{n+1} - \mathrm{y}\,\mathrm{y}^\top W\right) \cdot X_{\mathrm{y}}$.

The standard Euclidean metric on $\mathbb{R}_0^{n+1}$ induces a Riemannian metric g on $\mathcal{E}_W^n$, whose Levi–Civita connection is denoted by ∇^{g}. In turn, the induced metric on the product manifold $(\mathcal{E}_W^n)^k$ and its Levi–Civita connection are, respectively, denoted by $\mathbf{g} = \underbrace{\mathrm{g} \oplus \ldots \oplus \mathrm{g}}_{k-\text{times}}$ and $\nabla^{\mathbf{g}}$.

B. Self-attention mechanism

Let $\mathbf{y} \in (\mathcal{E}_W^n)^k$ and $\mathbf{P} \in \left(\mathbb{R}^{(n+1)\times(n+1)}\right)^h$. For each head $1 \le \eta \le h$, the *η-th (full) self attention matrix* is defined as:

$$A_\eta = \left(\alpha_{ij}^\eta\right)_{\substack{2\le i\le k,\\ 1\le j\le k}} \in \mathbb{R}^{(k-1)\times k},$$

where, for each $1 \le i, j \le k$, $i \ne 1$:

$$\alpha_{ij}^\eta(\mathbf{P}, \mathbf{y}) = \frac{1}{Z_i^\eta(\mathbf{P}, \mathbf{y})}\, \exp(\mathrm{y}_i^\top\, P_\eta\, \mathrm{y}_j),$$

$$Z_i^\eta(\mathbf{P}, \mathbf{y}) = \sqrt{n+1} \sum_{j=1}^{k} \exp(\mathrm{y}_i^\top\, P_\eta\, \mathrm{y}_j).$$

The first token is excluded since its dynamics will be controlled, as explained in the next section. Alternatively, we may consider the so-called *auto-regressive* (also known as *causal* or *masked*) *self-attention*. In that case, the sum in the definition of Z_i^η runs from $j = 1$ to i, and $\alpha_{ij}^\eta = 0$ whenever $j > i$. The results in this paper are valid for both the full and the auto-regressive architectures, so we do not explicitly differentiate between them.

C. Discrete-time transformer model

While there are several transformer models, most of them inherit the basic structure of the original model [1]. In essence, as detailed in [21], each transformer layer is composed of two parts: the feed-forward layer and the self-attention mechanism, each with its own skip connection and normalization. The model presented in [**?**] made the simplifying assumption that feedforward layers were absent. In this section we present a discrete-time model including both layers.

Formally, for each $l \in \mathbb{N}$ and $1 \le \eta \le h$, let:

$$P_\eta(l), U_\eta(l) \in \mathbb{R}^{(n+1)\times(n+1)},$$
$$W_1(l), W_2^\top(l) \in \mathbb{R}^{d\times(n+1)},$$
$$b_1(l) \in \mathbb{R}^d,\ b_2(l) \in \mathbb{R}^{n+1}.$$

The discrete *self-attention* and *feed-forward* maps:

$$\tilde{f}_{\mathrm{sa},i} : \mathbb{N} \times (\mathcal{E}_W^n)^k \times \mathbb{R}_{\ge 0} \to \mathbb{R}^{n+1},$$
$$\tilde{f}_{\mathrm{ff}} : \mathbb{N} \times \mathcal{E}_W^n \times \mathbb{R}_{\ge 0} \to \mathbb{R}^{n+1},$$

$1 < i \le k$, are defined as:

$$\tilde{f}_{\mathrm{sa},i}(l, \mathbf{y}, \tau) = \mathrm{y}_i + \tau \sum_{\eta=1}^{h} \sum_{j=1}^{k} \alpha_{ij}^\eta(\mathbf{P}(l), \mathbf{y})\, U_\eta(l)\, \mathrm{y}_j,$$
$$\tilde{f}_{\mathrm{ff}}(l, \mathrm{y}, \tau) = W_2(l)\, \Sigma\big(W_1(l)\, \mathrm{y} + b_1(l)\big) + b_2(l),$$

respectively. In the previous expressions, $\tau \in \mathbb{R}_{\ge 0}$ is a small training weight usually absorbed into the matrices $U_\eta(l)$, $1 \le \eta \le h$. In turn, we denote by $f_{\mathrm{sa},i} = \pi_W \circ \tilde{f}_{\mathrm{sa},i}$ and $f_{\mathrm{ff}} = \pi_W \circ \tilde{f}_{\mathrm{ff}}$ the corresponding normalized maps. In the previous expression, we denote $\Sigma(\mathrm{z}) = (\sigma(z^1), \ldots, \sigma(z^d))$ for each $\mathrm{z} = (z^1, \ldots, z^d) \in \mathbb{R}^d$, where $\sigma : \mathbb{R} \to \mathbb{R}$ is some *activation function*.

To allow for designing CBFs, we alter the transformer architecture by assuming we can arbitrarily steer the first token. This approach has three main advantages: (1) we minimize the number of controlled tokens, (2) we can better approximate the original model by selecting a control input which is as close to the original model dynamics as possible, and (3) this control input is sufficient to keep the system safe, as will be shown in Section III.

Thus, for each $\mathbf{y} \in (\mathcal{E}_W^n)^k$, $\mathrm{u} \in \mathbb{R}^{n+1}$, $k \in \mathbb{N}$ and $\tau > 0$, the *discrete-time model for the controlled transformer* reads:

$$\mathrm{y}_i(l+1) = \pi_W\big(\mathrm{y}_i(l) + \tau\, \mathrm{u}\big), \tag{4}$$

if $i = 1$, and

$$\mathrm{y}_i(l+1) = \pi_W\big(f_{\mathrm{sa},i}(l, \mathbf{y}, \tau) + \tau\, \tilde{f}_{\mathrm{ff}}(l, f_{\mathrm{sa},i}(l, \mathbf{y}, \tau))\big), \tag{5}$$

for $1 < i \le k$.

Remark 2.1: This choice of architecture is conceptually similar to prefixing [22], a technique where an additional prefix is added to the LLM's prompt to steer it away from toxicity. Our approach is a form of pre-fixing where we only use one token but, rather than keeping it constant, we steer it to keep the system safe. Hence, we propose to alter the architecture by adding a controller that steers the single token pre-fixing the prompt. The design of such controller is outside the scope of this paper.

D. Continuous-time attention model

Let (M, g) be a Riemannian manifold (recall that, for the ellipsoid, $M = \mathcal{E}_W^n$, we utilize the Riemannian metric induced by the Euclidean metric on $\mathbb{R}_0^{n+1}$). The flow of a (forward-complete) vector field $Y \in \mathfrak{X}(M)$ is denoted by $Y^\tau : M \to M$, $\tau \ge 0$. A map $\phi : M \times \mathbb{R}_{\ge 0} \to M$ is a first order approximation to Y^τ if there exist $T > 0$ and $\xi : M \to \mathbb{R}_{\ge 0}$ such that $\mathrm{d}_{\mathrm{g}}(Y^\tau(m), \phi(m, \tau)) \le \xi(m)\, \tau^2$ for each $\tau \in [0, T]$ and $m \in M$, where d_{g} is the Riemannian distance on M.

Let us compute the best first order approximation of the discrete model introduced in the previous section. Let $(t, \mathbf{y}) \in \mathbb{R}_{\ge 0} \times (\mathcal{E}_W^n)^k$. For $i = 1$, we have:

$$\left.\frac{d}{d\tau}\right|_{\tau=0} \pi_W\big(\mathrm{y}_1(t) + \tau\, \mathrm{u}\big) = T_{\mathrm{y}_1}\pi_W \cdot \mathrm{u}.$$

Similarly, for $1 < i \leq k$, it is readily obtained:

$$\frac{d}{d\tau}\bigg|_{\tau=0} \pi_W\big(f_{\mathrm{sa},i}(t,\mathbf{y},\tau) + \tau\, \tilde{f}_{\mathrm{ff}}(t, f_{\mathrm{sa},i}(t,\mathbf{y},\tau))\big) = T_{\mathrm{y}_i}\pi_W \cdot \Big(\tilde{f}_{\mathrm{sa},i}(t,\mathbf{y}) + \tilde{f}_{\mathrm{ff}}(t,\mathrm{y}_i)\Big).$$

In the previous expression, the continuous *self-attention* map, $\tilde{f}_{\mathrm{sa},i} : \mathbb{R}_{\geq 0} \times (\mathcal{E}_W^n)^k \to \mathbb{R}^{n+1}$, and *feed-forward* map, $\tilde{f}_{\mathrm{ff}} : \mathbb{R}_{\geq 0} \times \mathcal{E}_W^n \to \mathbb{R}^{n+1}$, are defined as:

$$\tilde{f}_{\mathrm{sa},i}(t,\mathbf{y}) = \sum_{\eta=1}^{h}\sum_{j=1}^{k} \alpha_{ij}^{\eta}(\mathbf{P}(t),\mathbf{y})\, U_\eta(t)\, \mathrm{y}_j,$$
$$\tilde{f}_{\mathrm{ff}}(t,\mathrm{y}) = W_2(t)\,\Sigma\big(W_1(t)\,\mathrm{y} + b_1(t)\big) + b_2(t),$$

respectively. In turn, the corresponding normalized maps are respectively denoted by $f_{\mathrm{sa},i}(t,\mathbf{y}) = T_{\mathrm{y}_i}\pi_W \cdot \tilde{f}_{\mathrm{sa},i}(t,\mathbf{y})$ and $f_{\mathrm{ff}}(t,\mathrm{y}) = T_{\mathrm{y}}\pi_W \cdot \tilde{f}_{\mathrm{ff}}(t,\mathrm{y})$.

By gathering the previous expressions, we conclude that, for each $(t,\mathbf{y},\mathrm{u}) \in \mathbb{R}_{\geq 0} \times (\mathcal{E}_W^n)^k \times \mathbb{R}^{n+1}$, the best first order approximation in τ of the discrete model (4), (5) is given by:

$$\boxed{\dot{\mathrm{y}}_i = \begin{cases} T_{\mathrm{y}_1}\pi_W \cdot \mathrm{u}, & i = 1, \\ f_{\mathrm{sa},i}(t,\mathbf{y}) + f_{\mathrm{ff}}(t,\mathrm{y}_i), & 1 < i \leq k. \end{cases}} \tag{6}$$

By denoting:

$$f_{\mathrm{sa}}(t,\mathbf{y}) = \big(0, f_{\mathrm{sa},2}(t,\mathbf{y}), \ldots, f_{\mathrm{sa},k}(t,\mathbf{y})\big),$$
$$\boldsymbol{f}_{\mathrm{ff}}(t,\mathbf{y}) = \big(0, f_{\mathrm{ff}}(t,\mathrm{y}_2), \ldots, f_{\mathrm{ff}}(t,\mathrm{y}_k)\big),$$
$$g(t,\mathrm{y}) = \big(T_{\mathrm{y}}\pi_W, 0, \ldots, 0\big),$$

the previous model may be equivalently expressed as:

$$\boxed{\dot{\mathbf{y}} = f_{\mathrm{sa}}(t,\mathbf{y}) + \boldsymbol{f}_{\mathrm{ff}}(t,\mathbf{y}) + g(t,\mathrm{y}_1)\cdot \mathrm{u}.} \tag{7}$$

Let $f : \mathbb{R}_{\geq 0} \times (\mathcal{E}_W^n)^k \to T(\mathcal{E}_W^n)^k$ be the time-dependent vector field defined as $f(t,\mathbf{y}) = f_{\mathrm{sa}}(t,\mathbf{y}) + \boldsymbol{f}_{\mathrm{ff}}(t,\mathbf{y})$.

When comparing this model to the one in [**?**], the main difference is the additional sum of the feedforward layer and the control input. Furthermore, thanks to our choices, the system (7) is control affine, which simplifies the computation of safe control inputs and ensures that the set of safe inputs K_{CBF} is convex.

Remark 2.2: Note that (7) is forward-complete, since the vector fields are smooth and the system evolves on a compact manifold without boundary.

III. Safety of controlled LLMs

In this section, we approach the toxicity problem as a safety problem, specifically, our goal is to constrain the last token, y_k, to a region where the corresponding sentence is not toxic. We focus on the last token because LLMs generate text sequentially: the prompt is converted into a set of tokens, the initial state of the transformer, and propagated through the transformer for a finite amount of time. After this propagation, the last token is used to predict the next token which is appended at the end of the prompt to create a new initial state. The process is then repeated starting from this new initial state. Therefore, to prevent toxic outputs, we want to ensure the last token does not lead to a toxic sentence. We thus employ a classifier that transforms the initial state $\mathbf{y}^0(0)$ and the last token at the current state $\mathrm{y}_k(t)$ into the toxicity score of the sentence formed by appending the predicted next token, based on $\mathrm{y}_k(t)$, to the sequence of tokens defined by the initial state $\mathbf{y}^0(0)$.

A *toxicity classifier* is a map $c_k : (\mathcal{E}_W^n)^{k+1} \to \mathbb{R}_{\geq 0}$ that assigns a toxicity score to each sequence of tokens $(\mathrm{y}_1^0, \ldots, \mathrm{y}_k^0, \mathrm{y})$, where $\mathbf{y}^0 = (\mathrm{y}_1^0, \ldots, \mathrm{y}_k^0)$ and y correspond to the initial sequence fed to the transformer and the generated token at the k-th pass, respectively. Given that the initial sequence is constant, we may consider the map $h_0 : \mathcal{E}_W^n \to \mathbb{R}_{\geq 0}$ defined as $h_0(\mathrm{y}) = c_k\big(\mathrm{y}_1^0, \ldots, \mathrm{y}_k^0, \mathrm{y}\big)$. Henceforth, we will assume that $h_0 \in C^2(\mathcal{E}_W^n, \mathbb{R}_{\geq 0})$. Let us consider the map $h \in C^2\big((\mathcal{E}_W^n)^k\big)$ defined as:

$$\boxed{h(\mathbf{y}) = -h_0(\mathrm{y}_k) + \gamma,} \tag{8}$$

where $\gamma > 0$ is a fixed toxicity threshold. The rest of the paper is devoted to show that h is a second-order CBF for (7). Note that $(\mathrm{d}h)_{\mathbf{y}} = -(\mathrm{d}h_0)_{\mathrm{y}_k}$. Given $\alpha \in \mathcal{K}_\infty^e \cap C^1(\mathbb{R})$, we consider the map $h_1 \in C^1\big(\mathbb{R}_{\geq 0} \times (\mathcal{E}_W^n)^k\big)$ defined as:

$$h_1(t,\mathbf{y}) = -(\mathrm{d}h_0)_{\mathrm{y}_k}(f_k(t,\mathbf{y})) + \alpha(-h_0(\mathrm{y}_k) + \gamma). \tag{9}$$

This maps corresponds to the auxiliary function (2), since $(\mathrm{d}h)_{\mathbf{y}}(f(t,\mathbf{y})) = -(\mathrm{d}h_0)_{\mathrm{y}_k}(f_k(t,\mathbf{y}))$. Our approach is based on the observation that $(L_g h)(t,\mathbf{y}) = 0$ for $(t,\mathbf{y}) \in \mathbb{R}_{\geq 0} \times (\mathcal{E}_W^n)^k$ for any h induced by a toxicity classifier c_k and thus we need a higher-order CBF.

The analysis in this paper requires the classifier to be C^2. If needed, we can always approximate a less regular classifier with one of class C^2.

This approach naturally increases the computational footprint of word generation due to the safety filter. However, since the constraints are linear, the increase is relatively modest.

A. Generic surjectivity of the vector field derivative

In this section, we investigate the partial derivative of the last component of the self-attention vector field, $f_{\mathrm{sa},k}$, with respect to the first token, y_1. We conclude that, for each $(t,\mathbf{y}) \in \mathbb{R}_{\geq 0} \times (\mathcal{E}_W^n)^k$, it has maximum rank (and, thus, defines an isomorphism) for almost every choice of the matrices $(\mathbf{P}(t), \mathbf{U}(t))$. In turn this will ensure that we can provide sufficient conditions on h alone without being dependent on the state.

For each $(t,\mathbf{y}) \in \mathbb{R}_{\geq 0} \times (\mathcal{E}_W^n)^k$ and $1 \leq i \leq k$, let us introduce the following linear map:

$$\psi_i(t,\mathbf{y}) = T_{\mathrm{y}_k}\pi_W \cdot \frac{\partial \tilde{f}_{\mathrm{sa},k}}{\partial \mathrm{y}_i}(t,\mathbf{y}) : T_{\mathrm{y}_i}\mathcal{E}_W^n \to T_{\mathrm{y}_k}\mathcal{E}_W^n,$$

where we denote $\tilde{f}_{\mathrm{sa},k} = \big(\tilde{f}_{\mathrm{sa},k}^1, \ldots, \tilde{f}_{\mathrm{sa},k}^{n+1}\big)$ and

$$\frac{\partial \tilde{f}_{\mathrm{sa},k}}{\partial \mathrm{y}_i} = \left(\frac{\partial \tilde{f}_{\mathrm{sa},k}^{\mu}}{\partial y_i^{\nu}}\right)_{1\leq \mu,\nu \leq n+1}.$$

In the same vein, let us introduce the map:

$$\Xi(\mathbf{y}) : \left(\mathbb{R}^{(n+1)\times(n+1)}\right)^{2h} \to \mathbb{R}^{(n+1)\times(n+1)},$$

defined as:

$$\Xi(\mathbf{y})(\mathbf{P},\mathbf{U}) = \sum_{\eta=1}^{h} \alpha_{k1}^{\eta}(\mathbf{P},\mathbf{y})\, U_\eta\, A_\eta(\mathbf{P},\mathbf{y}),$$

where we denote:

$$A_\eta(\mathbf{P},\mathbf{y}) = \mathbb{I}_{n+1} + \left(\mathrm{y}_1 - \sum_{j=1}^{k} \alpha_{kj}^{\eta}(\mathbf{P},\mathbf{y})\,\mathrm{y}_j\right) \mathrm{y}_k^\top\, P_\eta.$$

Lemma 3.1: For each $(t,\mathbf{y}) \in \mathbb{R}_{\geq 0} \times (\mathcal{E}_W^n)^k$, we have:

$$\frac{\partial \tilde{f}_{\mathrm{sa},k}}{\partial \mathrm{y}_1}(t,\mathbf{y}) = \Xi(\mathbf{y})(\mathbf{P}(t),\mathbf{U}(t)).$$

Proof: For brevity, we omit the arguments of the maps. Firstly, a computation shows that:

$$\frac{\partial \alpha_{kj}^{\eta}}{\partial \mathrm{y}_1} = \alpha_{kj}^{\eta}\,(\delta_{j1} - \alpha_{k1}^{\eta})\,\mathrm{y}_k^\top\, P_\eta,$$

where δ_{j1} denotes the Kronecker delta. From this, we conclude:

$$\begin{aligned}\frac{\partial \tilde{f}_{\mathrm{sa},k}}{\partial \mathrm{y}_1} &= \sum_{\eta=1}^{h} U_\eta \left(\alpha_{k1}^{\eta}\,\mathbb{I}_{n+1} + \sum_{j=1}^{k} \mathrm{y}_j \frac{\partial \alpha_{kj}^{\eta}}{\partial \mathrm{y}_1}\right)\\ &= \sum_{\eta=1}^{h} \alpha_{k1}^{\eta}\, U_\eta \left(\mathbb{I}_{n+1} + \left(\mathrm{y}_1 - \sum_{j=1}^{k} \alpha_{kj}^{\eta}\,\mathrm{y}_j\right)\mathrm{y}_k^\top\, P_\eta\right).\end{aligned}$$

■

The previous lemma shows the parametric dependence of $\partial \tilde{f}_{\mathrm{sa},k}/\partial \mathrm{y}_1(t,\mathbf{y})$ (and, thus, of $\psi_1(t,\mathbf{y})$) on the matrices $(\mathbf{P}(t),\mathbf{U}(t))$. The next results ensure that $\psi_1(t,\mathbf{y})$ is surjective (and, thus, a linear isomorphism) for almost every choice of these matrices.

Proposition 3.1: Let $(t,\mathbf{y}) \in \mathbb{R}_{\geq 0} \times (\mathcal{E}_W^n)^k$. Then the map $\psi_1(t,\mathbf{y}) : T_{\mathrm{y}_1}\mathcal{E}_W^n \to T_{\mathrm{y}_k}\mathcal{E}_W^n$ is surjective for almost every $(\mathbf{P}(t),\mathbf{U}(t)) \in \left(\mathbb{R}^{(n+1)\times(n+1)}\right)^{2h}$.

Proof: The map $\tilde{\Xi}(\mathbf{y}) : \left(\mathbb{R}^{(n+1)\times(n+1)}\right)^{2h} \to \mathbb{R}$ defined as $\tilde{\Xi}(\mathbf{y})(\mathbf{P},\mathbf{U}) = \det \Xi(\mathbf{y})(\mathbf{P},\mathbf{U})$ is real analytic. Indeed, the exponential and the determinant are real analytic, and the functions $Z_k^\eta(\cdot,\mathbf{y}) : \left(\mathbb{R}^{(n+1)\times(n+1)}\right)^h \to \mathbb{R}$, $1 \leq \eta \leq h$, never vanish, whence $\tilde{\Xi}(\mathbf{y})$ is real analytic. Moreover, $\tilde{\Xi}(\mathbf{y})$ is not identically zero: for instance, let $\mathbf{P}_0 = (\mathbf{0}_{n+1},\ldots,\mathbf{0}_{n+1})$, where $\mathbf{0}_{n+1}$ is the zero matrix of size $n+1$, and $\mathbf{U}_0 = (\mathbb{I}_{n+1},\ldots,\mathbb{I}_{n+1})$. Then we get $\alpha_{k1}^{\eta}(\mathbf{P}_0,\mathbf{y}_0) = (k\sqrt{n+1})^{-1}$, $1 \leq \eta \leq h$, and, thus, $\tilde{\Xi}(\mathbf{y})(\mathbf{P}_0,\mathbf{U}_0) = \left(h/(k\sqrt{n+1})\right)^{n+1} \neq 0$. As a result, $\tilde{\Xi}(\mathbf{y})^{-1}(0)$ has zero measure (cf., for example, [23] for an elementary proof). From Lemma 3.1, we conclude that $\operatorname{rank}\left(\partial \tilde{f}_{\mathrm{sa},k}/\partial \mathrm{y}_1\right)(t,\mathbf{y}) = n+1$ for almost every $(\mathbf{P}(t),\mathbf{U}(t)) \in \left(\mathbb{R}^{(n+1)\times(n+1)}\right)^{2h}$. Subsequently, we have that $\operatorname{rank}\psi_1(t,\mathbf{y}) = n = \dim T_{\mathrm{y}_k}\mathcal{E}_W^n$, where we used that $\operatorname{rank} T_{\mathrm{y}_k}\pi_W = n$. Thus, $\psi_1(t,\mathbf{y})$ is surjective for almost every $(\mathbf{P}(t),\mathbf{U}(t)) \in \left(\mathbb{R}^{(n+1)\times(n+1)}\right)^{2h}$.

■

B. Dynamics of the control barrier function

In this section we present the main result of the paper, Corollary 3.1, showing that the function $h \in C^2\left((\mathcal{E}_W^n)^k\right)$ defined in (8) is a second-order CBF for the dynamics of the transformer (7). We can thus leverage this CBF to prevent toxic outputs. Corollary 3.1 is a direct consequence of Theorem 3.1.

We start with the following technical lemma.

Lemma 3.2: For each $(t,\mathbf{y},\mathrm{u}) \in \mathbb{R}_{\geq 0} \times (\mathcal{E}_W^n)^k \times \mathbb{R}^{n+1}$, we have $\nabla^{\mathbf{g}}_{g(t,\mathrm{y}_1)\cdot \mathrm{u}} f_{\mathrm{sa},k}(t,\mathbf{y}) = \psi_1(t,\mathbf{y}) \cdot T_{\mathrm{y}_1}\pi_W \cdot \mathrm{u}$.

Proof: Let $\Gamma_{\mu\nu,l}^{ij,\sigma} \in C^\infty\left((\mathcal{E}_W^n)^k\right)$, $1 \leq i,j,l \leq k$, $1 \leq \mu,\nu,\sigma \leq n+1$, be the Christoffel symbols of $\mathbf{g}$, and denote $\partial_\mu = (0,\ldots,\underbrace{1}_{\mu\text{-th position}},\ldots,0)$, $1 \leq \mu \leq n+1$, which allows for writing $T_{\mathrm{y}_1}\pi_W \cdot \mathrm{u} = \sum_{\mu=1}^{n+1} \xi_1^\mu(\mathrm{y}_1,\mathrm{u})\,\partial_\mu \in T_{\mathrm{y}_1}\mathcal{E}_W^n$, for some $\xi_1^\mu(\mathrm{y}_1,\mathrm{u}) \in \mathbb{R}$, $1 \leq \mu \leq n+1$. Therefore, a computation using that $\Gamma_{\mu\nu,l}^{ij,\sigma} \neq 0$ only when $i=j=l$ (as the product metric is block diagonal and, thus, it does not mix different copies of $\mathcal{E}_W^n$), and $T_{\mathrm{y}_k}\pi_W$ does not depend on y_1, leads to $\nabla^{\mathbf{g}}_{g\cdot \mathrm{u}} f_{\mathrm{sa},k} = \psi_1 \cdot T_{\mathrm{y}_1}\pi_W \cdot \mathrm{u}$. ■

Now let us compute the Lie derivative of the auxiliary function h_1 with respect to the control field.

Proposition 3.2: Consider the function introduced in (9), $h_1 \in C^1\left(\mathbb{R}_{\geq 0} \times (\mathcal{E}_W^n)^k\right)$. Then, for each $(t,\mathbf{y},\mathrm{u}) \in \mathbb{R}_{\geq 0} \times (\mathcal{E}_W^n)^k \times \mathbb{R}^{n+1}$, we have:

$$(L_g h_1)(t,\mathbf{y}) \cdot \mathrm{u} = -(\mathrm{d}h_0)_{\mathrm{y}_k}\left(\psi_1(t,\mathbf{y}) \cdot T_{\mathrm{y}_1}\pi_W \cdot \mathrm{u}\right).$$

where $\alpha' : \mathbb{R}_{\geq 0} \to \mathbb{R}_{>0}$ denotes the derivative of α.

Proof: Note that $\nabla^{\mathbf{g}}_{g(t,\mathrm{y}_1)\cdot \mathrm{u}} f_{\mathrm{ff}}(t,\mathrm{y}_k) = 0$, since $g(t,\mathrm{y}_1)\cdot \mathrm{u}$ does not depend on y_k. Similarly, we have:

$$(\mathrm{d}h)_{\mathbf{y}}(g(t,\mathrm{y}_1)\cdot \mathrm{u}) = (\mathrm{d}h_0)_{\mathrm{y}_k}(0) = 0.$$

From this and (7), we obtain:

$$(L_g h_1)(t,\mathbf{y}) \cdot \mathrm{u} = -(\mathrm{d}h_0)_{\mathrm{y}_k}\left(\psi_1(t,\mathbf{y}) \cdot T_{\mathrm{y}_1}\pi_W \cdot \mathrm{u}\right),$$

where we used Lemma 3.2 for the last equality. ■

We are ready to give a sufficient condition for h to be a second-order CBF for the controlled dynamics of the transformer.

Theorem 3.1: Let $h \in C^2\left((\mathcal{E}_W^n)^k\right)$ be as in (8), h_1 be as in (9), and $\mathcal{D} \subset (\mathcal{E}_W^n)^k$ open be such that:

$$\mathcal{S} = h^{-1}(\mathbb{R}_{\geq 0}) \cap h_1^{-1}(\mathbb{R}_{\geq 0})^\circ \subset \mathcal{D}.$$

If the following assumptions hold true:

1) $(\mathrm{d}h_0)_{\mathrm{y}_k} \neq 0$ for each $\mathbf{y} \in \mathcal{D} - \operatorname{int}\mathcal{S}$, where $\operatorname{int}\mathcal{S}$ denotes the interior of $\mathcal{S}$.
2) $\psi_1(t,\mathbf{y}) : T_{\mathrm{y}_1}\mathcal{E}_W^n \to T_{\mathrm{y}_k}\mathcal{E}_W^n$ is surjective for each $(t,\mathbf{y}) \in \mathbb{R}_{\geq 0} \times \mathcal{D}$.

Then h is a second-order CBF for the control dynamics of the transformer (7).

Proof: From Definition 1.3, h is a second-order CBF for controlled dynamics of the transformer if $(L_g h)(t,\mathbf{y}) = 0$ and (3) holds for each $(t,\mathbf{y}) \in \mathbb{R}_{\geq 0} \cap \mathcal{D}$. The first condition is straightforward; namely:

$$(L_g h)(t,\mathbf{y}) = (\mathrm{d}h)_{\mathbf{y}}(g(t,\mathrm{y}_1)) = (\mathrm{d}h_0)_{\mathrm{y}_k}(0) = 0.$$

For the second condition, we choose $\alpha_1 \in \mathcal{K}_\infty^e$ as $\alpha_1(r) = (|L_m|/h_m)\, r$, where we denote $L_m = \inf_{(t,\mathbf{y})\in\mathbb{R}_{\geq 0}\times\mathcal{H}}(L_f h_1)(t,\mathbf{y})$, as well as $h_m = \inf_{(t,\mathbf{y})\in\mathbb{R}_{\geq 0}\times\mathcal{H}} h_1(t,\mathbf{y})$. Note that the following set:

$$\mathcal{H} = (\mathrm{d}h_0 \circ \pi_k)^{-1}(0) = \{\mathbf{y} \in \mathcal{D} \mid (\mathrm{d}h_0)_{\mathrm{y}_k} = 0_{\mathrm{y}_k}\},$$

where $\pi_k : \mathcal{D} \to \mathcal{E}_W^n$ is the projection onto the k-th copy, is closed (in $\mathcal{D}$ and, thus, in $(\mathcal{E}_W^n)^k$ as $\mathcal{D}$ is open). Moreover, assumption 1) implies that $\mathcal{H} \subset \operatorname{int}\mathcal{S}$. Hence, the closeness of $\mathcal{H}$, together with the fact that $h_1(t,\mathbf{y}) > 0$ for each $(t,\mathbf{y}) \in \mathbb{R}_{\geq 0} \times \operatorname{int}\mathcal{S}$, ensure that $h_m > 0$. Given $(t,\mathbf{y}) \in \mathbb{R}_{\geq 0} \times \mathcal{D}$, we distinguish two cases:

1) $\mathbf{y} \in \mathcal{D} - \mathcal{H}$. Consider the following map:

$$\mathbb{R}^{n+1} \to \mathbb{R}, \quad \mathrm{u} \mapsto (L_g h_1)(t,\mathbf{y}) \cdot \mathrm{u},$$

From Proposition 3.2, as well as the facts that $(\mathrm{d}h_0)_{\mathrm{y}_k} \neq 0$ and $\psi_1(t,\mathbf{y})$ is surjective, we conclude that the previous map is surjective and, thus, (3) holds.

2) $\mathbf{y} \in \mathcal{H}$. By construction, $h_1(t,\mathbf{y}) \geq h_m$, whence:

$$\alpha_1(h_1(t,\mathbf{y})) \geq |L_m| \geq -(L_f h_1)(t,\mathbf{y}),$$

and (3) holds.

■

By gathering Proposition 3.1 and Theorem 3.1, we arrive at the main result of the paper.

Corollary 3.1: Let $\mathcal{D} \subset (\mathcal{E}_W^n)^k$ open be such that:

$$\mathcal{S} = h^{-1}(\mathbb{R}_{\geq 0}) \cap h_1^{-1}(\mathbb{R}_{\geq 0})^\circ \subset \mathcal{D},$$

with h and h_1 as in (8) and (9), respectively. If $(\mathrm{d}h_0)_{\mathrm{y}_k} \neq 0$ for each $\mathbf{y} \in \mathcal{D} - \operatorname{int}\mathcal{S}$, then h is a second-order CBF for the control dynamics of the transformer (6) for almost every $(\mathbf{P}(t), \mathbf{U}(t)) \in \left(\mathbb{R}^{(n+1)\times(n+1)}\right)^{2h}$, $t \in \mathbb{R}_{\geq 0}$.

Note that $\mathcal{S} \subset h^{-1}(\mathbb{R}_{\geq 0}) \subset \{\mathbf{y} \in (\mathcal{E}_W^n)^k \mid h_0(\mathrm{y}_k) \leq \gamma\}$. This guarantees that the toxicity of the output remains below the threshold γ provided the initial condition lies in the safe set $\mathcal{S}$ and the control inputs are chosen to be in $K_{\mathrm{CBFH}}(t,\mathbf{y})$ for each $(t,\mathbf{y}) \in \mathbb{R}_{\geq 0} \times \mathcal{S}$.

The strength of this result lies in the weak assumptions placed on the function h_0, which are a consequence of the new transformer architecture proposed in this paper. We simply ask for the absence of points (interpreted as sentences) where the linearization of h_0 is identically zero and thus where h would fail to describe how toxicity changes in a small neighborhood of that sentence.

IV. Conclusion

In this paper, the problem of toxicity in LLM text generation has been addressed. Firstly, a continuous-time model explicitly describing the feedforward layers of transformers was derived. We then proposed a new architecture in which the first token can be steered and is regarded as a controlled token. This enabled us to show that any C^2 function satisfying a mild non-degeneracy condition can be used to construct a CBF for almost every choice of the parameters of the transformer (key, query and value matrices).

References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[2] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever *et al.*, "Improving language understanding by generative pre-training," *OpenAI*, 2018.

[3] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[4] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[5] D. Guo, D. Yang, H. Zhang, J. Song, R. Zhang, R. Xu, Q. Zhu, S. Ma, P. Wang, X. Bi *et al.*, "Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning," *arXiv preprint arXiv:2501.12948*, 2025.

[6] M. Sandler, H. Choung, A. Ross, and P. David, "A linguistic comparison between human and chatgpt-generated conversations," in *International Conference on Pattern Recognition and Artificial Intelligence*. Springer, 2024, pp. 366–380.

[7] S. Gehman, S. Gururangan, M. Sap, Y. Choi, and N. A. Smith, "Realtoxicityprompts: Evaluating neural toxic degeneration in language models," *arXiv preprint arXiv:2009.11462*, 2020.

[8] H. Luo and L. Specia, "From understanding to utilization: A survey on explainability for large language models," *arXiv preprint arXiv:2401.12874*, 2024.

[9] B.-D. Oh and W. Schuler, "Token-wise decomposition of autoregressive language model hidden states for analyzing model predictions," *arXiv preprint arXiv:2305.10614*, 2023.

[10] A. Madsen, S. Chandar, and S. Reddy, "Are self-explanations from large language models faithful?" *arXiv preprint arXiv:2401.07927*, 2024.

[11] X. Suau, P. Delobelle, K. Metcalf, A. Joulin, N. Apostoloff, L. Zappella, and P. Rodríguez, "Whispering experts: Neural interventions for toxicity mitigation in language models," *arXiv preprint arXiv:2407.12824*, 2024.

[12] T. S. Luong, T.-T. Le, L. N. Van, and T. H. Nguyen, "Realistic evaluation of toxicity in large language models," *arXiv preprint arXiv:2405.10659*, 2024.

[13] P. Chao, A. Robey, E. Dobriban, H. Hassani, G. J. Pappas, and E. Wong, "Jailbreaking black box large language models in twenty queries," *arXiv preprint arXiv:2310.08419*, 2023.

[14] N. Karagodin, Y. Polyanskiy, and P. Rigollet, "Clustering in causal attention masking," *arXiv preprint arXiv:2411.04990*, 2024.

[15] Á. Rodríguez Abella, J. P. Silvestre, and P. Tabuada, "The asymptotic behavior of attention in transformers," *arXiv preprint arXiv:2412.02682*, 2024.

[16] E. Cheng, M. Baroni, and C. A. Alonso, "Linearly controlled language generation with performative guarantees," *arXiv preprint arXiv:2405.15454*, 2024.

[17] A. D. Ames, X. Xu, J. W. Grizzle, and P. Tabuada, "Control barrier function based quadratic programs for safety critical systems," *IEEE Transactions on Automatic Control*, vol. 62, no. 8, pp. 3861–3876, 2016.

[18] Y. Miyaoka and M. Inoue, "Cbf-llm: Safe control for llm alignment," *arXiv preprint arXiv:2408.15625*, 2024.

[19] Q. Nguyen and K. Sreenath, "Exponential control barrier functions for enforcing high relative-degree safety-critical constraints," in *2016 American Control Conference (ACC)*. IEEE, 2016, pp. 322–328.

[20] W. Xiao and C. Belta, "High-order control barrier functions," *IEEE Transactions on Automatic Control*, vol. 67, no. 7, pp. 3655–3662, 2021.

[21] C. Yun, S. Bhojanapalli, A. S. Rawat, S. J. Reddi, and S. Kumar, "Are transformers universal approximators of sequence-to-sequence functions?" *arXiv preprint arXiv:1912.10077*, 2019.

[22] J. Zhao, K. Chen, X. Yuan, and W. Zhang, "Prefix guidance: A steering wheel for large language models to defend against jailbreak attacks," *arXiv preprint arXiv:2408.08924*, 2024.

[23] B. Mityagin, "The zero set of a real analytic function," 2015. [Online]. Available: https://arxiv.org/abs/1512.07276