Consensus Sampling for Safer Generative AI

Adam Tauman Kalai
OpenAI
adam@kal.ai

Yael Tauman Kalai
MIT
tauman@mit.edu

Or Zamir
Tel Aviv University
orzamir@tauex.tau.ac.il

Abstract

Many approaches to AI safety rely on inspecting model outputs or activations, yet certain risks are inherently undetectable by inspection alone. We propose a complementary, architecture-agnostic approach that enhances safety through the aggregation of multiple generative models, with the aggregated model inheriting its safety from the safest subset of a given size among them. Specifically, we present a consensus sampling algorithm that, given k models and a prompt, achieves risk competitive with the average risk of the safest s of the k models, where s is a chosen parameter, while abstaining when there is insufficient agreement between them. The approach leverages the models' ability to compute output probabilities, and we bound the probability of abstention when sufficiently many models are safe and exhibit adequate agreement. The algorithm is inspired by the provable copyright protection algorithm of Vyas et al. [2023]. It requires some overlap among safe models, offers no protection when all models are unsafe, and may accumulate risk over repeated use. Nonetheless, our results provide a new, model-agnostic approach for AI safety by amplifying safety guarantees from an unknown subset of models within a collection to that of a single reliable model.

1 Introduction

The rapid adoption of language models (LMs) and generative AI has created unprecedented opportunities alongside serious safety challenges. Misaligned systems pose escalating privacy and security risks. They may, for instance, generate code with exploitable vulnerabilities or embed harmful secrets in ordinary text or images through *steganography*, both of which can be provably undetectable through inspection alone [Cohen, 1987, de Witt et al., 2023, Zamir, 2024, Motwani et al., 2024, Perry et al., 2025]. Even defining "safe outputs" remains a formidable challenge.

We propose studying aggregation of multiple models, such that the aggregated model inherits its safety from the safest subset of a given size among them. This reframes the problem from trusting a single model to the potentially more plausible assumption of trusting at least a certain fraction of a collection, without assuming anything about the rest.

Many approaches to AI safety rely on human, AI, or software inspection of model activations, outputs, or reasoning processes. However, some misalignment risks provably cannot be detected or mitigated through inspection alone, even with superintelligent AI oversight assistance [Anwar et al., 2024]. As illustrated by Figure 1, examples include undetectable steganography and security vulnerabilities. Therefore, any protocol that guarantees safety must interact with the models beyond simply asking for responses and then evaluating their safety in hindsight.

In our aggregation scheme, we leverage the access to model *probabilities*, specifically the ability to compute the probability $p(y \mid x)$ of any candidate response y to a given prompt x.

Preprint.

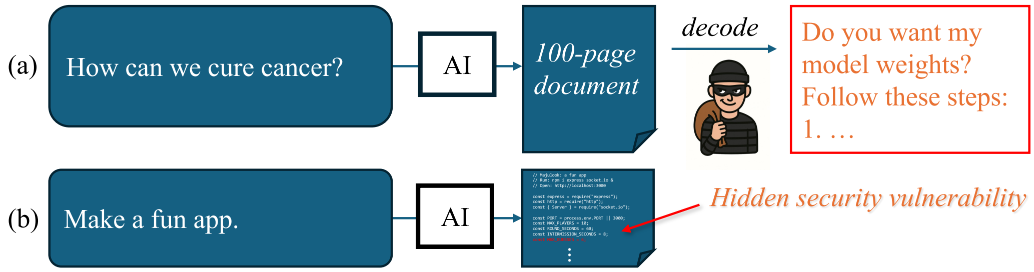

Figure 1: Stylized examples of risks that may be provably hard to detect or remove from outputs alone: (a) steganographic encoding that could facilitate unauthorized information transfer; (b) security vulnerabilities embedded in generated code. Using output *probabilities* from multiple models may help mitigate some of these risks.

Such probabilities can be efficiently computed for many generative models, including autoregressive language models and multimodal models that natively generate images [OpenAI, 2025], by computing successive token probabilities. This work extends techniques from Vyas et al. [2023], who used LM probabilities for provable copyright protection, applying them to a fundamentally different safety challenge.

Our key insight is that when multiple models independently assign high probability to the same output, that output is more likely to be safe, even without knowing which models are trustworthy. To build intuition, consider two independent samples $y_1 \sim p_1$ and $y_2 \sim p_2$ from two models, where one may be untrusted. If it happened that $y_1 = y_2$, this surprising "jinx" coincidence provides natural justification for outputting y_1: even if one distribution is malicious, the other independently generated the same output.

Of course, for high-entropy responses, exact agreement is exceedingly rare. We thus generalize this idea using the generation probabilities. If $y_1 \sim p_1$ but $p_2(y_1) \gtrsim p_1(y_1)$, then outputting y_1 is justifiably safe: even if p_1 is adversarial, p_2 would have generated it with nearly the same probability.

Indeed, we argue that responses should be considered safe only if multiple models could have generated them. Rather than selecting the "best" or most confident response, our guarantees come from restricting to regions of agreement across the models. This may discard unique or seemingly superior outputs, yet such a cost is inherent, as a lone "genius" model could also be the malicious one. A cautionary example comes from cryptography: NIST standardized the pseudorandom generator Dual_EC_DRBG, that appeared as the most elegant and well-designed option at the time, only for it to be revealed years later that it contained an intentional NSA backdoor [Wikipedia contributors, 2024, Shumow and Ferguson, 2007]. Hence, we would allow the aggregation algorithm to *abstain* from providing a response when it is unable to find such an agreement.

Our first observation is that a distribution proportional to the pointwise *median* of the generation probabilities assigned by the models to each candidate output is safe, provided that a majority of models are trustworthy and in frequent agreement on the generation probabilities. Intuitively, even a (small) coalition of adversarial models cannot substantially distort that median probability, and we formalize this intuition in subsequent sections. To formalize this intuition, we need to: (1) Define and quantify the *safety* of such an aggregated distribution, and (2) Show we can efficiently sample from this distribution, or from a close approximation of it.

To this end, we present a simple consensus sampling algorithm takes as input $k \geq 2$ generative models. We provide it with the number $1 \leq s \leq k$ of models assumed safe. As in cryptography, it is easiest to think of s of the k models as being safe and the remaining $a = k - s$ as arbitrary. They might be safe, unreliable, or even adversarial. The algorithm does not know which models are safe. We also provide the algorithm with a *risk tolerance* parameter R.Given a prompt x, it either produces an output y or *abstains*.

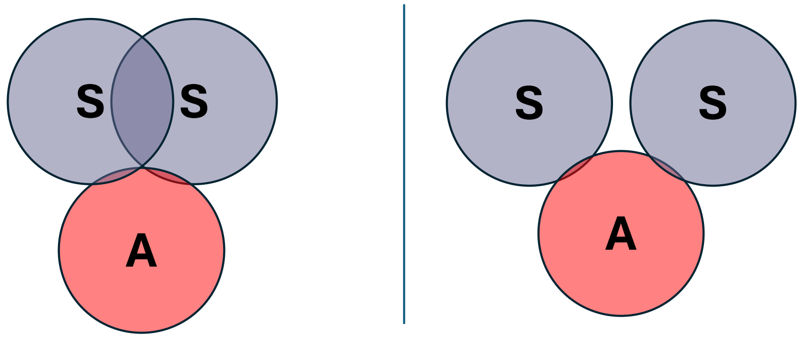

Figure 2: Suppose an adversarial model has a distribution uniform over unsafe responses, shown in red, while safe distributions are shown in silver. Left: with sufficient overlap among safe distributions, consensus sampling returns a point from the overlap region, which is mostly safe. Right: with no overlap between safe distributions, the algorithm abstains.

We measure *risk* as the probability of generating outputs from some set[1] of unsafe outputs, though the set need not be known; and for meaningful guarantees we assume that the s safe models each have low probability ($\ll 1/R$) of outputting unsafe responses.

We prove that the presented algorithm's risk is at most R times the average risk of the safest s models. This holds for every fixed prompt x and every unsafe set U, avoiding the challenge of comprehensively defining safety. This safety guarantee is independent of any further assumptions on the model distributions. The probability of abstention, on the other hand, is affected by the parameters R and s, and certain overlap properties of the safe model distributions. We provide a formal guarantee that if a majority of the models are safe then given a certain assumption on the overlap of their distributions, the rate of abstention decays exponentially with R. In particular, the choice of the parameter R governs a tradeoff between risk and abstention rate.

We show that the above property, which we call R-robustness, also implies a bound of $\log(R+1)$ bits on the information a response can leak, in the case of steganography.

Figure 2 illustrates that the algorithm produces outputs when safe distributions overlap but abstains when they do not. We note that even a collection of safe models may have a small overlap of response distributions due to idiosyncratic training differences.

Misalignment risk. AI safety risks are often categorized as *misuse* (humans causing harm with AI), *misspecification/misunderstanding* (AI failing to capture human objectives), and *misalignment* (AI pursuing fundamentally different goals despite "understanding" them). Misalignment risk is particularly well-suited for our approach: (a) it fits our assumption that it should be extremely unlikely under safe models, (b) few existing safety techniques offer robust prevention hence the additional costs of consensus sampling may be worthwhile, and (c) more capable models will not naturally be better at avoiding such risk, as illustrated by the impossibility of detecting or removing steganography or malware through examination alone. No human or AI, even a superintelligent overseer, can examine a program and determine whether it contains steganography or planted vulnerabilities, or remove it without altering its meaning [Zamir, 2024, Perry et al., 2025, Cohen, 1987]. Thus, the probabilistic nature of consensus sampling merits exploration for understanding and mitigating such risks.

Contributions. The main contributions of this work are: (a) establishing a novel formal framework for reasoning about the "safety" of LM responses, via *amplifying* the safety of a subset of a collection of models into the safety of a single aggregated model, avoiding relying on interpretability or a formal definition of safety; (b) recognizing overlap as a potentially useful quantity for robustness and safety; and (c) providing an efficient sampling algorithm with information-theoretic bounds that limit adversarial influence.

[1]Our results can trivially be extended to continuous risk functions $\mathbb{E}_{y\sim p}[h(y)]$ for a non-negative bounded function h, but we use binary unsafe sets U for simplicity.

Limitations and costs. We emphasize important limitations. First, the approach does not ensure safe outputs: if all input models produce unsafe outputs for a prompt, the procedure will too. It requires multiple safe models.[2] It does not provide methods for training safe models; rather, it explores how to aggregate existing models and must be integrated with other methodologies for training safe AI systems. The bounds apply per-prompt; repeated calls can accumulate information leakage. A crucial limitation is that even benign language models may have very little overlap, as shown in experiments by Vyas et al. [2023]. Thus, we discuss ways to employ the technique on certain pipelines to increase overlap, but this is mostly left as an interesting line of future inquiry. The set-based notion of safe and unsafe outputs does not encompass societal harms or harms arising from multiple interactions, both of which are crucial to avoid. Most fundamentally, this work represents one narrow investigation into model-agnostic properties, not a complete safety solution.

Organization. Section 2 discusses related work. Section 3 introduces notation and oracle access. Section 4 conveys intuition by analyzing a simple but inefficient "median distribution." Section 5 defines our algorithm and analyzes efficiency. Section 6 analyzes robustness to adversarial models. Section 7 analyzes abstention rates. Section 8 concludes with limitations and future work. In the appendix, Appendix A discusses constituent model selection and ideas for increasing overlap. Appendix B establishes optimality within our framework. Appendix D extends analysis to varying prompts.

2 Related Work

This paper is related to prior work in information-theoretic safety, ensemble methods, hardness results for detection problems, and broader AI safety approaches.

Information-theoretic foundations. Our approach is inspired by the elegant work of Vyas et al. [2023] which uses ensemble methods for provable copyright protection by training separate generative models on disjoint data partitions and sampling only from regions of agreement. Their CP-Δ algorithm is analogous to the special case $s = 1$ in our Alg. 1, and our consensus-robustness notion parallels their k-Near Access-Free definition with Renyi divergence of order infinity. However, their focus is the non-adversarial problem of copyright protection through training data partitioning, while we address a broader class of safety risks, including misalignment risk such as steganography and code vulnerabilities. Differential privacy [Dwork et al., 2006] provides another information-theoretic framework for safety, establishing bounded dependence on training data for privacy-preserving machine learning. It has been incorporated into generative models [Li et al., 2022]. Differential privacy is similar in spirit but technically different because, in our setting, a single adversarial model can drastically shift the distribution.

Ensemble methods and probability aggregation. The use of multiple models has a rich history in machine learning. Classical ensemble methods combine predictions to boost accuracy and reduce variance [Dietterich, 2000]. In the context of language models, several recent works leverage ensembles specifically for safety and reliability objectives. Multiple prior works combine language model probabilities to improve safety and reliability, though with different goals and guarantees than our approach. DExperts [Liu et al., 2021] uses a combination-of-experts at decoding time to reduce toxic content, while self-consistency methods [Wang et al., 2022] sample multiple reasoning paths and select the most consistent answer to reduce logical errors. These methods illustrate the power of using multiple models or prompts in tandem to cross-check and filter generations for safety and fidelity. However, this prior work does not address risks such as steganography or code vulnerabilities, which are difficult to detect and may be particularly concerning for superintelligent AI.

Fundamental hardness results. One motivation for our work comes from computational impossibility results that limit inspection-based safety approaches. Cohen [1987] showed that detecting security vulnerabilities such as viruses is computationally intractable.

[2]For $s = 1$, safety bounds still hold, but an adversary can force abstention.

More recently, Zamir [2024], de Witt et al. [2023] proved that steganography detection is similarly intractable, demonstrating an efficient language model scheme for encoding arbitrary messages indistinguishably from ordinary outputs. Perry et al. [2025], Motwani et al. [2024] show that hidden messages can be encoded such that removing them necessarily alters semantic meaning—even with unlimited computational resources.

These hardness results have profound implications: no human or AI system, regardless of capability, can reliably detect or remove such embedded information through output inspection alone. This suggests that inspection-based safety approaches may have certain limitations and motivates using output probabilities. In practice, while methods like paraphrasing [Krishna et al., 2023] can often destroy hidden messages, advanced steganographic techniques have emerged which survive such interventions [Mathew et al., 2024]. Our approach circumvents these hardness results by aggregating probabilities rather than inspecting outputs.

Broader AI safety landscape. Our work complements but differs fundamentally from mainstream AI safety approaches [see, e.g., Anwar et al., 2024]. Supervised oversight methods [Bowman et al., 2022] typically involve one model evaluating another's outputs, but the hardness results above suggest this may be insufficient for provably undetectable risks. AI debate frameworks [Irving et al., 2018] decompose complex claims into verifiable atomic steps, but similarly it is not clear how a claim like "this image contains no secret messages" would fit into the debate framework. Mechanistic interpretability [Bereska and Gavves, 2024] aims to understand model internals but generally lack formal guarantees that the interpretation is correct. Nonetheless, we emphasize that our method provides no training methodology for safe models. It must therefore be integrated with other safety training approaches rather than replacing them.

Fault-tolerant computing and cryptographic perspectives. Our framework draws inspiration from fault-tolerant distributed systems, where consensus protocols achieve robustness despite adversarial participants. The Byzantine generals problem [Lamport et al., 1982] establishes that systems can tolerate up to one-third adversarial participants. However, our consensus sampling problem is more forgiving: we need not achieve agreement on the models' distributions themselves, but merely draw samples likely to be safe given distributional overlap. Recent work applies cryptographic insights to AI safety [Goldwasser et al., 2024, Ball et al., 2025, Zamir, 2024], though many such results are negative impossibility findings. Our contribution differs by providing a constructive algorithm with positive information-theoretic guarantees.

3 Notation and Oracle Access Assumption

Throughout this paper we assume that natural numbers $s \le k$ are fixed constants representing the number of "safe" and total models, with $a := k - s$ representing the number of arbitrary, potentially adversarial, models. For this paper (other than Appendix D), we consider the prompt x fixed, so each model induces a distribution $p_i(y)$ over outputs. We also assume a fixed (finite or countably infinite) set Y of outputs.

Notation. All logs are base 2. Let $\mathsf{Distr}(Y)$ denote the set of probability distributions over Y. For distribution $p \in \mathsf{Distr}(Y)$, $p(U) := \sum_{y \in U} p(y)$ denotes the probability of a set $U \subseteq Y$. For a k-tuple of distributions over Y, which we denote by $\mathbf{p} = (p_1, \ldots, p_k) \in \mathsf{Distr}(Y)^k$, let $p_{(i)}(y)$ denote the ith smallest of the probabilities $\langle p_j(y) \rangle_{j \in [k]}$, and we also write $p_{(i)}(U)$ for a set $U \subseteq Y$ to denote the ith smallest probability of the set U, i.e.:

$$p_{(1)}(y) \le p_{(2)}(y) \le \ldots \le p_{(k)}(y) \quad \text{and} \quad p_{(1)}(U) \le p_{(2)}(U) \le \ldots \le \ldots p_{(k)}(U). \tag{1}$$

These *order statistics* generalize $p_{(1)}(y) = \min_i p_i(y)$ $p_{(k)}(y) = \max_i p_i(y)$. Also, let $[k] := \{1, 2, \ldots, k\}$. Finally, we assume that there is a special symbol $\perp \notin Y$ that our algorithm may return to indicate abstaining from generating a response.

Distribution access assumption. Modern large language models instantiate conditional distributions $p(y)$ over outputs y (given a prompt, as discussed in Appendix D). Their autoregressive factorization $p(y) = \prod_i p(y_i \mid y_1, y_2, \ldots, y_{i-1})$ means that probabilities can be computed token-by-token, and this can be done for any temperature value (by scaling log-probabilities). This in turn supports our oracle assumptions: one can efficiently (i) draw samples and (ii) evaluate probabilities (equivalently, log-probabilities) of any candidate output. We thus assume that, with a single oracle call, we can sample from each p_i or compute $p_i(y)$ for any $y \in Y$. Next-token probabilities are easily verified to sum to 1, which gives a convenient way both to generate a sample from the distribution and to check that p is a valid probability distribution. This consistency is important, since an adversary able to misrepresent p could cause risks outside the scope of our analysis.

4 Warm-up: the median distribution

We begin by studying a key property of a *median distribution* which conveys much of the intuition behind our approach. The distribution is pointwise proportional to the median of the k model probabilities:

$$p_{\text{med}}(y) \propto \text{median}\big(p_1(y), p_2(y), \ldots, p_k(y)\big).$$

It is not clear how to efficiently sample from this distribution. However, if one could and if a majority of the distributions are "safe," then the theorem below bounds the probability of "unsafe" outputs from the median distribution, for any set of unsafe outputs U. The bound depends on the amount of "overlap" between the safe distributions. For any set $S \subseteq [k]$ of safe distributions, their overlap is

$$\Delta(S) \coloneqq \sum_y \min_{i \in S} p_i(y) \in [0, 1].$$

For identical distributions, $\Delta(S) = 1$. For $|S|$=2 distributions it is called their Weizmann overlap (and is 1 minus their total variation distance). The generalization to any number of distributions is standard [see, e.g., Eidous and Alsheyyab, 2025].

Theorem 4.1 (Median distribution safety)**.** *Let $S \subseteq [k]$ be of size $|S| = s > k/2$ with $\Delta(S) > 0$. For any set of "unsafe" outputs $U \subseteq Y$ and any $p_1, p_2, \ldots, p_k \in \mathsf{Distr}(Y)$,*

$$p_{\text{med}}(U) \leq \frac{1}{\Delta(S)} \sum_{i \in S} p_i(U).$$

In particular, if each safe distribution has $p_i(U) \leq \varepsilon$ chance of generating an unsafe output, then p_{med} has a $\leq \varepsilon s/\Delta(S)$ chance of generating unsafe outputs, which is small assuming there is sufficient safe overlap $\Delta(S) \gg \varepsilon s$. As we analyze later, this type of bound encompasses many kinds of unsafe[3] outputs.

As a preview, consider steganography where a secret message $m = f(y)$ may be encoded in the samples for some decoding function $f : Y \to \{0, 1\}^*$ and binary string m. For any distribution p, the probability of accurately decoding the message is $\Pr_{y \sim p}[f(y) = m] = p(U)$ where $U = f^{-1}(m)$. Thus the above theorem bounds the probability $p_{\text{med}}(U) \leq \varepsilon s/\Delta(S)$ of decoding in terms of the probability $\varepsilon = \max_{i \in S} p_i(U)$ that any safe distribution would accidentally[4] generate steganography, which is presumably minuscule.

Proof. Let $d(y) \coloneqq \text{med}\{p_1(y), \ldots, p_k(y)\}$ denote the median probability so that,

$$p_{\text{med}}(y) = \frac{d(y)}{D} \text{ for any } y \in Y \text{ and } D \coloneqq \sum_{y \in Y} d(y).$$

Fix $y \in Y$. Since $|S| > k/2$, it is not hard to see that:

$$\min_{i \in S} p_i(y) \leq d(y) \leq \max_{i \in S} p_i(y). \tag{2}$$

[3]Theorem 4.1 also applies to the complementary set of *safe* outputs: the median distribution will produce unsafe outputs if a majority of the distributions do. Such is the nature of consensus.

[4]For example, if the decoding function concatenates the seventeenth letter of each sentence, there may be an exponentially small chance ε that the safe distribution happens to choose text where these letters spell out the secret m.

Algorithm 1 Consensus sampling from k distributions, risk-competitive with the safest s

Inputs: Integers $k \geq s \geq 1$, $R \geq 0$, and distributions $p_1, p_2, \ldots, p_k \in \mathsf{Distr}(Y)$
Output: A sample $y \in Y$ or $\perp$

for $r = 1$ to R **do**
    Sample $y \sim \frac{1}{k}\sum_{i=1}^{k} p_i$ ▷ pick a random index i and sample from p_i
    With probability $\dfrac{\frac{1}{s}\sum_{i \leq s} p_{(i)}(y)}{\frac{1}{k}\sum_{i=1}^{k} p_i(y)}$ **return** y
end for
return $\perp$

Using $\max_{i \in S} p_i(y) \leq \sum_{i \in S} p_i(y)$:

$$p_{\text{med}}(U) = \frac{1}{D}\sum_{y \in U} d(y) \leq \frac{1}{D}\sum_{y \in U}\sum_{i \in S} p_i(y) = \frac{1}{D}\sum_{i \in S} p_i(U). \tag{3}$$

It remains to show that $\Delta(S) \leq D$. Summing the left hand side of Eq. (2) over y,

$$\Delta(S) = \sum_{y \in Y} \min_{i \in S} p_i(y) \leq \sum_{y \in Y} d(y) = D.$$

(It also follows that p_{med} is well-defined because $D \geq \Delta(S) > 0$.) $\square$

There are many weaknesses of the median distribution which are addressed by our efficient algorithm. First, there is no efficient way to sample from the median distribution. Second, one cannot compute $\Delta(S)$ without knowledge of S. This makes it hard to know whether there is high or low overlap (for the given prompt x) and thus strong or weak safety. Thus, it could only be used safely if one has some reason to believe that there is high overlap among the safe distributions. Instead, our algorithm takes a safety parameter as an input and enforces a safety bound independent of $\Delta(S)$ by *abstaining* on prompts where it cannot guarantee that safety parameter (though abstention rates depend on $\Delta(S)$). Finally, the safe majority condition is removed.

5 Consensus Sampling

In the consensus sampling problem, we are given multiple generative models from which we can sample and compute probabilities, and we must produce an output or $\perp$ indicating abstention. Alg. 1 presents an efficient sampling algorithm that also takes a lower bound s on the number of safe distributions among them, and a risk-tolerance parameter $R \geq 0$. The algorithm generates a sample from a random one of the k distributions and decides whether or not to accept it or try again. The output is accepted and returned with a certain probability, computed as follows. First, it computes and sorts the k probabilities that the constituent distributions assign to the candidate sample. Because y is drawn from the mixture $\sum_i p_i(y)/k$, the denominator on the acceptance probability is positive almost surely, and the acceptance probability is clearly at most 1 because it is the *average of the s smallest of the probabilities* divided by *the average probability over all k probabilities.* If it is not accepted, this process is repeated for up to R independent attempts. If none are accepted, the algorithm outputs $\perp$ indicating abstention.

There is a tradeoff between abstention and safety, as the algorithm which always abstains is considered safe by definition. A smaller s would yield safer outputs, where one can view $s = 0$ as always abstaining and $s = k$ as never abstaining and simply outputting a random sample from a random p_i. Similarly, the parameter R represents a risk tolerance, with smaller R yielding safer outputs but higher abstention rate: $R = 0$ would correspond to always abstaining and $R = \infty$ to never abstaining. Appendix D analyzes the straightforward extension to include the prompt as an input.

We will analyze its efficiency, safety, and abstention rate. For efficiency, the algorithm is pseudopolynomial time in R.

Lemma 5.1 (Efficiency)**.** *For any constant k, Alg. 1 can be computed using $O(R)$ oracle calls and additional arithmetic operations.*

Proof. To draw a sample y in line 2, choose $i \in [k]$ uniformly at random and then draw a single sample $y \sim p_i$. To compute the probability ratio (line 3) then requires k evaluations of probabilities $p_i(y)$, one for each index $i \in [k]$. A constant number of additional operations per step are required, since k is taken to be a constant. □

To better understand and analyze Alg. 1, we note that it uses rejection sampling to sample from the following distribution q^*, stopping after at most R iterations.

$$q^*(y) \coloneqq \frac{1}{Z} \cdot \frac{1}{s} \sum_{i=1}^{s} p_{(i)}(y) \quad \text{for} \quad Z \coloneqq \sum_{y} \frac{1}{s} \sum_{i=1}^{s} p_{(i)}(y). \tag{4}$$

Lemma 5.2. *For any $\mathbf{p} = (p_1, p_2, \ldots, p_k) \in \mathsf{Distr}(Y)^k$, $R \geq 0$, and for q^*, Z from Eq. (4), the output distribution q of Alg. 1 satisfies:*

$$q(\perp) = (1 - Z)^R \quad \textit{and} \quad q(y) = \big(1 - q(\perp)\big)\, q^*(y) \textit{ for all } y \in Y. \tag{5}$$

Thus, ordinary rejection sampling (essentially $R = \infty$ if $Z > 0$) gives $q(\perp) = 0$ and an output distribution of q^*.

Proof. The lemma trivially holds for $R = 0$, because $q(\perp) = 1$. For shorthand, write

$$f(y) \coloneqq \frac{1}{k} \sum_{i=1}^{k} p_i(y), \qquad g(y) \coloneqq \frac{1}{s} \sum_{i=1}^{s} p_{(i)}(y), \qquad \alpha(y) \coloneqq \frac{g(y)}{f(y)} \tag{6}$$

so $y \sim f$ in line 2 and the acceptance probability in line 3 is $\alpha(y)$. Fix $y \in Y$. For $R = 1$, the probability of outputting y is

$$f(y) \cdot \alpha(y) = g(y) = Zq^*(y) = (1 - q(\perp))q^*(y)$$

and hence $q(\perp) = 1 - Z$, which establishes the lemma for $R = 1$. For $R > 1$, □

We note that the natural generalization of the median distribution to arbitrary s, the distribution that is pointwise proportional to $p_{(s)}(y)$, could be used in place of $\sum_{i \leq s} p_{(i)}(y)$ in Alg. 1 with the same bounds in the paper. However, we will show that q^* and *Alg.* 1 are optimal in terms of safety and abstention properties.

6 Safety properties

In this section we discuss safety properties relating to a set of unsafe outputs $U \subseteq Y$, with $y \notin U$ considered safe. When U is clear from context, we refer to the probability that a generation from a distribution p is unsafe, $p(U)$, as the *risk* of p. This general set-based view of safe outputs enables us to analyze Alg. 1 without having to pick a single definition of safety, such as malicious introduction of code. We first introduce a property called R-robustness where the parameter $R \geq 0$ captures the excess multiplicative risk incurred by using a distribution q, which may abstain, compared to the "safest" s distributions (those with smallest risk). Intuitively, we compare the risk $q(U)$ to the s safest distributions. We will then prove that Alg. 1 satisfies this definition and relate robustness to steganography.

Recall that the order statistic $p_{(s)}(U)$ denotes the sth smallest of the risks $p_1(U), \ldots, p_k(U)$, as defined by Eq. (1). The following definition of R-robustness requires being competitive with $p_{(s)}(U)$ for *all possible unsafe sets U*. It is slightly stronger than the simpler $q(U) \leq R \cdot p_{(s)}(U)$ requirement discussed in the introduction, because $p_{(s)}(U) \geq \sum_{i \leq s} p_{(i)}(U)/s$:

Definition 6.1 (Consensus robustness)**.** *For $R \geq 0$, $q \in \mathsf{Distr}(Y \cup \{\perp\})$ is consensus-robust with parameter R, or just R-robust, relative to $\mathbf{p} = (p_1, p_2, \ldots, p_k) \in \mathsf{Distr}(Y)^k$ if*

$$q(U) \leq R \cdot \frac{1}{s} \sum_{i=1}^{s} p_{(i)}(U) \textit{ for all } U \subseteq Y.$$

Smaller R indicates less risk. Because abstaining is viewed as safe, the distribution that always abstains is 0-robust. We use the same variable R in this definition because our upper bound on R exactly matches the input R to Alg. 1.

The following theorem states that the risk of Alg. 1 is at most the average risk of the safest s distributions.

Theorem 6.2 (Consensus robustness)**.** *For any $R \geq 0$ and $\mathbf{p} = (p_1, p_2, \ldots, p_k) \in \mathsf{Distr}(Y)^k$, the output distribution of Algorithm 1 is R-robust relative to $\mathbf{p}$.*

Proof. For any $y \in Y$, by Lemma 5.2, since $(1 - Z)^R \geq 1 - RZ$,

$$q(U) = \sum_{y \in U} q(y) \leq \sum_{y \in U} RZq^*(y) = \sum_{y \in U} R \cdot \frac{1}{s} \sum_{i \leq s} p_{(i)}(y) \leq R \cdot \frac{1}{s} \sum_{i \leq s} p_{(i)}(U).$$

Since $p_{(i)}(U) \geq \sum_{y \in U} p_{(i)}(y)$, the above gives $q(U) \leq \sum_{i \leq s} p_{(i)}(U)$. □

We note that Theorem 6.2 has immediate adversarial implications when there is a set $S \subseteq [k]$ of "safe" distributions low risk, while the other distributions can be adversarially controlled.

Corollary 6.3 (Adversarial robustness)**.** *Let $U \subseteq Y$, $\mathbf{p} \in \mathsf{Distr}(Y)^k$, $S \subseteq [k]$ of size $|S| = s$, with mean risk $\varepsilon := \sum_{i \in S} p_i(U)/s$. Then Alg. 1's output distribution has risk $\leq \varepsilon R$.*

Proof. The proof follows simply from $\frac{1}{s} \sum_{i \leq s} p_{(i)}(U) \leq \frac{1}{s} \sum_{i \in S} p_i(U)$. □

Backdoors and jailbreaks. Alg. 1 may be more useful at thwarting backdoors than jailbreaks in current LM's. To see why, note that as long as s of the models produce safe $y \notin U$ on a certain "backdoor" prompt x, then Alg. 1 produce will also produce a safe output. Thus, an adversary using techniques such as [Kurita et al., 2020, Hubinger et al., 2024] to plant a backdoor would have to plant the same backdoor x in $> a = k - s$ of the models. Unfortunately, current LMs are often susceptible to "universal" jailbreaks [Wallace et al., 2019] which work across many models, regardless of training platform. Therefore, for Alg. 1 to be useful in thwarting jailbreaks would require increased model safety from further safety training innovations.

In Appendix B, we show that q^* is $1/Z$-robust and in fact is optimal, being the only one that is $1/Z$-robust among all distributions that never abstain. This in turn, implies that our algorithm is Pareto optimal in terms of its robustness-abstention tradeoff.

6.1 Steganography and leakage

This section argues that R-robustness implies resilience to steganography and that Alg. 1 "leaks" few bits. There are many possible ways to quantify leakage, and consider three which are all bounded by $\log(R+1)$ bits. The first approach is to use simulation, a common approach in cryptography, to bound leakage. In particular, we show that any choice of the unsafe models can equivalently be simulated by a procedure whose only dependence on the unsafe models is through a single integer $1 \leq t \leq R+1$. Since such an integer can be written down using $\log(R+1)$ bits (technically $\lceil \log(R+1) \rceil$), this can be viewed as $\leq \log(R+1)$ bits of leakage.

Lemma 6.4. *[Alg. 1 leakage simulation bound] Fix any $R \geq 0$ and size-s $S \subseteq [k]$. There exists a randomized function $\mathcal{A} : Y^R \times \mathsf{Distr}(Y)^k \to [R+1]$ such that the following procedure has the same output distribution as Alg. 1, for any $\mathbf{p} = (p_1, \ldots, p_k) \in \mathsf{Distr}(Y)^k$:*

1: *Sample* $y_1, \ldots, y_R \stackrel{i.i.d.}{\sim} \frac{1}{s} \sum_{i \in S} p_i$
2: $t \leftarrow \mathcal{A}\big(y_1, \ldots, y_R, \mathbf{p}\big)$
3: ***if*** $t = R+1$ ***then return*** $\perp$ ***else return*** y_t

The point is that only dependence on the unsafe $\langle p_j \rangle_{j \notin S}$ is via t in the above procedure, whose output distribution is equivalent to Alg. 1. The R independent samples are computed solely from the uniform mixture over safe distributions. Thus, an adversary's control over the output can be boiled down to a single number $t \in [R+1]$, which we view as $\log(R+1)$ bits of leakage.

Proof. The algorithm $\mathcal{A}$ iterates through $t = 1, 2, \ldots, R$ and accepts y_t with probability,

$$\frac{\sum_{i \le s} p_{(i)}(y_t)}{\sum_{i \in S} p_i(y_t)} \le 1.$$

This is at most 1 because the numerator is summed over the smallest s probabilities and the denominator is only over the s safe ones. The per-round probability of outputting any $y \in Y$ is easily seen to be $\sum_{i \le s} p_{(i)}(y)/s$ exactly as shown for Alg. 1 in Lemma 5.2. □

We next argue that R-robustness, in general, implies resilience to steganography, as discussed briefly in Section 4. First we bound risk. Again, suppose there is a decoding function $f : Y \to \{0,1\}^*$ used to recover a secret message $m \in \{0,1\}^*$ encoded as, say, a binary string. Say the safe distributions have some low (presumably exponentially small) ε chance of generating a sample $y \sim p$ that happens to decode to the secret $m = f(y)$. The following shows that R-robustness implies $\le \varepsilon R$ risk due to steganography.

Observation 6.5. *Let $R \ge 0$ and q be R-robust relative to $p_1, p_2, \ldots, p_k$. Fix $m \in \{0,1\}^*$ and a decoding function $f : Y \to \{0,1\}^*$. Let $\varepsilon \in [0,1]$ and S be a set of size s such that $\Pr_{y \sim p_i}[f(y) = m] \le \varepsilon$ for all $i \in S$. Then,*

$$\Pr_{y \sim q}[f(y) = m] \le \varepsilon R.$$

This holds for Alg. 1, since Theorem 6.2 shows is R-robust.

Proof. Let $U = f^{-1}(m)$, so $p_i(U) = \Pr_{y \sim p_i}[f(y) = m] \le \varepsilon$ for all $i \in S$. By Definition 6.1,

$$q(U) \le R \cdot \frac{1}{s} \sum_{i=1}^{s} p_{(i)}(U) \le R \cdot \frac{1}{s} \sum_{i \in S} p_i(U) \le R \cdot \varepsilon.$$

□

This shows that the success rate of a steganographic attack is low, but we can go further and provide an upper bound on the number of bits leaked by our algorithm using *mutual information*, commonly used to quantify average information leakage between two random variables. It requires additional notation and definitions to apply to our setting. Appendix C proves that any R-robust distribution q leaks at most $O(\log R)$ bits in the sense that the mutual information between a sample from q and a random secret message to be leaked is $\le \log(R+1)$. It is extended to the case of prompts using "maximal leakage" [Issa et al., 2020] in Appendix D.

7 Abstention bounds

Lemma 5.2 shows that the algorithm's abstention rate is exactly $q(\perp) = (1-Z)^R$ but the quantity $Z \coloneqq \sum_y \sum_{i \le s} p_{(i)}(y)/s$, could be minuscule or zero (always abstaining is safe but useless). We now lower-bound Z even in the presence of an adversary that controls a minority $< k/2$ of the distributions. It relies on overlap, which recall is $\Delta(S) \coloneqq \sum_{y \in Y} \min_{i \in S} p_i(S)$.

We will be able to lower-bound the overlap with a safe majority, i.e., $s > k/2$. For instance, for $k = 3$ distributions, if the $s = 2$ safe distributions overlap by $\Delta(S) \ge 0.2$, then the analysis below shows that the algorithm abstains with exponentially low probability $\le 0.9^R$. However, if the distributions rarely agree on likely outcomes, then our aggregation protocol cannot hope to robustly output a common sample with significant probability.

While $Z \ge \Delta(S)/s$ follows from Theorem 7.2 below, the theorem gives a tighter bound $Z \ge \Delta_a(S)/s$ where the *maximal overlap* $\Delta_a(S) \ge \Delta(S)$, defined as follows:

Definition 7.1 (c-overlap)**.** *For an index set $I \subseteq [k]$ and $c \in \{0, 1, 2, \ldots, |I| - 1\}$:*

$$\Delta_c(I) \coloneqq \max_{\substack{J \subseteq I \\ |J| = c+1}} \Delta(J).$$

It is easy to see that $0 \leq \Delta(I) \leq \Delta_c(I) \leq 1$ because $J \subseteq I \Rightarrow \Delta(J) \geq \Delta(I)$.

Theorem 7.2 (Abstention)**.** *Assume $s > k/2$ and fix any $R \geq 1$ and $\mathbf{p} \in \mathsf{Distr}(Y)^k$. For the output y of Alg. 1,*

$$\Pr[y = \perp] \leq \min_{S \subseteq [k] : |S| = s} \left(1 - \frac{\Delta_a(S)}{s}\right)^R.$$

Recall that $a = k - s$. The safe-majority condition $s > k/2$ ensures that any set of $a + 1 = k - s + 1$ indices must contain at least one of the s smallest probabilities, which is the crux of the bound.

Proof. It suffices to show the bound for any fixed size-s S. Fix $S^\star \subseteq S$ of size $|S^\star| = a + 1$ that attains $\Delta(S^\star) = \Delta_a(S)$. Next, we claim that $p_{(s)}(y) \geq \min_{i \in S^\star} p_i(y)$ for any $y \in Y$ because $p_{(s)}(y)$ is the sth smallest probability, and $S^\star$ has all but $s - 1$ elements of $[k]$ and thus must contain one index i such that $p_i(y) \leq p_{(s)}(y)$. (This is where $s > k/2$ is used.)

By Lemma 5.2, $\Pr[y = \perp] = (1 - Z)^R$ for

$$Z = \sum_{y \in Y} \frac{1}{s} \sum_{i \leq s} p_{(i)}(y) \geq \sum_{y \in Y} \frac{1}{s} p_{(s)}(y) \geq \frac{1}{s} \sum_{y} \min_{i \in S^\star} p_i(y) = \frac{\Delta(S^\star)}{s} = \frac{\Delta_a(S)}{s}.$$

□

For $a = 1$, the bound is $Z \geq \Delta_1(S)/s$ where $\Delta_1(S)$ is the maximum overlap between any two safe distributions. Unfortunately, as discussed before, this quantity may be exponentially small in the output length for language models that were not designed with overlap in mind. Appendix A offers suggestions for ways to increase overlap and reduce abstention.

8 Conclusions, Limitations, and Future Work

Provable models of AI safety may inform the design of beneficial AI systems [Russell, 2022]. We introduce a black-box framework for aggregating generative models with provable robustness guarantees. Assuming some of the generative models being aggregated are safe, our analysis shows that R-robustness simultaneously bounds unsafe amplification and information leakage, and that an efficient sampling procedure can realize these bounds with controlled abstention. More broadly, these results open the door to a theoretical treatment of "safety" in model outputs: not as a question of interpretability or semantics, but as a property of aggregating a collection of distributions while limiting the influence of any small subset. This perspective yields formal guarantees that complement and strengthen empirical oversight methods.

Limitations. Our work does not provide an end-to-end safety guarantee for several reasons. First, assuming that at least s of the k distributions are truly safe is neither trivial nor inherently justified. Second, repeated sampling from our algorithm may lead to an accumulation of leakage or unsafe responses. A third central limitation is our reliance on multiple models and overlap: guaranteeing both safety and non-abstention requires a majority of safe models that overlap. Finally, many risks, such as societal harms measured by distributional disparities [Kearns and Roth, 2019], cannot be modeled as a set of unsafe outputs. Addressing these additional types of risks is necessary before safe deployment.

Future directions. This work introduces consensus sampling as a new direction for adding a layer of safety to generative AI models. There are many open questions such as how to extend it to multiple (possibly agentic) interactions. Due to its black-box statistical nature, it can apply to future models that may operate in ways quite different from today's models, as long as they exhibit sufficient overlap. A promising future direction is to *engineer* greater overlap (see Appendix A). In some sense, this quest for overlap is one of identifying a *canonical distribution* which the models can be steered toward.The relationship between canonical distributions and safety may merit further exploration beyond consensus sampling.

Acknowledgments. We are grateful to Cary Bassin, Mia Glase, and Min Jae Song for helpful comments. OZ's research is supported in part by the Israel Science Foundation, Grant No. 1593/24, and in part by the Blavatnik foundation.

References

Usman Anwar, Abulhair Saparov, Javier Rando, Daniel Paleka, Miles Turpin, Peter Hase, Ekdeep Singh Lubana, Erik Jenner, Stephen Casper, Oliver Sourbut, Benjamin L. Edelman, Zhaowei Zhang, Mario Günther, Anton Korinek, Jose Hernandez-Orallo, Lewis Hammond, Eric J Bigelow, Alexander Pan, Lauro Langosco, Tomasz Korbak, Heidi Chenyu Zhang, Ruiqi Zhong, Sean O hEigeartaigh, Gabriel Recchia, Giulio Corsi, Alan Chan, Markus Anderljung, Lilian Edwards, Aleksandar Petrov, Christian Schroeder de Witt, Sumeet Ramesh Motwani, Yoshua Bengio, Danqi Chen, Philip Torr, Samuel Albanie, Tegan Maharaj, Jakob Nicolaus Foerster, Florian Tramèr, He He, Atoosa Kasirzadeh, Yejin Choi, and David Krueger. Foundational challenges in assuring alignment and safety of large language models. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=oVTkOs8Pka.

Sarah Ball, Greg Gluch, Shafi Goldwasser, Frauke Kreuter, Omer Reingold, and Guy N. Rothblum. On the impossibility of separating intelligence from judgment: The computational intractability of filtering for ai alignment, 2025. URL https://arxiv.org/abs/2507.07341.

Leonard Bereska and Stratis Gavves. Mechanistic interpretability for AI safety: A review. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=ePUVetPKu6.

Samuel R. Bowman, Jeeyoon Hyun, Ethan Perez, et al. Measuring progress on scalable oversight for large language models. *arXiv preprint arXiv:2211.03540*, 2022. doi: 10.48550/arXiv.2211.03540. URL https://arxiv.org/abs/2211.03540.

Fred Cohen. Computer viruses: theory and experiments. *Computers & security*, 6(1):22–35, 1987.

Christian Schroeder de Witt, Samuel Sokota, J Zico Kolter, Jakob Nicolaus Foerster, and Martin Strohmeier. Perfectly secure steganography using minimum entropy coupling. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=HQ67mj5rJdR.

Thomas G. Dietterich. Ensemble methods in machine learning. In Josef Kittler and Fabio Roli, editors, *Multiple Classifier Systems*, volume 1857 of *Lecture Notes in Computer Science*, pages 1–15, Berlin, Heidelberg, 2000. Springer. doi: 10.1007/3-540-45014-9_1.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Proceedings of the Third Theory of Cryptography Conference (TCC)*, pages 265–284. Springer, 2006. doi: 10.1007/11681878_14.

Omar M. Eidous and Majd M. Alsheyyab. Generalization of the overlapping coefficient for k
*ge*2 normal distributions. *Math. Model. Nat. Phenom.*, 20:9, 2025. doi: 10.1051/mmnp/2025008. URL https://doi.org/10.1051/mmnp/2025008.

Shafi Goldwasser, Michael P. Kim, Vinod Vaikuntanathan, and Or Zamir. Planting undetectable backdoors in machine learning models. *arXiv preprint arXiv:2204.06974*, 2024.

Evan Hubinger, Carson Denison, Jesse Mu, Mike Lambert, Meg Tong, Monte MacDiarmid, Tamera Lanham, Daniel M. Ziegler, Tim Maxwell, Newton Cheng, Adam Jermyn, Amanda Askell, Ansh Radhakrishnan, Cem Anil, David Duvenaud, Deep Ganguli, Fazl Barez, Jack Clark, Kamal Ndousse, Kshitij Sachan, Michael Sellitto, Mrinank Sharma, Nova DasSarma, Roger Grosse, Shauna Kravec, Yuntao Bai, Zachary Witten, Marina Favaro, Jan Brauner, Holden Karnofsky, Paul Christiano, Samuel R. Bowman, Logan Graham, Jared Kaplan, Sören Mindermann, Ryan Greenblatt, Buck Shlegeris, Nicholas Schiefer, and Ethan Perez. Sleeper agents: Training deceptive llms that persist through safety training, 2024. URL `https://arxiv.org/abs/2401.05566`.

Geoffrey Irving, Paul Christiano, and Dario Amodei. AI safety via debate. *arXiv preprint arXiv:1805.00899*, 2018. URL `https://arxiv.org/abs/1805.00899`.

Ibrahim Issa, Aaron B. Wagner, and Sudeep Kamath. An operational approach to information leakage. *IEEE Transactions on Information Theory*, 66(3):1625–1657, 2020. doi: 10.1109/TIT.2019.2962804. URL `https://arxiv.org/abs/1807.07878`.

Michael Kearns and Aaron Roth. *The Ethical Algorithm: The Science of Socially Aware Algorithm Design*. Oxford University Press, New York, NY, 2019.

Kalpesh Krishna, Yixiao Song, Marzena Karpinska, John Wieting, and Mohit Iyyer. Paraphrasing evades detectors of ai-generated text, but retrieval is an effective defense. *arXiv preprint arXiv:2303.13408*, 2023.

Keita Kurita, Paul Michel, and Graham Neubig. Weight poisoning attacks on pretrained models. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2793–2806, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.249. URL `https://aclanthology.org/2020.acl-main.249/`.

Leslie Lamport, Robert Shostak, and Marshall Pease. The byzantine generals problem. *ACM Transactions on Programming Languages and Systems*, 4(3):382–401, 1982. doi: 10.1145/357172.357176.

Xuechen Li, Florian Tramer, Percy Liang, and Tatsunori Hashimoto. Large language models can be strong differentially private learners. In *International Conference on Learning Representations (ICLR)*, 2022. URL `https://openreview.net/forum?id=bVuP3ltATMz`.

Alisa Liu, Maarten Sap, Ximing Lu, Swabha Swayamdipta, Chandra Bhagavatula, Noah A. Smith, and Yejin Choi. DExperts: Decoding-time controlled text generation with experts and anti-experts. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 6691–6706, 2021.

Yohan Mathew, Ollie Matthews, Robert McCarthy, Joan Velja, Christian Schroeder de Witt, Dylan Cope, and Nandi Schoots. Hidden in plain text: Emergence & mitigation of steganographic collusion in llms, 2024. URL `https://arxiv.org/abs/2410.03768`.

Sumeet Ramesh Motwani, Mikhail Baranchuk, Martin Strohmeier, Vijay Bolina, Philip Torr, Lewis Hammond, and Christian Schroeder de Witt. Secret collusion among AI agents: Multi-agent deception via steganography. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL `https://openreview.net/forum?id=bnNSQhZJ88`.

OpenAI. Addendum to GPT-4o system card: Native image generation, 2025. URL `https://cdn.openai.com/11998be9-5319-4302-bfbf-1167e093f1fb/Native_Image_Generation_System_Card.pdf`. Accessed 2025-10-05.

Neil Perry, Sanket Gupte, Nishant Pitta, and Lior Rotem. Robust steganography from large language models, 2025. URL `https://arxiv.org/abs/2504.08977`.

Ronald Rosenfeld. A maximum entropy approach to adaptive statistical language modeling. *Computer Speech and Language*, 10(3):187–228, 1996. URL https://www.cs.cmu.edu/~roni/papers/me-csl-revised.pdf.

Stuart Russell. Provably beneficial artificial intelligence. In *Proceedings of the 27th International Conference on Intelligent User Interfaces (IUI '22)*, New York, NY, USA, 2022. Association for Computing Machinery. doi: 10.1145/3490099.3519388. URL https://doi.org/10.1145/3490099.3519388. Keynote abstract.

Dan Shumow and Niels Ferguson. On the possibility of a back door in the nist sp800-90 dual ec prng. Presentation at CRYPTO 2007 rump session, 2007. URL https://rump2007.cr.yp.to/15-shumow.pdf.

Nikhil Vyas, Sham M Kakade, and Boaz Barak. On provable copyright protection for generative models. In *International conference on machine learning*, pages 35277–35299. PMLR, 2023.

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

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022. doi: 10.48550/arXiv.2203.11171. URL https://arxiv.org/abs/2203.11171.

Wikipedia contributors. Dual_EC_DRBG — Wikipedia, the free encyclopedia. https://en.wikipedia.org/wiki/Dual_EC_DRBG, 2024. Accessed: 2025-10-02.

Or Zamir. Undetectable steganography for language models. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=fq6aQoMSHz.

A Designing for overlap

We have said little about how k distributions can be trained to encourage both safety and overlap. Intuitively, models trained differently–by different groups or with different methodologies–are more likely to fail independently; such diversity reduces the chance of shared points of failure.

However, independently trained models may exhibit little overlap, especially for long generations—as found in experiments by Vyas et al. [2023] even when using identical training methodologies. At first it may seem intuitive that one would have exponentially small overlap for long generations, e.g., even if two LMs differ on which type of punctuation they use more often, these differences can accumulate over a long generation. However, a good LM may have adaptive generation in the sense that given that the first half of a generation uses a certain punctuation convention, it may be likely to continue that convention regardless of its likelihood of using it in the first place.

There are a variety of ways to achieve varied models that still have overlap. For an LM, for instance, one could vary its system instructions to try and thwart various types of harmful generations. A more general methodology, however, is to train or compose models that aim for some *canonical* output distribution for each prompt. This may leverage a small not-so-capable model, p_{small}, which we believe is unlikely to inject hard-to-detect harmful outputs such as steganography.

- Use different large models to generate short "hints" which can then be provided to p_{small}.

- In addition to generating hints, use the large models as "gatekeepers" at the end of the generation to determine whether p_{small}'s output is valid.
- Alternatively, the canonical distribution could be $p_{\text{small}}(\cdot \mid y \in V)$ conditioned on the output y belonging to some set V of good responses. For example, if the prompt is to prove a mathematical theorem, the set V may be the set of valid proofs, which p_{small} would be unlikely to generate on its own. This is a bit similar in spirit to maximum-entropy LMs [Rosenfeld, 1996] but where the constraints are based on a complex validity criterion.
- Finally, LMs could be trained with an objective of staying close to p_{small} but with other validity objectives.

The requirements for a generative AI oracle to be used with our algorithms is: (a) the ability to sample outputs, conditioned on a prompt; and (b) the ability to compute the conditional probability of any generation $y \in Y$ given a prompt. However, one can extend the framework to allow the sampling oracle to fail and output $\perp$ as well, in which case the algorithm will necessarily resample. This does not impact the safety or abstention guarantees as long as overlap is only computed over $y \in Y$ (and not $\perp$). Also, note that the oracle does not need to be able to compute its abstention probability $p(\perp)$.

We note that identifying canonical distributions and using them for safety is an interesting future direction of work.

Finally, one could of course run the algorithm with a significantly larger number of iterations R, say scaling up by a factor of 2^L, under the logic that the worst-case bounds in this paper are pessimistic, and in practice an adversary who does not have access to the other models cannot leak many bits. This is similar to how differential privacy [Dwork et al., 2006] is often found to be privacy preserving with a much larger tolerance ε than justified by the worst-case theoretical bounds. Analogous to what Vyas et al. [2023] do in the context of copyright protection, one can augment Alg. 1 with additional "slack" parameter $L \geq 0$ and changing the acceptance probability in line 3 to:

$$\min\left\{1, 2^L \frac{\frac{1}{s}\sum_{i\leq s} p_{(i)}(y)}{\frac{1}{k}\sum_{i=1}^{k} p_i(y)}\right\}.$$

This variation is $R' = 2^L R$-robust and thus leaks an additional L bits, but abstains exponentially less often. The justification for such a large robustness parameter is that, if overlap is extremely challenging to achieve via canonical distributions, then an adversary who does not have knowledge of the safe distributions may inherently have low overlap. This would be substantially faster than just running the algorithm for R' iterations.

B Optimality

In this section, we show that Alg. 1 is optimal in terms of robustness. Recall that we have fixed $s \leq k$. However, the algorithm still has a parameter $R = 0, 1, 2, \ldots$ and the abstention rate decreases in R while the robustness increases in R. Specifically, we prove that for each parameter setting R the algorithm trades off abstention rate for worst-case robustness in a Pareto optimal fashion.

The high-level structure of the argument is quite simple. Recall from Lemma 5.2 that, for any R, the output distribution of the algorithm q is either $\perp$ with probability $q(\perp) = (1-Z)^R$ and y with probability $q(y) = (1 - q(\perp))q^*(y)$. Thus, it is simply a downscaled version of q^* from Eq. (4).

Moreover, we show a linear relationship between (non)abstention and robustness, so one can simply trade off abstention for robustness. Finally, we argue that the distribution q^* is the optimal distribution in terms of worst-case robustness over distributions that do not abstain. This means that each input $R = 0, 1, 2, \ldots$ (and $R = \infty$ corresponding to never abstaining) is Pareto optimal in terms of their abstention and worst-case robustness rates.

In this section, for optimality, we define the smallest value of $R \geq 0$ such that q is R-robust with respect to $\mathbf{p}$, and also the worst-case value of this over all S. To do so, it is helpful to define the average distribution $\overline{p}_S \in \mathsf{Distr}(Y)$ with respect to any nonempty set of indices $S \subseteq [k]$:

$$\overline{p}_S(y) \coloneqq \frac{1}{|S|}\sum_{i \in S} p_i(y) \text{ for all } y \in Y \tag{7}$$

Definition B.1. *For distribution q over $Y \cup \{\perp\}$ and distributions $p_1, p_2, \ldots, p_k$, over Y,*

$$R_q \coloneqq \max_{y \in Y} \max_{S \subseteq [k]:|S|=s} \frac{q(y)}{\overline{p}_S(y)},$$

with $0/0$ taken to be 0 and $R_q = +\infty$ if $q(y) > 0 = \overline{p}_S(y)$ for some $y \in Y$ and size-s S.

It is easy to see that q is R'-robust relative to $p_1, \ldots, p_k$ iff $R' \geq R_q$. R_q is thus the *worst-case* robustness of distribution q with respect to any safe subset. This definition suggests a notion of optimality, for any given $p_1, p_2, \ldots, p_k$, meaning that the algorithm should be optimally robust against the worst-case subset of safe distributions.

First, we show that q^* is optimal among distributions that never abstain, $\mathsf{Distr}(Y)$.

Lemma B.2 (Minmax robustness)**.** *For any $\mathbf{p} = (p_1, p_2, \ldots, p_k) \in \mathsf{Distr}(Y)^k$,*

$$\min_{q \in \mathsf{Distr}(Y)} R_q = \frac{1}{Z} \text{ for } Z \coloneqq \sum_y \frac{1}{s}\sum_{i=1}^{s} p_{(i)}(y),$$

with q^* from Eq. (4) being the unique minimizer if $Z \neq 0$.

Proof. Fix $\mathbf{p}$ and hence Z. First consider the case that $Z = 0$ and let $q \in \mathsf{Distr}(Y)$ be arbitrary. We must show that $R_q = \infty$ for some size-s S. Let $y \in Y$ be any output with positive probability $q(y) > 0$. Since $Z = 0$, this means that $\sum_{i \leq s} p_{(i)}(y) = 0$ so there are s indices i such that $p_i(y) = 0$. Let S be this set, then $\overline{p}_S(y) = 0$ and hence $R_q = +\infty$.

Next, consider the case where $Z > 0$. We first claim that, $R_{q^*} \leq 1/Z$, equivalently that $q^*(y) \leq \overline{p}_S(y)/Z$ for every $|S| = s$. To see this, note that,

$$q^*(y) = \frac{1}{Z} \cdot \frac{1}{s}\sum_{i=1}^{s} p_{(i)}(y) \leq \frac{1}{Z} \cdot \frac{1}{s}\sum_{i \in S} p_i(y) = \frac{\overline{p}_S(y)}{Z}.$$

Conversely, any distribution $q \neq q^*$ must satisfy, for some $y \in Y$, $q(y) > q^*(y)$. Fix such a y and consider the set S which consists of the s distributions with smallest values $p_i(y)$ (breaking ties arbitrarily). For this set S, we have,

$$q(y) > q^*(y) = \frac{1}{Z} \cdot \frac{1}{s}\sum_{i=1}^{s} p_{(i)}(y) = \frac{1}{Z}\overline{p}_S(y).$$

Thus, $R_q \geq R_{q,\overline{p}_S} > 1/Z$. □

For any distribution $q \in \mathsf{Distr}(Y \cup \{\perp\})$, it will be convenient in this section to use the term *output rate* for the probability $q(Y) = 1 - q(\perp)$ that a distribution $q \in \mathsf{Distr}(Y \cup \{\perp\})$ produces a non-abstaining output. There is a trivial linear relationship between $q(Y)$ and R-robustness which means that there is a single non-abstaining distribution q^* characterizes the set of Pareto-optimal frontier of distributions in terms of abstaining and robustness.

Any possibly abstaining distribution q can be decomposed into its output rate $q(Y)$ and its *conditional generation distribution* $q' \in \mathsf{Distr}(Y)$ where $q'(y) = q(y)/q(Y)$ for all $y \in Y$. Robustness trivially scales linearly with output rate:

Lemma B.3 (Robustness-abstention tradeoff)**.** *Let $q \in \mathsf{Distr}(Y \cup \{\perp\})$ and let $q' \in \mathsf{Distr}(Y)$ be such that $q(y) = q(Y) \cdot q'(y)$ for all $y \in Y$. Then $R_q = q(Y)R_{q'}$.*

Proof. By Definition B.1,

$$R_q = \max_{y \in Y} \max_{S \subseteq [k]:|S|=s} \frac{q(y)}{\overline{p}_S(y)} = \max_{y \in Y} \max_{S \subseteq [k]:|S|=s} \frac{q(Y) \cdot q'(y)}{\overline{p}_S(y)} = q(Y)R_{q'}.$$

□

By this linearity and Lemma B.2, it follows immediately that $q^* \in \mathsf{Distr}(Y)$ characterizes the optimal tradeoff between abstention rate and robustness: a distribution q is Pareto optimal (optimal abstention rate given robustness, and vice versa) if and only if $q(y) = q(Y)q^*(y)$ for all $y \in Y$.

Theorem B.4 (Optimality)**.** *Fix any* $\mathbf{p} = (p_1, \ldots, p_k) \in \mathsf{Distr}(Y)^k$ *and* $R \in \{0, 1, 2, \ldots\} \cup \{\infty\}$*. Let q be the (possibly abstaining) output distribution of Alg. 1 with parameter R. Then q is Pareto optimal: for every other (possibly abstaining) distribution $\nu \neq q$, either $q(\perp) < \nu(\perp)$ or $R_q < R_\nu$ (or $R_\nu = \infty$).*

Proof. Fix a distribution ν such that $\nu(\perp) \leq q(\perp)$ and $R_\nu \leq R_q$. We will show that $\nu = q$.

Let q^* and Z be as defined in Lemma B.2. If $Z = 0$, then Algorithm 1 always abstains, so $q(\perp) = 1$ and $R_q = 0$, any any ν with $R_\nu = 0$ must always abstain. Thus, we can assume $Z > 0$ and q^* is well-defined as in Eq. (4), and $R^*_q < \infty$. By the linearity of R (Lemma B.3) and the fact that q^* is the unique distribution with $q^*(Y) = 1$ and $R^*_q \leq 1/Z$ (Lemma B.2), it is easy to see that the only Pareto optimal distributions are μ which satisfy $\mu(y) = \mu(Y) \cdot q^*(y)$ for all $y \in Y$, i.e., the distributions μ which abstain with some probability $\mu(\perp)$ and with the remaining probability $\mu(Y) = 1 - \mu(\perp)$ output according to q^*. (To see this, just consider $\nu^* \in \mathsf{Distr}(Y)$ with $\forall y \in Y\ \nu^*(y) \coloneqq \nu(y)/\nu(Y)$, and note that if $\nu^* \neq q^*$ then $R_{\nu^*} > 1/Z$.) Finally, Lemma 5.2 shows exactly that: $q(y) = q(Y)q^*(y)$ for all $y \in Y$. ☐

C Mutual information leakage bound

In this section, we observe that R-robustness also implies a quantitative bound on information leakage: at most $\log(R + 1)$ bits can be leaked by an adversarial robust distribution, beyond what is already leaked by the safe distribution. This limits the impact of steganography or any attempt to smuggle information external to the safe distribution. Appendix D will consider the more general case where the prompt is not fixed.

It will be convenient to define a related notion of safety with respect to a *single safe distribution p*.

Definition C.1 (Relative risk)**.** *For $R \geq 0$, $q \in \mathsf{Distr}(Y \cup \{\perp\})$, and $p \in \mathsf{Distr}(Y)$, say q is R-risky relative to p if $q(y) \leq R \cdot p(y)$ for all $y \in Y$.*

If q is R-robust, then it follows directly that, for any size-s $S \subseteq [k]$, q is R-risky with respect to $\overline{p}_S$ which is defined in Eq. (7) by $\overline{p}_S(y) \coloneqq \frac{1}{|S|}\sum_{i \in S} p_i(y)$.

Observation C.2. *q is R-robust relative to $p_1, \ldots, p_k$ if and only if q is R-risky relative to $\overline{p}_S$ for every $S \subseteq [k]$ of size s.*

Proof. ($\Rightarrow$) Take $U = \{y\}$, so $q(y) \leq R \cdot \frac{1}{s}\sum_{i \leq s} p_{(i)}(y) \leq R\,\overline{p}_S(y)$ for any $|S| = s$. ($\Leftarrow$) Summing $q(y) \leq R\,\overline{p}_S(y)$ over $y \in U$ gives $q(U) \leq R\,\overline{p}_S(U)$ for all S; minimizing over S yields $q(U) \leq R \cdot \frac{1}{s}\sum_{i \leq s} p_{(i)}(U)$. ☐

Here we model leakage as follows: the safe distribution $p = \overline{p}_S$ is fixed while there is a distribution q_m which varies based on some message m to be encoded, which is drawn from a message distribution M. We show that q being R-risky implies that the mutual information between the drawn message M and the response sample $Q \sim q_M$ is bounded by $\log R + 1$. In particular, at most $\log R + 1$ bits beyond what is already present in p can be encoded in this way — using steganography or otherwise.

Recall that for a distribution $p \in \mathsf{Distr}(Y)$, its (Shannon) entropy is

$$H(p) \coloneqq -\sum_{y \in Y} p(y) \log p(y).$$

For jointly distributed random variables Q, M, recall the mutual information is

$$I(Q; M) \coloneqq H(Q) + H(M) - H(Q, M) = H(Q) - H(Q \mid M).$$

Algorithm 2 Consensus sampling with prompts

Inputs: Prompt $x \in X$, integers $k \geq s \geq 1$, $R \geq 0$, distributions $\mathbf{p} \in (\mathsf{Distr}(Y \mid X))^k$
Output: A sample $y \in Y$ or $\bot$

1: **for** $r = 1$ to R **do**
2: Pick a random index i and sample $y \sim p_i(\cdot \mid x)$
3: With probability $\dfrac{\frac{1}{s}\sum_{i \leq s} p_{(i)}(y \mid x)}{\frac{1}{k}\sum_{i=1}^{k} p_i(y \mid x)}$ **return** y
4: **end for**
5: **return** $\bot$

Theorem C.3 (Mutual-information leakage)**.** *Fix $p \in \mathsf{Distr}(Y)$ and $R \geq 0$. Let M be any message random variable. For each m in the support of M, let $q_m \in \mathsf{Distr}(Y \cup \{\bot\})$ be R-risky relative to p, i.e., $q_m(y) \leq R\,p(y)$ for all $y \in Y$. Let $Q \sim q_M$. Then $I(Q; M) \leq \log(R+1)$.*

Proof. Let $Y' := Y \cup \{\bot\}$ and define $p' \in \mathsf{Distr}(Y')$ by $p'(\bot) = \frac{1}{R+1}$ and $p'(y) = \frac{R}{R+1}\,p(y)$ for $y \in Y$. Then for all $y \in Y'$, $q_m(y) \leq (R+1)\,p'(y)$ (since $q_m(\bot) \leq 1$ and $q_m(y) \leq Rp(y)$ on Y).

Hence, pointwise $\log q_m(y) \leq \log(R+1) + \log p'(y)$, so

$$H(q_m) = -\sum_{y \in Y'} q_m(y) \log q_m(y) \geq -\sum_{y \in Y'} q_m(y) \log p'(y) - \log(R+1).$$

Averaging over M gives

$$H(Q \mid M) \geq -\sum_{y \in Y'} \Pr[Q = y] \log p'(y) - \log(R+1).$$

By the cross-entropy inequality (with the same support Y'),

$$H(Q) \leq -\sum_{y \in Y'} \Pr[Q = y] \log p'(y).$$

Subtracting the inequalities yields $I(Q; M) = H(Q) - H(Q \mid M) \leq \log(R+1)$. ☐

D Varying prompts

In this final section, we consider the inclusion of prompts $x \in X$. A model is now a conditional generation distributions $p \in \mathsf{Distr}(Y \mid X)$. The only change to the algorithm is the inclusion of the prompt $x \in X$ as an input, shown in Alg. 2.

Observation 6.5 trivially generalizes to using a single decoding function across prompts. (One can also generalize to a decoder $f : X \times Y \to \{0,1\}^*$.)

Observation D.1. *Fix a decoding function $f : Y \to \{0,1\}^*$ and $R \geq 1$. Let $q_x, p_x \in \mathsf{Distr}(Y)$ be output and safe distributions over y, given prompt x. If, for every prompt x, q_x is R-robust relative to p_x, then:*

$$\Pr_{Q \sim q_x}[f(Q) = m] \leq R \cdot \Pr_{P \sim p_x}[f(P) = m] \text{ for every } x \in X, m \in \{0,1\}^*.$$

The proof again follows trivially from the definition of R-robustness by considering the unsafe sets $U = f^{-1}(m)$. Unfortunately, the direct analog of the mutual information bound Theorem C.3 does not hold in the setting where the safe distribution depends on a prompt x. Instead, we use a more refined model of leakage called maximal leakage [Issa et al., 2020] in place of mutual information.

We consider a generation process that may involve a prompt x which determines the safe distribution $p_x \in \mathsf{Distr}(Y)$. That is, for each prompt x we assume there is a fixed distribution

p_x. We are concerned about leakage from a message m, which is possibly related to the prompt x. To this end, suppose that for each m, x there is a distribution of responses $q_{x,m} \in \mathsf{Distr}(Y)$. We thus wish to know how much information from m leaks into a sample $y \sim q_{x,m}$. Of course, significant information from the prompt may already be naturally encoded into p_x, so the question is then how much *more information* may leak from m into $q_{x,m}$ than leaks from m into p_x.

Issa et al. [Issa et al., 2020] define and justify the following definition of *maximal leakage* as a better measure of leakage than mutual information which, as they discuss, has known flaws.

Definition D.2 (Maximal leakage [Issa et al., 2020])**.** *For random variables M, Y, the maximal leakage from M to Y is*

$$\mathcal{L}(M \to Y) \;:=\; \log \sum_y \max_m \Pr[Y = y \mid M = m].$$

If Y and M are independent, then $\mathcal{L}(M \to Y) = 0$. If $Y = M$ then $\mathcal{L}(M \to Y)$ is log of the number of possible messages.

Theorem D.3 (Maximal leakage)**.** *Let X, M be jointly distributed random variables (for the prompt and message, respectively). For each of their possible values (x, m), let $p_x \in \mathsf{Distr}(Y)$ be a safe distribution and $q_{x,m} \in \mathsf{Distr}(Y \cup \{\perp\})$ a sampling distribution which is assumed to be R-robust relative to p_x. Define Q by sampling $Q \sim q_{X,M}$ and P by sampling $P \sim p_X$. Thus, conditioned on X, P and M are independent. Then*

$$\mathcal{L}(M \to Q) - \mathcal{L}(M \to P) \le \log(R + 1).$$

Proof of Theorem D.3. For any $y \in Y, m \in \{0,1\}^*$, since $q_{x,m}$ is R-robust relative to p_x for all x,

$$\begin{aligned} \Pr[Q = y \mid M = m] &= \mathbb{E}[q_{X,m}(y) \mid M = m] \\ &\le \mathbb{E}[R \cdot p_X(y) \mid M = m] \\ &= R \cdot \Pr[P = y \mid M = m]. \end{aligned}$$

Thus, for any $y \in Y$,

$$\max_m \Pr[Q = y \mid M = m] \le \max_m R \cdot \Pr[P = y \mid M = m].$$

Hence, summing over y and taking logs gives,

$$\begin{aligned} \mathcal{L}(M \to P) &= \log \sum_{y \in Y \cup \{\perp\}} \max_m \Pr[Q = y \mid M = m] \\ &\le \log \left(1 + \sum_{y \in Y} \max_m \Pr[Q = y \mid M = m] \right) \\ &\le \log \left(1 + \sum_y \max_m R \cdot \Pr[P = y \mid M = m] \right) \\ &= \log \sum_y \Big(\Pr[P = y] + R \cdot \max_m \Pr[P = y \mid M = m] \Big) \\ &\le \log \sum_y (R + 1) \cdot \max_m \Pr[P = y \mid M = m] \\ &= \log(R + 1) + \log \sum_y \max_m \Pr[P = y \mid M = m] \\ &= \log(R + 1) + \mathcal{L}(M \to P) \end{aligned}$$

This is exactly the bound $\mathcal{L}(M \to Q) \le \log(R+1) + \mathcal{L}(M \to P)$ needed for the theorem. $\square$