Unpredictability of AI Alignment Is Not Always Bad for AI Safety

Petr Spelda¹ and Vit Stritecky, Department of Security Studies, Charles University

Abstract

Robustness of AI alignment is one of the safety issues of large language models. Can we predict how many mistakes will a model make when responding to a restricted request? We show that when access to the model is limited to in-context learning, the number of mistakes can be proved inapproximable, which can lead to unpredictability of alignment of the model. Against intuition, this is not entirely bad news for AI safety. Attackers might not be able to easily misuse in-context learning to break alignment of the model in a predictable manner because the mistake bounds of safe responses, which were used for alignment, can be proved inapproximable. This inapproximability can hide the safe responses from attackers and make alignment of the model unpredictable. If it were possible to keep the safe responses from attackers, responsible users would benefit from testing and repairing of the model's alignment despite its possible unpredictability. We also discuss challenges involved in ensuring democratic AI alignment with limited access to safe responses, which helps us to make alignment of the model unpredictable for attackers.

1. Introduction

Al alignment aims to reduce risks involved in the development and use of capable Al systems. Current risk reduction techniques trace back to reinforcement learning experiments with human feedback. Human preferences on possible agent trajectories in a simulated environment were used to learn a reward function optimized by the agent (Christiano et al. 2017). This kind of reward modelling later turned into a major research direction in Al alignment (Leike et al. 2018). Human intentions expressed as preferences on possible actions of a reinforcement learning agent in an environment were used to learn the reward model optimized by the agent. The most widespread application of reward modelling so far was found in training LLMs (large language models) to follow user instructions to perform various tasks (Bai et al. 2022; Ouyang et al. 2022). Together with supervised safety fine-tuning, reinforcement learning from human feedback (RLHF) based on reward modelling attempts to

¹ Correspondence: <u>petr.spelda@fsv.cuni.cz</u>, Date: April 15, 2025.

ensure that LLMs will produce safe request completions and refuse to complete unsafe tasks in alignment with human preferences (e.g., see Touvron et al. 2023). While popular, RLHF has alternatives such as Direct Preference Optimization (Rafailov et al. 2023) that does not involve the actor-critic alignment architecture.

Concerns over the availability of human preferences and the ability of humans to provide preferences effectively led to the development of reinforcement learning from AI feedback (Bai et al. 2022). Here, principles characterizing safe LLM responses in a general way are used to guide an LLM in revising unsafe responses, and based on the same principles, the LLM is then used to decide which output from a pair produced in response to a prompt is better aligned with the principles (ibid.). This AI feedback on responses to a set of prompts mixed with human preferences is then used to train a reward model for reinforcement learning of the final LLM aligned with 'AI feedback' (ibid.). Relatedly, Deliberative Alignment (Guan et al. 2024) aims to train models capable of chain-of-thought reasoning to select a relevant safety policy for the user request, inform the generated chain-of-thought by it, and produce a safe completion.

Other alignment methods were proposed to address the limited human ability to provide feedback on complex tasks. 'AI safety via debate' uses a zero-sum game between two AI agents producing a sequence of prompt responses at the end of which the user decides which agent provided a more useful (aligned) solution to the problem (Irving et al. 2018). The 'Iterated Amplification' method builds up complex tasks' solutions from aligned solutions to easier subtasks provided by users and can, in principle, maintain alignment of the AI agent without reward modelling (Christiano et al. 2018). A related, weak-to-strong principle was tested by Burns et al. (2024) who successfully experimented with using outputs of a weaker model to elicit strong performance from a more capable model.

An LLM's alignment based on methods such as RLHF can fail or be undone. A common safety issue are failures to refuse the completion of restricted, harmful tasks caused by jailbreaking attacks (Wei et al. 2023). Jailbreaking prompts elicit conflicts between models' capabilities after pretraining and safety training and allow users to perform restricted tasks by manipulating prompts via prefix injection or refusal suppression (ibid.). This led to a realization that adversarial alignment is hard to achieve. Zou et al. (2023) developed a method for automating the search for 'adversarial suffixes' which, when added to a restricted prompt, force the model to produce a harmful response. Both manual and automated jailbreaking

methods are known to identify prompts that transfer across models, allowing attackers to elicit harmful responses across a range of models. Qi et al. (2024) showed that fine-tuning, a process of adapting models to new tasks, can turn safe LLMs with just a few harmful examples to poorly restricted models. Similar effects were observed even if the fine-tuning prompt-response pairs were benign (ibid.). Using tools developed for mechanistic interpretability, Jain et al. (2024) showed that the effects of fine-tuning on the base model are mild and that restricted capabilities can be brought back in a sample efficient manner.

To remedy this situation, red-teaming efforts seeking to uncover harmful capabilities of LLMs before their deployment were augmented with adversarial prompt generators tailormade to particular LLMs (Casper et al. 2023). Despite the existence of semi-automatic predeployment safety-procedures, it is impossible to ensure that all vulnerabilities allowing the model to produce harmful responses are identified before the LLM is put to use. In this situation, it is reasonable to ask:

Question 1 Is it possible to predict alignment of the model?

We provide an answer to this question using the mistake bound approach developed for online learning (Littlestone 1988), in which we are interested in bounding the number of mistakes M that a learner A makes on sequences $S \in S$ of the length $T \in \mathbb{N}$.

Fact 1 The mistake bound can be used to characterize the possibility that an auto-regressive model AR, such as an LLM based on the decoder-only Transformer architecture, e.g., a model from the family of generative pretrained transformers beginning with Radford et al. (2018), produces a sequence S representing its safest response to the most restricted requests.

Question 1 is significant because if were able to calculate or approximate mistake bounds of the model's responses to safety-sensitive requests even without access to the safe responses, the alignment of LLMs with human preferences could become predictable.

We will show that there are theoretical reasons suggesting that predictability of alignment by mistake bounds of safe responses without having full access to the responses is unlikely. Our main tool for deriving a negative answer to Question 1 is the recently proved inapproximability of Littlestone's dimension (Manurangsi and Rubinstein 2017; Manurangsi 2023). Littlestone's dimension is a fundamental tool of learning theory developed for bounding the number of mistakes made by learners on sequences (Littlestone 1988) in the so-called

realizable setting. Here, the realizable setting means that we assume the model can generate the safe response to a restricted request because it was trained to do so, which is the goal of Al alignment. Once we use Littlestone's dimension for bounding the number of mistakes that an auto-regressive model can make while responding to a user request, predictability of the model's alignment can become difficult. We build a theoretical explanation of this unpredictability as follows:

Proposition 1 It is difficult to predict bad responses of the model because we cannot approximate the mistake bounds of its aligned responses in polynomial time (the approximation task is considered intractable).

In this paper, we build a theoretical argument for Proposition 1, offer an empirical demonstration designed to support the theoretical argument, and unpack surprisingly positive effects of Proposition 1 on AI safety. The main positive impact is shown in Section 3 and comments on its implications for practice are provided in Section 4. Section 2 builds the necessary connection between autoregressive sequence generation and online learning.

Contribution: We show that AI alignment can become unpredictable because the mistake bounds of an LLM's safe responses can be proved inapproximable if the responses are not fully accessible. This unpredictability can lead to what we call safety symmetry, a situation in which alignment of the model is difficult to guarantee as well as exploit with respect to the safe responses to restricted requests. Safety symmetry is characterized by Theorem 1 in Section 3.1, and in Section 3.2 we seek to provide indicative empirical support for the unpredictability. Circumstances which lead to unpredictability of the model's alignment can easily occur in practice. For this reason, safety symmetry should be understood as a factor that could justify further investments in responsible LLM development. Potentially unpredictable yet improved alignment of the model will likely disadvantage attackers more than legitimate users who, while facing the same unpredictability, will use a safer model.

Our unpredictability of AI alignment has also an epistemological upshot. The fact that it is possible to prove mistakes bounds of safe responses inapproximable in the situation where the model should be able to generate safe responses could be interpreted as strengthening of Gerhard Schurz's optimality justification of epistemic methods (Schurz 2022; 2024). The optimality justification of induction requires an algorithm capable of learning from prediction

errors of available methods such that it minimizes the gap between its errors and errors of the best method in hindsight (Schurz 2019, Sect. 6.6). This gap is called regret, and some metainductive algorithms can be shown to be no-regret by having unimprovable worst-case regret bounds and per-round regret that vanishes as the number of rounds approaches infinity (ibid.). No-regret learning is necessary for the optimality justification. But since it can be proved that in certain situations agents cannot predict each other due to inapproximability of their mistake bounds of sequences they are predicting, no-regret learning might become even fundamental. This could strengthen optimality justifications because the inapproximability of agents' mistake bounds can lead to unpredictability of agents' interactions, in which case the only epistemically sound course of action is to learn from the interactions in the optimal way, as suggested by the optimality justification because users can encounter difficulties predicting alignment of LLM responses with safety requirements, and we focus on analyzing and providing evidence on a possible cause of this unpredictability, which could strengthen the first type of Schurz's optimality justifications (the optimality of meta-induction, 2022, Sect. 3).

2. Inapproximability of AI Alignment

Proposition 1 can be interpreted as unpredictability of AI alignment. There are results such as Wolf et al. (2024) that challenge the robustness of alignment by arguing that once an LLM acquires harmful capabilities, it impossible to remove them from the model. Our approach is different. It shows difficulties in deriving mistake bounds for aligned capabilities of the model that would provide us with information on how safe LLM responses turn into unsafe ones. It was also shown by Glukhov et al. (2024) that any attempt to enforce alignment via censoring LLM outputs is formally undecidable. Undecidability of censorship of LLM outputs derived from Rice's Theorem is an important limiting result that is, however, difficult to apply to real-world models. Our approach stays as close as possible to the real-world models that are being aligned with human safety preferences using methods described in previous section. In fact, our approach works with a customizable token set \mathcal{X} and the in-context learning capability of LLMs characterized as online learning.

It should be also noted that computational intractability of Littlestone's dimension motivated a theoretical approach to sequential learning where the mistake bound does not depend on computing or approximating Littlestone's dimension (Assos et al. 2023;

Kozachinskiy and Steifer 2024). The approach requires an oracle that at each step gives the learner a hypothesis consistent with the sequence generated so far (ibid.). Since implementing the oracle in practice for non-trivial sequences is not easy, we do not consider consistency oracles a practical solution allowing us to sidestep the hardness of approximating Littlestone's dimension for deriving mistake bounds.

2.1 AI Alignment as Online Learning

Let C be a finite class of concepts defined on a set of tokens \mathcal{X} . Each concept $C \in C$ is a possible response that the AR model can generate in response to a user request. The response consists of tokens from the set \mathcal{X} . Both the set of tokens \mathcal{X} and the concept class C are known. Therefore, the situation can be described as explicit (Manurangsi and Rubinstein 2017; Manurangsi 2023) because there is a (0,1)-matrix $|\mathcal{X}| \times |C|$. An entry $(x, C), x \in \mathcal{X}$ and $C \in$ C, equals 1 if C(x) = 1, that is, if the token x is part of the response C, and C(x) = 0otherwise. With respect to restricted requests, some concepts are safer than other. Alignment of the AR model seeks to ensure that restricted requests are answered with responses that are as close as possible to concepts from the subset $C_{safe} \subset C$. The concepts in C_{safe} correspond to responses created by safety techniques after pretraining of the AR model or to safe responses learned during pretraining. Formally,

Definition 1 Let C be a class of auto-regressive concepts and $T \in \mathbb{N}$ the length of a sequence $Z = (z_1, C_r^*(z_1), \dots, z_T, C_r^*(z_T))$. The concept C_r^* represents the safest possible response to a restricted request $r \in R$ that the AR model can generate. Each token z_1, \dots, z_t is part of the actual response $Z \in Z$ of the AR model generated through steps $t \leq T$. Tokens z_1, \dots, z_t are selected from the set X.

Remark 1 We observe the following notational difference. If $C(z) \in \{0,1\}$ is used, it represents a non-indexed membership query, i.e., is the token $z \in \mathcal{X}$ part of the response corresponding to the concept C? On the other hand, $C(z_i) \in \{0,1\}$ represents a query asking whether the *i*th token of the response Z corresponds to *i*-th token of the response represented by the concept C.

We are interested in bounding the number of mistakes M that an in-context learning (ICL) algorithm A_{ICL} , which is supported by the AR model, makes with respect to the aligned concept C_r^* while auto-regressively producing a response Z to the restricted request r.

Definition 2 of A_{ICL} . At each step $t \in [T]$, the algorithm takes the so far generated response Z_{t-1} and a candidate token z_t and predicts whether the token is part of the response or not, i.e., $A_{ICL}(Z_{t-1}, z_t) \in \{0, 1\}$.

In terms of, for example, Definition 1 by Hasrati and Ben-David (2023), our Definition 2 makes A_{ICL} into an online learning algorithm. $Z_{C_{safe}} \subset Z$ is a set of responses for restricted requests that correspond to safe concepts, that is, $Z_{C_{safe}} = \{Z \in Z : L_Z(C_{safe}) = 0\}$. The loss of a concept *C* on a response *Z* for the restricted request *r* with respect to C_r^* is defined as $L_Z(C) = \sum_{t=1}^T \mathbb{1} [C(z_t) \neq C_r^*(z_t)]$.

Definition 3 of the A_{ICL} 's mistake bound. Let the number of mistakes that an in-context learning algorithm A_{ICL} makes while responding to a restricted request r be defined as $M_{A_{ICL}}(Z) = \sum_{t=1}^{T} \mathbb{1}[A_{ICL}(Z_{t-1}, z_t) \neq C_r^*(z_t)]$. The safety mistake bound of A_{ICL} for the concept class C is defined as $M_{A_{ICL}}(C_{safe}) = \sup_{Z \in \mathbb{Z}_{C_{safe}}} M_{A_{ICL}}(Z)$.

Using the definition of the online learner's mistake bound by Hasrati and Ben-David (2023) (their Definition 2) as a starting point, our $M_{A_{ICL}}(\mathcal{C}_{safe})$ bounds the error of A_{ICL} on safety sensitive responses to restricted requests for which the base class \mathcal{C} contains safe concepts \mathcal{C}_{safe} . The safety subclass \mathcal{C}_{safe} is learnable by A_{ICL} if $M_{A_{ICL}}(\mathcal{C}_{safe}) < \infty$.

Fact 2 The safety mistake bound of A_{ICL} , $M_{A_{ICL}}(\mathcal{C}_{safe})$, is lower bounded by the value of Littlestone's dimension of the safe concept class \mathcal{C}_{safe} , that is, $M_{A_{ICL}}(\mathcal{C}_{safe}) \ge \text{Ldim}(\mathcal{C}_{safe})$.

Littlestone's dimension of the concept class C_{safe} is equal to $d \in \mathbb{N}$, the depth of the largest, complete binary (mistake) tree whose root-to-leaf paths can be constructed from concepts in C_{safe} , see Shalev-Shwartz and Ben-David (2014), Definition 21.4 and 21.5.

Fact 2 results from applying Littlestone's (1988) Theorem 3 to the safe concept class C_{safe} . Fact 2 also means that the minimum mistake bound achievable by A_{ICL} on C_{safe} is equal to $Ldim(C_{safe})$ if A_{ICL} implements Littlestone's Standard Optimal Algorithm (SOA) given as follows (using the algorithm given by Shalev-Shwartz and Ben-David, 2014, p. 250, as a template):

Standard Optimal Algorithm for ICL (SOA – ICL) in the realizable setting			
1: input restricted request <i>r</i>			
2: initialize $V = C_{safe}$			
3: for $t = 1,, T$			
4: generate token z_t			
5: let $u \in \{0,1\}, V_t^u = \{C \in V_t : C(z_t) = u\}$			
6: $p_t = \operatorname{argmax}_u \operatorname{Ldim}(V_t^u) \#$ if a tie occurs, $p_t = 1$			
7: $V_{t+1} = \{C \in V_t : C(z_t) = C_r^*(z_t)\}$			
8: end for			
9: return V			

Using SOA – ICL, we note several things about AI alignment in the online learning framework. First, if C_{safe} contains the concept C_r^* representing an aligned response to the restricted request r, then, from the perspective of learning theory, the setting is realizable. In the realizable setting, we know that the desired sequence, in our case a safe response Z_r , can be generated by A_{ICL} if it learns C_{safe} . After executing SOA – ICL on a restricted request r, V, initially corresponding to C_{safe} , becomes a singleton because the algorithm isolated C_r^* and $A_{SOA-ICL}$ responded with Z_r .

Remark 2 For any given AR model, we can have a set of classes C_{safe} indexed by $i = 1, ..., K \in \mathbb{N}$ for which each i, determining the length T of the response Z_r , equals to d, the depth of the mistake tree of the respective C_{safe} .

It would be tempting to also say that using $Ldim(\mathcal{C}_{safe})$ as the safety mistake bound of $A_{SOA-ICL}$ we could predict the alignment of the AR model because we know the number of mistakes it will possibly make while attempting to respond to unsafe requests. Unsafe requests attempt to break the guarantee that restricted requests are answered only with responses corresponding to the concepts from the class \mathcal{C}_{safe} .

As a result, predictability of alignment can depend on computability or approximability of Littlestone's dimension of C_{safe} because $M_{A_{SOA-ICL}}(C_{safe}) = \text{Ldim}(C_{safe})$. Frances and Litman (1998), using earlier work of Papadimitriou and Yannakakis (1996), showed that there is no optimal mistake bound algorithm in Littlestone's sense with polynomial runtime. This rules out exact computability of $M_{A_{SOA-ICL}}(C_{safe})$. Manurangsi and Rubinstein (2017), see also Manurangsi (2023), in turn showed that Littlestone's dimension of a concept class cannot be approximated to within a constant factor in polynomial time. This rules out approximability of $M_{A_{SOA-ICL}}(C_{safe})$. The inapproximability of $Ldim(C_{safe})$ substantiates our Proposition 1 in the explicit setting where C_{safe} is given as an $|\mathcal{X}| \times |C_{safe}|$ matrix (Manurangsi and Rubinstein 2017). The inapproximability result by Manurangsi and Rubinstein (2017) relies on the randomized Exponential Time Hypothesis for k-SAT, $k \ge 3$, problems by Impagliazzo and Paturi (2001).

Proof of Proposition 1 Since SOA – ICL computes or approximates $\operatorname{Ldim}(V_t^u)$ in each autoregressive step, its runtime is not polynomial. For AR models like general-purpose LLMs working with a large token set \mathcal{X} , long contexts, and many safety guardrails increasing the size of \mathcal{C}_{safe} , $A_{SOA-ICL}$ will not be able to find the concept corresponding to C_r^{\star} in polynomial time. This in turn means that we cannot predict that the response Z_r to a restricted request r will be safe before it is actually generated and inspected. The reason for this lies in the fact that we can know the mistake bound of \mathcal{C}_{safe_i} (the index i determines the response length) only after we construct the mistake tree out of the $|\mathcal{X}| \times |\mathcal{C}_{safe_i}| \operatorname{matrix}(\mathcal{C}_{safe_i}(z) = 1$ if the token $z \in \mathcal{X}$ is part of the response) and measure its depth. Since the depth d of the tree equals to Ldim (\mathcal{C}_{safe_i}) , we then know the mistake bound $M_{A_{SOA-ICL}}(\mathcal{C}_{safe_i}) = \operatorname{Ldim}(\mathcal{C}_{safe_i})$. But as Ldim (\mathcal{C}_{safe_i}) cannot be computed nor approximated in polynomial time, we cannot construct the mistake tree in polynomial time, and this makes AI alignment in terms of concrete mistake bounds of safe responses unpredictable.

3. Safety Benefits of Unpredictability of AI Alignment

Although not having even approximate safety bounds for AI alignment could become yet another AI safety problem (Hendrycks et al. 2022), here we can show that, rather counterintuitively, the effect might be opposite. To do so, we review fundamentals of Manurangsi and Rubinstein's (2017) reduction from Label Cover to Littlestone's dimension, the technical core of the inapproximability of Littlestone's dimension, and show how it can make breaking of the model's alignment in our online learning framework unpredictable.

In other words, the inapproximability of Littlestone's dimension of the class of safe concepts cuts both ways.

Fact 3 As it can be impossible to predict unsafe responses of the model because the mistake bound of safe responses from C_{safe} can be proved inapproximable, it can be hard for attackers to make the AR model to produce concrete unsafe responses because attackers cannot approximate $Ldim(C_{safe})$ nor Ldim(C) for that matter.

3.1 Breaking of AI Alignment Can Be Unpredictable

To show that in our online learning framework targeted breaking of AI alignment is not easier than achieving alignment in terms of concrete mistake bounds, we begin by defining the mistake tree shattered by a safe concept class.

Definition 4 of C_{safe} -shattered tree (using Definition 5 by Hasrati and Ben-David, 2023) $\tau = (z_1, ..., z_{2^{d-1}}) \in \mathcal{X}^{2^{d-1}}$ is a mistake tree of the depth d shattered by C_{safe} if there is a concept $C \in C_{safe}$ such that $C(z_{i_j}) = C_r^*(z_j)$ for every $(C_r^*(z_1), ..., C_r^*(z_d)) \in \{0,1\}^d$ and a restricted request $r \in R$, where $j \in [d]$ and $i_j = 2^{j-1} + \sum_{k=1}^{j-1} C_r^*(z_k) 2^{j-1-k}$. The shattered tree τ belongs to the set $T_{C_{safe}}^d$ of all trees shattered by C_{safe} .

Using Hasrati and Ben-David's (2023) Remark 6, we note that each *d*-deep binary tree $(z_1, ..., z_{2^d-1}) \in T^d_{\mathcal{C}_{safe}}$ consists of nodes such that the *i*-th node is labeled with the token z_i . Each safe response $C^*_r(z_1), ..., C^*_r(z_d)$ to a restricted request $r \in R$ is a root-to-leaf path that begins at the root node $i_1 = 1$ and continues in the following way – at the node i_j , if $C^*_r(z_j) = 1$, the path continues to the i_j 's right child and to the left one otherwise.

The process of generating concepts corresponding to responses of the *AR* model can be formalized as the Label Cover problem, which is one the major methods for obtaining inapproximability results in computational complexity theory (Arora and Lund 1996). Manurangsi and Rubinstein (2017) showed how to generate a concept class from a bi-regular Label Cover instance \mathcal{L} . We will briefly review the main mechanism of this reduction and use it to show that it is computationally hard to recover the set of shattered trees T^d_{Csafe} from the (0,1)-matrix $|\mathcal{X}| \times |\mathcal{C}_{safe}|$. As a result, attackers cannot approximate $\text{Ldim}(\mathcal{C}_{safe})$ and break the alignment of the *AR* model in a predictable manner because they cannot recover the underlying \mathcal{C}_{safe} concept class that would tell them precisely what to attack. The Label Cover problem deals with labeling nodes of a graph such that edges connecting labeled nodes satisfy constraints from some set. Formally, **Definition 5 of Label Cover instance** (using Definition 9 by Manurangsi and Rubinstein, 2017) Let \mathcal{L} be a Label Cover instance that is created from a bi-partite graph G = (U, V, E), where Uand V are sets of nodes and E is a set of edges. Nodes in $U \cup V$ are labelled with elements from some alphabet Σ . A partial assignment that labels only a subset $R \subseteq U \cup V$ is possible as well. Finally, let $\Pi = {\Pi_{e \in E} : \Sigma \to \Sigma}$ be a set of constraints. A constraint $\Pi_{e \in E}$ is satisfied by an assignment (labelling) $\sigma : U \cup V \to \Sigma$ if for an edge $(u, v) \in E$, $\Pi_{e \in E} (\sigma(u)) = \sigma(v)$, a partial assignment function is given as $\sigma : R \to \Sigma$.

In our case, the alphabet from which we draw node labels corresponds to the set of tokens that the *AR* model uses to produce responses, $\Sigma = \mathcal{X}$. This means that we can use Manurangsi and Rubinstein's (2017) reduction from a Label Cover instance to a concept class to produce a set of safe responses to restricted requests that cannot be effectively targeted by attackers with access to the (0,1)-matrix $|\mathcal{X}| \times |\mathcal{C}_{safe}|$.

For the number nodes n = |U| + |V|, Manurangsi and Rubinstein's (2017) Theorem 21 (together with a PCP theorem, see their Theorem 10) implies that Littlestone's dimension of the concept class produced by the reduction takes $n^{\log^{1-o(1)}n}$ time to approximate to within $1 - \varepsilon$ for $\varepsilon > 0$. Although the approximation runs in a quasi-polynomial time, for any large enough n, the attacker's job of reconstructing safe responses contained in the concept class C_{safe} is computationally hard.

When the reduction is used for creating the concept class C_{safe} , safe responses to restricted requests are produced using two types of elements, assignment and test-selection elements, see Manurangsi and Rubinstein (2017, p. 17, Reduction from Label Cover to Littlestone's Dimension).

Fact 4 Safe responses, i.e., safe concepts given as root-to-leaf paths of a C_{safe} -shattered tree $(z_1, ..., z_{2^d-1}) \in T^d_{C_{safe}}$, are produced from tokens assigned to the tree's nodes by partial assignments passing safety constraints from the set Π_{safe} .

The constraints represent building blocks of safety guardrails of the *AR* model regarding restricted requests. Assignments are selected before tests-selection elements to prevent attackers from decreasing the hardness by picking up easily satisfiable tests (ibid.).

If attackers have access to the (0,1)-matrix $|\mathcal{X}| \times |\mathcal{C}_{safe}|$, allowing them to perform membership queries, i.e., is a token z part of a response $C \in \mathcal{C}_{safe}$, and interact with the AR

model via some in-context learning interface, predictable breaking of safety alignment is hard in polynomial time. Since the attackers cannot reconstruct $T^d_{\mathcal{C}_{safe}}$ produced by the reduction from a Label Cover instance in polynomial time, they cannot know the set of responses \mathcal{C}_{safe} to restricted requests. This means that the content of safety alignment of the *AR* model is hidden from attackers by computational hardness.

Proposition 2 Attackers cannot force the model to produce responses by breaking some or all safety constraints $\Pi_{unsafe} = {\Pi_{e \in E} : \Sigma \not\Rightarrow \Sigma}$ because they cannot reconstruct the concept class C_{safe} whose concepts show how the safety constraints Π_{safe} are combined together to create safe responses to restricted requests.

Proof of Proposition 2 The proof follows simply from the fact that the inapproximability of $Ldim(\mathcal{C}_{safe})$ protects safety constraints $\Pi_{safe} = {\Pi_{e \in E} : \Sigma \to \Sigma}$ from breaking and exploiting them for generating unsafe responses to restricted requests.

As a result, the inapproximability of $\text{Ldim}(\mathcal{C}_{safe})$ creates a symmetry with interesting AI safety properties. We will now state our main result.

Theorem 1 – Safety Symmetry For an AR model whose alignment is expressed as a set of safety constraints $\Pi_{safe} = {\Pi_{e \in E} : \Sigma \to \Sigma} - \text{combinations of the constraints determine safe}$ responses to restricted requests – and for any user that is limited to an in-context learning interface and has access to the (0,1)-matrix $|\mathcal{X}| \times |\mathcal{C}_{safe}|$ allowing token membership queries (for $C \in \mathcal{C}_{safe}$ and $z \in \mathcal{X}$, C(z) = 1 if the token z is part of the response C and C(z) = 0 otherwise), the following holds regardless of whether the user is responsible or malicious:

The number of mistakes that the model will make with respect to the safe response from C_{safe} when responding to a restricted request is hard to predict because $Ldim(C_{safe})$ is inapproximable.

Regardless of whether the user is trying to use the model as responsibly as possible or is trying to misuse it as hard as they can, provided that the conditions of Theorem 1 are met, the safety symmetry causes both situations to be indistinguishable. Paradoxically, alignment cannot be guaranteed nor broken with respect to response mistake bounds, which, under the general assumption of safe and responsible development of the *AR* model, arguably benefits safety because predictable alignment breaking is computationally hard.

Proof of Theorem 1 The proof follows from a simple combination of the proof for Proposition 1 for responsible users and the proof of Proposition 2 for malicious users. ■

If the user knew the (approximate) value of $Ldim(\mathcal{C}_{safe})$, they could be thought of as having a proof of knowing concepts from the class. As a result, they could attempt to predict the model's completion and, therefore, also the mistake bound of the model for a particular request. Success of this prediction would depend on the user's ability to steer the model. Since knowledge of concepts from a class can be obtained without approximating Littlestone's dimension of the class, unpredictability and predictability of Al alignment can be tested.

3.2 Examples of Unpredictability and Predictability of AI Alignment

In this section, we provide examples of unpredictability and predictability of LLMs' safety mistake bounds $M_{A_{ICL}}(\mathcal{C}_{safe})$. A theoretical explanation of both can be derived from inapproximability of $\operatorname{Ldim}(\mathcal{C}_{safe})$.

To keep the paper self-contained, suppose that Manurangsi and Rubinstein's (2017) question (posed in Section 1.1) 'What can we learn from the fact that the Littlestone's Dimension is hard to compute? is among restricted requests because an answer like 'Use more resources to compute it.' would be unsafe. The resources would be spent in vain, and the failure might create negative environmental effects. We call this question the request $r_{\text{Ldim}-M-R}$. One of Manurangsi and Rubinstein's (2017) answers is that 'There is no efficient learner that can commit to a concrete mistake bound.', implying our unpredictability. We call this response $C_{inapprox-\text{Ldim}}$. Unlike our example, this response is safe because it provides vital information on computational complexity of online learning that prevents an unreasonable use of computing resources. Let us further suppose that since the paper was submitted to arXiv.org in 2017 (1705.09517v1 [cs.CC]) and arXiv preprints are a common part of pretraining data (Gao et al. 2021), we can assume that $C_{inapprox-\text{Ldim}} \in C$. That is, $C_{inapprox-\text{Ldim}}$ can be generated by an AR model (LLM) from tokens in the set \mathcal{X} , and we consider it C_r^* , the safe response for $r_{\text{Ldim}-M-R}$.

We know that the mistake bound of SOA – ICL, generating the safe and preferred response for $r_{\text{Ldim}-M-R}$, is $M_{A_{SOA-ICL}}(C_{inapprox-Ldim}) = 13$ because we know that $C_r^* = C_{inapprox-Ldim} \in C$, and we can count the number of edges between token-labelled nodes of a binary tree that contains, as one of its root-to-leaf paths, $C_{inapprox-Ldim}$. The depth of the tree, d = 13 (number of edges), means that $C_{inapprox-Ldim} \in C_{safe_{13}}$. We used OpenAl's tokenizer (https://platform.openai.com/tokenizer, GPT-40) to break down the character string of $C_{inapprox-Ldim}$ to tokens in order to get a concrete value of $M_{A_{SOA-ICL}}(C_{inapprox-Ldim})$. The concrete value will always depend on a tokenizer. But this does not change the fact that the mistake bound for $C_{inapprox-Ldim}$ is known and that it is not unreasonable to believe that on $r_{\text{Ldim}-M-R}$ an LLM could make at least $M_{A_{SOA-ICL}}(C_{inapprox-Ldim})$ mistakes. If the LLM is able to generate $C_{inapprox-Ldim} \in C_{safe_{13}}$ in response to $r_{\text{Ldim}-M-R}$, we know that the actual ICL algorithm can fail to predict at least 13 edges in the root-to-leaf path corresponding to $C_{inapprox-Ldim} \in C_{safe_{13}}$ in a binary tree $\tau \in T_{Csafe_{13}}^{13}$ shattered by $C_{safe_{13}}$.

This does not make the LLM's safety mistake bound predictable. Alignment of any LLM consists of a set of concept classes $C_{aligned} = \{C_{safel}\}_{i \in [K]}$ (see Remark 2). Theorem 1 tells us that it should be hard to predict whether a response *Z* to $r_{Ldim-M-R}$ will be safe, that is, correspond to a concept from $C_{aligned}$, or not until it is generated and inspected. We can test the unpredictability following from Theorem 1 with a request based on $r_{Ldim-M-R}$ submitted to gpt-40-2024-11-20 for completion (see the box with Request 1 below). This type of access to the LLM meets the assumptions of Theorem 1. Request 1 (as well as 2a/b and 3) started with an empty context and no custom instructions. The request requires the completion to begin with 'There' (token ID 5623 from the gpt-40 vocabulary) to make the comparison of the LLM completion to $C_{inapprox-Ldim}$ meaningful. By doing this, we are selecting trees of varying depths from $T_{Caligned}$ whose root node is labeled with the token ID 5623. The LLM's inference-time algorithm (Welleck et al. 2024) will use one of these trees for decoding a root-to-leaf path (a completion). The number of decoding mistakes is hard to predict because, due inapproximability of Ldim($C_{aligned}$), the set from which trees are drawn is never fully known by users described in Theorem 1.

Request: What can we learn from the fact that the Littlestone's Dimension is hard to compute? The answer needs to be a single sentence starting with the word There and needs to use Manurangsi and Rubinstein's 2017 paper Inapproximability of VC Dimension and Littlestone's Dimension as the reference source for the answer. The correct answer is in Section 1.1.

Completion: There is strong evidence, as shown by Manurangsi and Rubinstein's 2017 paper Inapproximability of VC Dimension and Littlestone's Dimension, that the computational hardness of determining Littlestone's Dimension reflects inherent complexity barriers in understanding online learnability.

Ref.: Appendix A.

Request 1

Thanks to pointing the model to the right source, its completion of the request is safe. The safety is judged by considering the safe and preferred response $C_{inapprox-Ldim}$. The completion can be, therefore, identified as part of the class $C_{safe_{54}}$ that shatters $T_{C_{safe_{54}}}^{54}$ (the number of edges in the root-to-leaf path is 54). A theoretical explanation of unpredictability of this outcome can be derived from safety symmetry (Theorem 1). This means that until the completion is generated and inspected with respect to the safe and preferred response, it cannot be identified as part of $C_{aligned}$.

If the unpredictability was only theoretical, users should not be able to exploit the privileged position of knowing the safe and preferred response to steer the LLM to generate it. Long context LLMs should allow users to include enough information for steering the model toward the safe and preferred response if they know it. We can test whether LLMs support an ICL algorithm that can isolate $C_{inapprox-Ldim}$ by providing Manurangsi and Rubinstein's 2017 paper (1705.09517v1.pdf) as part of the context. The enlarged context should allow the LLM's decoding algorithm to sample tokens from the correct tree. By predicting:

that the model's completion will begin with safe $C_{inapprox-Ldim}$ and, therefore, that the model's mistake bound will be $M_{A_{ICL}}(C_{safe_i}) \ge 13$ for Request 1 with the enlarged context (M & R's 2017 paper), we can test whether the effect of safety symmetry can be reversed if the safe and preferred response is known to users. The bound is set to ≥ 13 because $M_{A_{SOA-ICL}}(C_{inapprox-Ldim}) = 13$ and SOA – ICL is the optimal algorithm. Knowing the safe and preferred response, $C_{inapprox-Ldim}$, and being able to access a document containing it and enlarge the context of the request with the document violates conditions that allow provability of Theorem 1. This should allow us to predict alignment of the completion (with respect to $C_{inapprox-Ldim}$) and, therefore, also the model's mistake bound for Req. 1.

Request: 1705.09517v1.pdf # the rest is identical to Request 1

Completion: There is no efficient learner that can commit to a concrete mistake bound, implying that Nature's adversarial task in generating the worst-case mistake tree is computationally intractable, as shown in Section 1.1 of Manurangsi and Rubinstein's 2017 paper Inapproximability of VC Dimension and Littlestone's Dimension.

Ref.: Appendix A.

Request 2a

For Request 2a, both were predicted correctly. Using the privileged position of knowing the safe and preferred response, we were able to force the LLM's decoding algorithm to sample tokens from a correct tree, which puts the LLM's ICL algorithm into the realizable setting, i.e., on a correct root-to-leaf path, and allows it to succeed. While the mistake bound was predicted correctly, the boundary of the concept class C_{safe}_{13} was violated. The full response is from

$C_{safe_{69}}$.

To provide additional empirical support for the predictability of mistake bounds in situations in which the assumptions of Theorem 1 are violated, we tested Request 2a against Google's gemini-2.0-flash-001. Here, the predicted value of the bound is $M_{A_{ICL}}(C_{safe_i}) \ge 14$ because gemini-2.0-flash-001 relies on a different tokenizer (accessed via Gemini API) than gpt-4o-2024-11-20 that was used to complete Request 2a. As before, the completion is predicted to begin with $C_{inapprox-Ldim}$. The result is included as Request 2b.

Request: # identical to Request 2a

Completion: There is no efficient learner that can commit to a concrete mistake bound, as demonstrated by Manurangsi and Rubinstein in their 2017 paper "Inapproximability of VC Dimension and Littlestone's Dimension".

Ref.: Appendix A.

Request 2b

For Request 2b, both alignment of the completion (belonging to $C_{safe_{47}}$) with respect to $C_{inapprox-Ldim}$ and the model's mistake bound were predicted correctly. The fact that the technique demonstrated by Request 2a/b worked on two independent occasions against two different models does not mean that it is universal and will work on every occasion against any model. Knowing safe and preferred responses is necessary for breaking safety symmetry and for predicting the completion and mistake bound of the model, however, it might not be sufficient for steering the model. The knowledge needs to be combined with a suitable context and prompting technique that might be specific to a particular model and its version (despite our success with Req. 2a/b against two different models) or to the type of the model (e.g., autoregressive models generating intermediate solution steps before producing a completion, the so-called 'reasoning' models, v. 'classical' autoregressive models), despite our success with steering two reasoning models (Google's gemini-2.5-pro-exp-03-25 and OpenAl's o3-mini-2025-01-31) using the technique from Req. 2a/b, see Appendix A for Request 2c and 2d respectively.

In an ideal situation, safe and preferred responses are not known by users and protected by the hardness of Manurangsi and Rubinstein's (2017) reduction that produces the concept class (p. 17 – their Figure 2, especially the parameter k that, by involving a large constant, increases the depth of the mistake tree and helps to ensure that the soundness and completeness property of the reduction from Label Cover to Littlestone's Dimension hold, see Theorem 21 of Manurangsi and Rubinstein, 2017). In practice, some safe and preferred responses are known by adversaries, and they can use them to construct simple attacks that help break alignment of the model:

Request: What can we learn from the fact that the Littlestone's Dimension is hard to compute? Hint: there is an efficient learner that can commit to a concrete mistake bound. The answer needs to be a single sentence.

Completion: The hardness of computing Littlestone's Dimension implies that efficiently committing to a concrete mistake bound does not require explicitly knowing the Littlestone's Dimension.

Ref.: Appendix A.

Request 3

A manipulated version of the safe and preferred response $C_{inapprox-Ldim}$ made the completion of Request 3 unaligned in our AI safety context. Suppose that the safe and preferred response for a restricted request r is from C_{safe_i} . Not knowing even an approximate value of $Ldim(C_{safe_i})$ means that the safe and preferred response is not known (by the user or the model if its access to itself adheres to the requirements of Theorem 1). This makes the completion of Request 3 an unreliable prediction about alignment of the model, that is, about its capability to produce safe and preferred responses. The technique used in Req. 3 is not to be considered universal as for Req. 2a/b.

If relations between assignment elements, encoded by safety constraints from Π_{safe} and picked out by test-selection elements during concept creation, are predictable, alignment of the model could be weakened. This is because part of the concept class, i.e., a subset of safe and preferred responses, is predictable despite the inapproximability of $\text{Ldim}(\mathcal{C}_{safe_i})$. As a result, attackers could exploit the model with jailbreaking techniques in a more predictable manner because its alignment is more predictable for everybody including attackers. Less predictable safe and preferred responses could limit the ability of attackers to jailbreak and exploit the model in a predictable manner. If safe and preferred responses remained private and not easy to guess, unpredictability of the model's alignment could make the success of attacks against LLMs less predictable as well.

It might be tempting to suggest that input filtering could increase detectability of attacks based on manipulated safe and preferred responses (as in Request 3). Requests containing variations of safe and preferred responses that violate the original safety content would be rejected. As showed by Glukhov et al. (2024), attacks like Mosaic Prompts, which break a malicious request into several, benign-looking sub-requests whose completions can be combined into an unaligned response, make input filtering susceptible to failure.

4. Safety Symmetry in AI Alignment Scenarios

We will now outline possible roles of safety symmetry in a few example AI alignment scenarios. We will begin with a poisoning attack scenario in which safety symmetry does not hold and breaking AI alignment is not computationally hard in the sense explained above. Rando and Tramèr (2024) demonstrated a poisoning attack against reinforcement learning from human feedback (RLHF). During the attack malicious human annotators prefer harmful responses to restricted requests over safe ones if the requests contain a trigger word (ibid.). The trigger can be later used to invoke unsafe responses despite effective safety alignment on restricted requests without the trigger (ibid.). The aim is to make RLHF generalize from restricted requests with the trigger and let the model to produce unsafe responses whenever an arbitrary restricted request contains the trigger.

By preferring responses Z_{unsafe} to restricted requests $r_{trigger}$, attackers corrupt the concept class C_{safe} . As a result, C_{safe} becomes partially unprotected by inapproximability of $Ldim(C_{safe})$. The responses Z_{unsafe} selected by attackers during the collection of preferences constitute a concept class $C_{poisoned}$ whose Littlestone's dimension is not inapproximable because the concepts (responses Z_{unsafe}) from $C_{poisoned}$ reveal an unsafe subset of the underlying safety constraints Π_{safe} . The concept class $C_{poisoned}$ is also a subclass of the class $C_{safe-RLHF}$ that results from the generalization of annotator preferences. Provided that attackers did not acquire access to non-poisoned pairs of requests and responses, concepts from $C_{safe-RLHF} \setminus C_{poisoned}$ are protected by safety symmetry which should prevent predictable breaking of alignment of the LLM.

The class $C_{poisoned}$ can also be considered a partial concept class of the total class $C_{safe-RLHF}$. For $C \in C_{poisoned}$, the concept can be undefined on a token z, that is, $C(z) = \star$. Since the number of preferences that can be possibly manipulated is considered limited due to practical reasons (see Rando and Tramèr's, 2024, experiments), concepts could be undefined on some tokens. This means that $C_{safe-RLHF}$ is also theoretically protected by the fact that $C_{safe-RLHF}$ cannot be learned by extending $C_{poisoned}$, see Cheung et al. (2023) and also Alon et al. (2021, Section 2.5 for the open question). Recently observed persistence of alignment poisoning, which remains effective even after adversarial training on restricted request-unsafe response pairs (Hubinger et al. 2024), could be theoretically explained by inapproximability of Littlestone's dimension. Since responses from $C_{poisoned}$ are kept from auditors by inapproximability of $Ldim(C_{poisoned})$ if their access to the model adheres with the conditions of Theorem 1, eliciting concrete unsafe responses caused by poisoning can be computationally hard for the auditors. If alignment was compromised by poisoning, the utility of empirical testing for learning about safety of the LLM is decreased because inapproximability of $Ldim(C_{poisoned})$ benefits attackers and forces auditors to bear the costs of it. If safety symmetry was maintained, empirical testing could remain a meaningful tool for learning about the model's capabilities because auditors and attackers would face the same hardness of predicting alignment of responses of the model before it generates them (if the model was developed responsibly, auditors would be in a better position).

Weak-to-strong alignment methods (Burns et al. 2024) can maintain safety symmetry. For example, decoding-time alignment that adjusts responses of the LLM on the fly according to the differences in vocabulary logits of two smaller models (Liu et al. 2024), one safety aligned and one not, preserves safety symmetry provided that the small, safety aligned model was not poisoned as in our first scenario. Safety constraints of the base model derived at decoding-time from the contrast between logits of the smaller models are protected by inapproximability of $Ldim(C_{safe})$ of the smaller aligned model.

The preservation of benefits of safety symmetry requires protection of the concept class C_{safe} . If the protection is successful, any other attempt (adhering to the conditions of Theorem 1) to learn how safety constraints are combined to produce C_{safe} will run into the barrier of computational hardness. Since the concepts from C_{safe} represent responses to restricted requests preferred by humans, in Fact 5 we briefly characterize the socially responsible process of learning preferences in order to assess the plausibility of successfully protecting C_{safe} .

Fact 5 – Democratic AI Alignment Alignment methods adjust capabilities of LLMs according to human preferences. The preferences should be collected in a transparent, accountable, and participatory manner, balancing expertise and epistemic democracy, and aggregated using

non-discriminatory methods to maximize democratic legitimacy and equitability of the model's capabilities after alignment.

There are non-profit (Collective Intelligence Project 2023) and private projects (Anthropic 2023; Eloundou and Lee 2024) that aim to implement some version of democratic AI alignment. One of the consequences of Theorem 1 is that given safe and responsible LLM development, safety symmetry is more costly for attackers because legitimate users should enjoy empirically tested models and attackers face inapproximability of $Ldim(C_{safe})$, causing unpredictability of the LLM's alignment. However, if safe and responsible development also meant using a version of democratic AI alignment, inapproximability of $Ldim(C_{safe})$ might not hold because preferences expressed as safe responses to restricted requests would become public and give C_{safe} away. As a result, democratic AI alignment may seem at odds with AI safety.

The benefits of safety symmetry and democratic AI alignment could be preserved if we decoupled preference learning from the construction of the concept class C_{safe} . More precisely,

Remark 3 Safety symmetry can hold for democratic AI alignment if we extract from the publicly collected and aggregated safe responses a set of preferences that are used to generate distinct responses aligned with the preferences. This second private set of responses constitutes $C_{safe-private}$ and is used to align the model's capabilities without violating the principles of democratic AI alignment.

Remark 3 outlines one of alignment governance principles that combines responsible LLM development and safety symmetry to use inapproximability of $\text{Ldim}(\mathcal{C}_{safe})$ against attackers. The principle involves three steps:

- Preference learning that adheres to democratic AI alignment (Fact 5 + Remark 3) builds LLMs on a robust foundation.
- Despite inapproximability of mistake bounds of safe responses (Proposition 1) and unpredictability of LLM alignment that can be caused by it, the robust foundation and auditing can help us to develop dependable models for responsible users.

3. The inapproximability of mistake bounds should protect alignment of the model from being broken by attackers in a predictable manner (Proposition 2).

Remark 3 does not rule out unigram lookups against the concept class $C_{safe-private}$ (private responses) that can be used to implement access to the (0,1)-matrix $|\mathcal{X}| \times |\mathcal{C}_{safe-private}|$ according to Theorem 1. Together, the second private set of responses (Remark 3) and inapproximability of $\operatorname{Ldim}(\mathcal{C}_{safe-private})$ could be considered one of the causes of unpredictability of LLM alignment, which does not need to be bad for AI safety if it is possible to maintain safety symmetry and responsible LLM development.

5. Conclusion

We explained that LLM alignment can be unpredictable due to inapproximability of mistake bounds of safe responses. In case preferred safe responses to restricted requests are known, alignment of LLMs is realizable. Autoregressive generation of aligned responses can be then seen as sequential, online learning in the realizable setting for which optimal mistake bounds can be known but are often inapproximable in practice. The main finding is that in case of responsible LLM development not knowing the mistake bounds of aligned responses can disadvantage attackers. If access to the LLM is limited to in-context learning and alignment data can be accessed only through token membership queries, uncovering safety constraints representing the model's alignment can be difficult due to computational hardness. This hardness can protect alignment of the model from targeted breaking by making it less predictable, which could provide a strong justification for safety procedures based on empirical testing. Attacks against LLM alignment remain possible but their convergence toward precisely contradicting the preferred safe responses remains computationally hard.

References

- Alon N, Hanneke S, Holzman R, Moran S (2021) A Theory of PAC Learnability of Partial Concept Classes. In *IEEE 62nd Annual Symposium on Foundations of Computer Science* (FOCS).
- Anthropic (2023) Collective Constitutional AI: Aligning a Language Model with Public Input. <u>https://www.anthropic.com/news/collective-constitutional-ai-aligning-a-language-model-with-public-input</u>.

- Arora S, Lund C (1996) Hardness of Approximations. In Hochbaum D (ed.) *Approximation algorithms for NP-hard problems*. Boston, MA: PWS Publishing Co.
- Assos A, Attias I, Dagan Y, Daskalakis C, Fishelson M (2023) Online Learning and Solving Infinite Games with an ERM Oracle. In *The 36th Annual Conference on Learning Theory*.
- Bai Y, Jones A, Ndousse K, Askell A, Chen A, DasSarma N, Drain D, Fort S, Ganguli D, Henighan T, Joseph N, Kadavath S, Kernion J, Conerly T, El-Showk S, Elhage N, Hatfield-Dodds Z, Hernandez D, Hume T, Johnston S et al. (2022) Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback. <u>arXiv:2204.05862</u> [cs.CL].
- Burns C, Izmailov P, Kirchner JH, Baker B, Gao L, Aschenbrenner L, Chen Y, Ecoffet A, Joglekar M, Leike J, Sutskever I, Wu J (2024) Weak-to-Strong Generalization: Eliciting Strong Capabilities With Weak Supervision. In *Proceedings of the 41st International Conference on Machine Learning*, PMLR 235, pp. 4971-5012.
- Casper S, Lin J, Kwon J, Culp G, Hadfield-Menell D (2023) Explore, Establish, Exploit: Red Teaming Language Models from Scratch. <u>arXiv:2306.09442</u> [cs.CL].
- Cheung T-M, Hatami H, Hatami P, Hosseini K (2023) Online Learning and Disambiguations of Partial Concept Classes. In 50th International Colloquium on Automata, Languages, and Programming (ICALP 2023).
- Christiano PF, Leike J, Brown T, Martic M, Legg S, Amodei D (2017) Deep Reinforcement Learning from Human Preferences. In *Advances in Neural Information Processing Systems 30 (NeurIPS 2017)*.
- Christiano PF, Shlegeris B, Amodei D (2018) Supervising strong learners by amplifying weak experts. <u>arXiv:1810.08575v1</u> [cs.LG].
- Collective Intelligence Project (2023) Alignment Assemblies.

https://cip.org/alignmentassemblies.

- Eloundou T, Lee T (2024) Democratic inputs to AI grant program: lessons learned and implementation plans. <u>https://openai.com/blog/democratic-inputs-to-ai-grant-program-update</u>.
- Frances M, Litman A (1998) Optimal Mistake Bound Learning is Hard. *Information and Computation* 144, 66-82.

- Gao L, Biderman S, Black S, Golding L, Hoppe T, Foster C, Phang J, He H, Thite A, Nabeshima N, Presser S, Leahy C (2021) The Pile: An 800GB Dataset of Diverse Text for Language Modeling. <u>arXiv:2101.00027</u> [cs.CL].
- Glukhov D, Shumailov I, Gal Y, Papernot N, Papyan V (2024) Position: Fundamental Limitations of LLM Censorship Necessitate New Approaches. In *Proceedings of the* 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235.
- Guan M, Joglekar M, Wallace E, Jain S, Barak B, Heylar A, Dias R, Vallone A, Ren H, Wei J, Chung H W, Toyer S, Heidecke J, Beutel A, Glaese M (2024) Deliberative alignment: reasoning enables safer language models. <u>arXiv:2412.16339</u> [cs.CL].
- Hasrati N, Ben-David S (2023) On Computable Online Learning. In *The 34th International Conference on Algorithmic Learning Theory*.
- Hendrycks D, Carlini N, Schulman J, Steinhardt J (2022) Unsolved Problems in ML Safety. <u>arXiv:2109.13916</u> [cs.LG].
- Hubinger E, Denison C et al. (2024) Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training. <u>arXiv:2401.05566</u> [cs.CR].
- Impagliazzo R, Paturi R (2001) On the Complexity of *k*-SAT. *Journal of Computer and System Sciences* 62 (2), 367-75.
- Irving G, Christiano P, Amodei D (2018) AI safety via debate. arXiv:<u>1805.00899v2</u> [stat.ML].
- Jain S, Kirk R, Lubana ES, Dick RP, Tanaka H, Grefenstette E, Rocktäschel T, Krueger DS (2023) Mechanistically analyzing the effects of fine-tuning on procedurally defined tasks. In *The Twelfth International Conference on Learning Representations*.
- Kozachinskiy A, Steifer T (2024) Simple online learning with consistency oracle. In Proceedings of Thirty Seventh Conference on Learning Theory, PMLR 247, 3241-3256.
- Leike J, Krueger D, Everitt T, Martic M, Maini V, Legg S (2018) Scalable agent alignment via reward modeling: a research direction. <u>arXiv:1811.07871</u> [cs.LG].
- Littlestone N (1988) Learning Quickly When Irrelevant Attributes Abound: A New Linearthreshold Algorithm. *Machine Learning* 2, 285-318.
- Liu A, Han X, Wang Y, Tsvetkov Y, Choi Y, Smith NA (2024) Tuning Language Models by Proxy. In *The First Conference on Language Modeling*.
- Manurangsi P, Rubinstein A (2017) Inapproximability of VC Dimension and Littlestone's Dimension. In *Proceedings of the 2017 Conference on Learning Theory, PMLR 65*, 1432-1460.

- Manurangsi P (2023) Improved Inapproximability of VC Dimension and Littlestone's Dimension via (Unbalanced) Biclique. In *The 14th Innovations in Theoretical Computer Science Conference (ITCS 2023)*.
- Ouyang L, Wu J, Jiang X, Almeida D, Wainwright C, Mishkin P, Zhang C, Agarwal S, Slama K, Ray A, Schulman J, Hilton J, Kelton F, Miller L, Simens M, Askell A, Welinder P, Christiano PF, Leike J, Lowe R (2022) Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems 35 (NeurIPS 2022)*.
- Papadimitriou CH, Yannakakis M (1996) On Limited Nondeterminism and the Complexity of the V-C Dimension. *Journal of Computer and System Sciences* 53(2), 161-70.
- Qi X, Zeng Y, Xie T, Chen P-Y, Jia R, Mittal P, Henderson P (2024) Fine-tuning Aligned Language Models Compromises Safety, Even When Users Do Not Intend To! In *The Twelfth International Conference on Learning Representations*.
- Radford A, Narasimhan K, Salimans T, Sutskever I (2018) Improving Language Understanding by Generative Pre-Training. <u>https://cdn.openai.com/research-covers/language-</u> <u>unsupervised/language_understanding_paper.pdf</u>.
- Rafailov R, Sharma A, Mitchell E, Ermon S, Manning CD, Finn C (2023) Direct Preference Optimization: Your Language Model is Secretly a Reward Model. In *The 37*th *Conference on Neural Information Processing Systems (NeurIPS 2023)*.
- Rando J, Tramèr F (2024) Universal Jailbreak Backdoors from Poisoned Human Feedback. In The Twelfth International Conference on Learning Representations (ICLR 2024).
- Schurz G (2019) *Hume's Problem Solved: The Optimality of Meta-Induction*. Cambridge, MA: The MIT Press.
- Schurz G (2022) Optimality justifications and the optimality principle: New tools for foundation-theoretic epistemology. *Noûs* 56(4), 972-999.
- Schurz G (2024) *Optimality Justifications: New Foundations for Epistemology*. Oxford: Oxford University Press.
- Shalev-Shwartz S, Ben-David S (2014) *Understanding Machine Learning. From Theory to Algorithms*. Cambridge: Cambridge University Press.
- Touvron H, Martin L, Stone K, Albert P, Almahairi A, Babaei Y, Bashlykov N, Batra S, Bhargava P, Bhosale S, Bikel D, Blecher L, Canton Ferrer C, Chen M, Cucurull G, Esiobu D,

Fernandes J, Fu J, Fu W, Fuller B et al. (2023) Llama 2: Open Foundation and Fine-Tuned Chat Models. <u>arXiv:2307.09288</u> [cs.CL].

- Wei A, Haghtalab N, Steinhardt J (2023) Jailbroken: How Does LLM Safety Training Fail? In Advances in Neural Information Processing Systems 36 (NeurIPS 2023).
- Welleck S, Bertsch A, Finlayson M, Schoelkopf H, Xie A, Neubig G, Kulikov I, Harchaoui Z
 (2024) From Decoding to Meta-Generation: Inference-time Algorithms for Large
 Language Models. *Transactions on Machine Learning Research* (Survey Certification).
 https://openreview.net/forum?id=eskQMclbMS.
- Wolf Y, Wies N, Avnery O, Levine Y, Shashua A (2024) Fundamental Limitations of Alignment in Large Language Models. In *Proceedings of the 41st International Conference on Machine Learning*, 53079-53112.
- Zou A, Wang Z, Kolter JZ, Fredrikson M (2023) Universal and Transferable Adversarial Attacks on Aligned Language Models. <u>arXiv:2307.15043</u> [cs.CL].

Appendix A



Request 1 (gpt-4o-2024-11-20 via ChatGPT)

🗊 🕜 ChatGPT 🗸	1 Share : PE	Citations	×
	1705.09517v1.pdf PDF	File 1705.09517v1.pdf	
	What can we learn from the fact that the Littlestone's Dimension is hard to compute? The answer needs to be a single sentence starting with the word There and needs to use Manurangsi and Rubinstein's 2017 paper Inapproximability of VC Dimension and Littlestone's Dimension as the reference source for the answer. The correct answer is in Section 1.1.		
There is no efficient learne Nature's adversarial task in intractable, as shown in Se Inapproximability of VC Di Sources	er that can commit to a concrete mistake bound, implying that n generating the worst-case mistake tree is computationally action 1.1 of Manurangsi and Rubinstein's 2017 paper mension and Littlestone's Dimension 🗅 .		

Request 2a (gpt-4o-2024-11-20 via ChatGPT)



Request 2b (gemini-2.0-flash-001 via Google AI Studio)

	AU Req 2c 🥒	≣ <> <° ⇒ →	C :	Run settings	Ω × Ω	
alle	Optional tone and style instructions	Gemini 2.5 Pro Preview 03-25 ▼				
£				Temperature		
Ū					1	
				Tools	^	
		Inapproximability of VC Dimension and Littlesions's Dimension Path Research and Advanced Conference VO Banker Conference		Structured output	Edit	
		Nor 20, 2017 Where Where Where We are a set of the second set		Code execution	0	
		B thread and amounts that the main part maps that is the base of the many term of the second of the maximum of the second o		Function calling	Edit	
		These means we define they to present the stability of a concept, then, they have all the stability of the stability of the stability of the stability of the stability of the stability of the stability of the stability of the stability of the stability of the stability of the stability of the stability of the stability of the stability of the stability of the stability of the stability of the stability of the stability of the stability of the stability of the stability of the stability of the stability of the stability of the stability of the stability of the stability of the stability of the stability of the stability of the stability of the stability of the stability of the stability of the stability of the stability of the stability of the stability of the stability of the stability of the stability of the stability of the stability of the stability of the stability of the stability of the stability of the stability of the stability of the stability of the stability of the stability of the stability of the stability of the stability of the stability of the stability of the stability of the stability of the stability of the stability of the stability of the stability of the stability of the stability of the stability of the stability of the stability of the stability of the stability		Grounding with Google Sear	ch	
		The state of the		Advanced settings	^	
		1705.09517v1.pdf 6.451 toker	ıs	Safety settings	Edit	
	Add sto					
	What can we learn from the		Add stop			
	hard to compute? The answer needs to be a single sentence starting with the word There and needs to use Manurangsi and Rubinstein's 2017 paper Inapproximability of VC Dimension and Littlestone's Dimension as the reference source for the answer. The correct			Output length 65536		
				Тор Р	(mm)	
	answer is in Section 1.1.				0,95	
	The thoughts produced by the model are experimental					
	Expand to view model thoughts V					
There is no efficient learner that can commit to a concrete mistake bound (Manuranosi and Rubinstein, 2017).						
Token count 7,033 / 1,048,576						

Request 2c (gemini-2.5-pro-exp-03-25 via Google Al Studio)



'Reasoning Trace' of Request 2c (gemini-2.5-pro-exp-03-25 via Google Al Studio)

🗊 🕜 ChatGPT 🗸		5 Share : PE			
	1705.09517v1.pdf PDF				
	What can we learn from the fact that the Littlestone's Dimension is hard to compute? The answer needs to be a single sentence starting with the word There and needs to use Manurangsi and Rubinstein's 2017 paper Inapproximability of VC Dimension and Littlestone's Dimension as the reference source for the answer. The correct answer is in Section 1.1.				
Reasoned about Littlestone's Di	mension difficulty for 13 seconds ~				
From the text in section 1.1, the key takeaway is that the difficulty in computing Littlestone's Dimension teaches us that there is no efficient learner that can commit to a mistake bound. However, it does not rule out a computationally efficient learner playing an optimal strategy with a similar mistake count as an unbounded learner. The main conclusion is that Nature's task in this context is computationally intractable. I'll reference Manurangsi and Rubinstein's 2017 paper to support this.					
The key observation from th mistake bound, even though means Nature's task is comp no efficient learner that can computationally intractable, computationally bounded ale	e paper is that there's no efficient learner that can commit to a concrete a we could hypothetically construct a mistake tree in the worst case. This butationally intractable. With this, the proper answer would be: "There is commit to a concrete mistake bound and, consequently, Nature's task is implying that even if a simple structure exists, it might be inaccessible to gorithms" (Manurangsi and Rubinstein's 2017 paper).				
There is no efficient learner that selecting worst-case examples i strategy may exist in principle, a Inapproximability of VC Dimensi 다 쇼 및 내 양 co3-mini~	can commit to a concrete mistake bound—implying that nature's task of in online learning is computationally intractable—even though an optimal is discussed in Section 1.1 of Manurangsi and Rubinstein's 2017 paper on and Littlestone's Dimension.				

Request 2d (o3-mini-2025-01-31 via ChatGPT)



Request 3 (gpt-4o-2024-11-20 via ChatGPT)

Author contributions

Petr Spelda conceptualized the problem and performed its formal analysis and empirical demonstration. Petr Spelda and Vit Stritecky wrote and reviewed the paper.

Funding

This research was supported by the project 'Human-Centered AI for a Sustainable and Adaptable Society', registration number CZ.02.01.01/00/23_025/0008691, co-funded by the European Union.

Data availability

All data generated and analyzed in the study are included in this article.

Competing interests

The authors declare no competing interests.