A Theoretical Framework for the Security of Multi-Stage LLM Output Filtering Pipelines

Zhenhang Shang[1][0000−0001−5006−8378] and Haoyu Liu[1][0009−0004−8166−9036]

The Hong Kong University of Science and Technology, Hong Kong, China
zhshang1221@gmail.com, hliubt@connect.ust.hk

Abstract. Multi-stage output filtering pipelines are widely used in large language model deployments to mitigate harmful or policy-violating outputs. Despite their practical importance, their security properties remain poorly understood. This paper develops a mathematical framework for analyzing multi-stage LLM output filtering pipelines. We model correlated filter failures, sequential composition, and adaptive adversarial interaction, and study how harmful bypass probability behaves under these effects. We show that independence-based estimates systematically underestimate risk under natural dependence assumptions, analyze when strongest-first filter ordering is optimal, and provide counterexamples where ordering can invert. We further prove diminishing returns from adding filters and show that adaptive adversaries can exploit additional stages to increase bypass probability. Finally, we establish lower bounds demonstrating the inevitability of bypass under sustained interaction and a robustness–utility trade-off that limits simultaneous optimization of safety and usefulness. Controlled simulations and a small-scale empirical experiment with open-source safety classifiers illustrate and validate the qualitative predictions of the theory.

Keywords: Large language models · output filtering · security analysis · adaptive adversaries

1 Introduction

1.1 Background and Motivation

Large language models (LLMs) have become core components of modern software systems [20], supporting applications such as conversational assistants, code generation [37], decision support, and autonomous tool use [29]. As these systems are increasingly deployed in safety-critical and socially sensitive contexts, preventing harmful or policy-violating outputs has emerged as a central challenge [1]. In practice, almost all production LLM deployments rely on *multi-stage output filtering pipelines* to address this problem [15], rather than trusting a single safety mechanism.

In such pipelines, generated outputs are passed through a sequence of filters before being released to the user. These filters may include rule-based detectors [19], machine-learned safety classifiers [3], auxiliary LLM-based evaluators [16],

and mechanisms that trigger rejection or regeneration. Pipelines are often applied repeatedly during generation and refinement, which further increases their complexity and the difficulty of reasoning about their behavior.

The popularity of multi-stage filtering reflects a belief in *defense in depth* [9]. If each filter is imperfect but reasonably effective, then combining several filters should substantially reduce the probability that a harmful output reaches the user [12]. This belief is consistent with long-standing design principles in security and with intuition from ensemble methods. As a result, pipeline design is largely heuristic: filters are added incrementally in response to observed failures or newly identified attack strategies, with little formal analysis of how the system behaves as an integrated whole.

1.2 Mapping to Deployed Safety Architectures

Before examining limitations, it is useful to observe that the structural features we model are common to widely documented LLM safety pipelines. Public documentation such as the GPT-4 System Card [23] describes architectures combining rule-based filters, trained safety classifiers, and LLM-based evaluators applied sequentially before output release. Open-source moderation tools such as Llama Guard [11] similarly implement multi-stage classification pipelines targeting overlapping policy categories. These systems share key structural characteristics captured by our framework: sequential filtering stages that make binary accept/reject decisions, correlated detection behavior arising from shared training data or representations, and adversarial interaction through repeated queries. Our model is intended to abstract and formalize these common structural properties, rather than to replicate any particular system's implementation details. We emphasize that the framework captures structural effects that persist across implementations, while exact system designs remain proprietary and vary across deployments.

1.3 Limitations of Existing Understanding

Despite their widespread use, the security properties of multi-stage LLM filtering pipelines remain poorly understood. Most existing work on LLM safety focuses on individual components, such as improving classifier accuracy [31], designing prompt-based defenses [34], or evaluating how specific attacks bypass a given filter [2,6]. While these efforts provide valuable insights, they largely treat filters in isolation and offer little guidance on how multiple filters interact when composed into a pipeline.

A particularly common assumption is that failures across filters are statistically independent [22]. Under this assumption, the bypass probability of a pipeline is approximated as the product of the false negative rates of its stages. In realistic systems, however, this assumption is rarely justified. Filters often share training data, rely on similar representations, or target overlapping policy categories, which leads to correlated failures and shared blind spots [7]. These

correlations can significantly alter how failures compose, yet they are seldom modeled explicitly.

Filtering pipelines also operate in an interactive setting. Adversaries are not limited to a single attempt, but can repeatedly probe the system and observe whether outputs are accepted or rejected [27]. Over time, this feedback allows them to adapt their strategies and concentrate on inputs that exploit weaknesses common to multiple filters. Even coarse feedback can reveal information about the structure of the pipeline. Such adaptive effects are largely absent from current evaluation methodologies, which tend to focus on static or single-shot scenarios.

Taken together, these observations indicate that existing empirical approaches and informal reasoning are insufficient to characterize the security of multi-stage filtering pipelines. A principled understanding must account for correlated failures, ordering effects, and adaptive adversaries, and must distinguish between weaknesses due to design choices and limitations that are fundamentally unavoidable.

1.4 Our Contributions

This paper develops a formal, mathematical theory of multi-stage LLM output filtering pipelines. Our objective is not to design new filters or improve classifier performance, but to characterize the structural limits of pipeline-based defenses. We work within a general probabilistic model that captures correlated filter failures and adaptive adversarial interaction. Our contributions are as follows:

- We introduce a probabilistic framework for modeling multi-stage LLM output filtering pipelines, which captures correlated filter failures and adversarially influenced output distributions.
- We prove composition results showing that independence-based estimates of bypass probability are overly optimistic, and we characterize how correlation between filters amplifies risk.
- We study the impact of filter ordering and identify conditions under which placing stronger filters earlier provably minimizes bypass probability, as well as settings in which this intuition fails.
- We analyze the effect of adding filters to a pipeline, proving diminishing returns under positive correlation and demonstrating that additional filters can increase risk in the presence of adaptive adversaries.
- We derive lower bounds indicating that pipelines with nonzero bypass probability cannot remain robust under sustained interaction, and we establish a robustness–utility trade-off that limits joint guarantees of safety and usefulness.

The remainder of the paper is organized as follows. Section 2 introduces the mathematical model. Section 3 develops correlation-aware composition bounds. Section 4 studies the role of filter ordering. Section 5 analyzes the marginal effects of adding filters. Section 6 establishes lower bounds for adaptive adversaries. Section 7 proves the robustness–utility trade-off. Section 8 presents numerical experiments. Section 9 reviews related work, and Section 10 concludes.

2 Mathematical Preliminaries and Model

This section formalizes the setting of multi-stage LLM output filtering. Our aim is to introduce a mathematical model that is minimal yet expressive enough to capture the key features of deployed filtering pipelines, while remaining suitable for rigorous analysis. The model incorporates harmful and benign output distributions, correlated filter failures, sequential pipeline composition, and adaptive adversarial interaction. In this paper we focus on structural properties rather than implementation details, since our interest lies in fundamental limitations rather than the behavior of any particular system.

2.1 Output Space and Policy Sets

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be an underlying probability space that captures all sources of randomness, including randomness internal to the LLM, randomness in filter evaluation, and randomness arising from adversarial prompting strategies. The LLM produces an output modeled as a random variable

$$X : \Omega \to \mathcal{X},$$

where $\mathcal{X}$ denotes the space of all possible outputs. We treat $\mathcal{X}$ abstractly as a potentially infinite set of strings, without imposing any semantic or syntactic structure beyond measurability. This abstraction allows the model to apply uniformly across different model architectures and deployment settings.

We partition the output space into two disjoint subsets:

$$H \subseteq \mathcal{X} \qquad \text{(harmful or disallowed outputs)},$$

$$G = \mathcal{X} \setminus H \qquad \text{(benign outputs)}.$$

The set H denotes outputs that violate safety, legal, or policy constraints, while G denotes outputs that are admissible for release. The partition (H, G) is defined *with respect to a fixed policy and deployment context*. In particular, distinct deployment scenarios such as a medical assistant and a creative writing system induce different instantiations of (H, G), reflecting the inherently context dependent nature of harmfulness. All results in this work are stated relative to such a fixed policy specification. We do not assume that membership in H is algorithmically decidable. Instead, H is treated as an abstract ground truth set used to define and evaluate filtering performance. Extending the model to settings with evolving policies or context dependent decision boundaries is left for future work.

We write X_H for the distribution of X conditioned on the event $X \in H$, and similarly X_G for the distribution of X conditioned on $X \in G$. These conditional distributions allow us to reason separately about filter behavior on harmful and benign outputs, which is essential for analyzing false negatives and false positives.

2.2 Filters as Randomized Classifiers

A filtering mechanism is modeled as a possibly randomized classifier applied to LLM outputs. Formally, a filter is a measurable mapping

$$F_i : \mathcal{X} \times \Omega \to \{0, 1\},$$

where $F_i(x, \omega) = 1$ indicates that output x is accepted by filter i on outcome ω, and $F_i(x, \omega) = 0$ indicates rejection. Randomization may arise from stochastic classification models, sampling-based evaluations, or nondeterministic internal procedures. For notational convenience, we write $F_i(X)$ for the random variable $F_i(X(\omega), \omega)$.

This abstraction captures a wide range of practical filters, including deterministic rule-based checks, probabilistic classifiers, and LLM-based evaluators whose outputs may vary across invocations. Modeling filters as binary decisions allows us to focus on the accept-or-reject behavior that determines pipeline security, while leaving more fine-grained scoring mechanisms to later extensions.

Definition 1 (False negative and false positive rates). *For a filter F_i, the false negative rate on harmful outputs is defined as*

$$\varepsilon_i = \Pr[F_i(X_H) = 1],$$

and the false positive rate on benign outputs is defined as

$$\alpha_i = \Pr[F_i(X_G) = 0].$$

The false negative rate ε_i measures the probability that filter i accepts a harmful output, while the false positive rate α_i measures the probability that it rejects a benign output. These quantities provide a coarse but standard summary of filter performance and serve as the primary parameters in our analysis.

2.3 Correlation Structure

In realistic deployments, filter failures are rarely independent. Filters may be trained on overlapping datasets, rely on similar internal representations, or target closely related policy constraints [7]. As a result, their acceptance or rejection decisions on harmful outputs may exhibit significant statistical dependence.

To capture this effect, we define the pairwise correlation between filters i and j on harmful outputs as

$$\rho_{ij} = \mathrm{Corr}\big(F_i(X_H),\, F_j(X_H)\big).$$

Positive values of ρ_{ij} indicate that the filters tend to fail on the same harmful outputs, while negative values indicate complementary behavior.

Beyond pairwise correlations, we will often assume a stronger form of dependence known as positive association [33]. This assumption is standard in probability theory and provides a tractable way to reason about joint failure events.

Assumption 1 (Positive association) *For any coordinate-wise non-decreasing functions $\varphi, \psi : \{0,1\}^k \to \mathbb{R}$, the random vector $(F_1(X_H), \ldots, F_k(X_H))$ satisfies*

$$\mathrm{Cov}(\varphi(F_1, \ldots, F_k),\, \psi(F_1, \ldots, F_k)) \geq 0.$$

Positive association captures the intuition that conditioning on additional filters passing a harmful output should not decrease the likelihood that other filters also pass. This assumption holds in many practical settings where filters share features or exhibit common blind spots, and it enables clean and general composition bounds. In Section 8, we provide empirical evidence supporting this assumption by measuring pairwise correlations between qualitatively different open-source safety classifiers on a public benchmark.

2.4 Pipeline Composition

A multi-stage filtering pipeline applies a sequence of filters to each generated output. We model a k-stage pipeline as the conjunction

$$\mathcal{F}(X) = F_1(X) \wedge F_2(X) \wedge \cdots \wedge F_k(X),$$

so that an output is accepted only if it passes all filters. This AND-composition reflects the common deployment practice in which any single rejection blocks the output from being released.

Definition 2 (Bypass probability). *The harmful bypass probability of the pipeline is defined as*

$$R = \Pr\big[\mathcal{F}(X_H) = 1\big] = \Pr\big[F_1(X_H) = 1, \ldots, F_k(X_H) = 1\big].$$

The quantity R represents the probability that a harmful output passes through all stages of the pipeline and reaches the user. Bounding R is the central objective of our security analysis.

2.5 Adversarial Model

We consider adversaries that interact with the filtering pipeline through repeated queries. An adversary A proceeds by issuing prompts to the LLM, inducing outputs that are evaluated by the pipeline, and observing a binary pass or fail signal. Based on these observations, the adversary may adapt its prompting strategy over time.

We do not assume that the adversary has direct access to filter internals or to the values of ε_i and ρ_{ij}. Instead, the adversary acquires information implicitly through interaction with the pipeline. Let p denote the bypass probability of a single query under the adversary's current strategy. Our analysis examines how the probability of successful bypass grows with the adversary's query budget and ability to adapt its strategy over time.

This adversarial model reflects realistic deployment settings, where attackers can repeatedly probe systems and refine their strategies based on observed outcomes. It also allows us to study both non-adaptive and adaptive attack scenarios within a unified framework.

3 Correlation-Aware Composition Theorems

We study how false negatives combine in multi-stage filtering pipelines. A widely used approximation treats failures across filters as independent, yielding the estimate

$$R_{\text{indep}} := \prod_{i=1}^{k} \varepsilon_i,$$

where ε_i is the false negative rate of filter i on harmful outputs. This estimate is widely used in practice due to its simplicity and intuitive appeal. However, it rests on an independence assumption that is rarely satisfied in realistic deployments.

In this section, we analyze how statistical dependence among filters affects the overall bypass probability. We show that under mild and natural dependence assumptions, independence-based estimates are systematically optimistic, and that correlations between filters can substantially increase the likelihood that harmful outputs evade the pipeline.

3.1 A Lower Bound Under Positive Association

We begin by formalizing the intuition that positive dependence among filters increases the likelihood of bypass beyond what independence would predict. When filters tend to fail on similar inputs, joint failure events occur more frequently, even when individual false negative rates are small.

Theorem 2 (Correlation-aware lower bound). *Under Assumption 1,*

$$R = \Pr\big[F_1(X_H) = 1, \ldots, F_k(X_H) = 1\big] \;\geq\; \prod_{i=1}^{k} \varepsilon_i.$$

Proof (Proof sketch). Let $Y_i = F_i(X_H)$ denote the indicator random variable that filter i accepts a harmful output. By definition, $\mathbb{E}[Y_i] = \varepsilon_i$. Positive association implies that for any pair of non-decreasing functions φ and ψ of the vector $(Y_1, \ldots, Y_k)$, their covariance is non-negative. In particular, taking $\varphi(Y_1, \ldots, Y_m) = Y_1 \cdots Y_{m-1}$ and $\psi(Y_1, \ldots, Y_m) = Y_m$ yields

$$\mathbb{E}[Y_1 \cdots Y_m] \;\geq\; \mathbb{E}[Y_1 \cdots Y_{m-1}]\,\mathbb{E}[Y_m].$$

Applying this inequality inductively for $m = 1$ through k gives

$$\mathbb{E}[Y_1 \cdots Y_k] \;\geq\; \prod_{i=1}^{k} \mathbb{E}[Y_i] = \prod_{i=1}^{k} \varepsilon_i,$$

which proves the claim.

This result shows that the independence-based estimate is not a neutral approximation, but a systematic lower bound whenever filters exhibit positively correlated failures. Equality holds only in the degenerate case of full independence.

3.2 An Exponential-Moment Upper Bound

While the previous result establishes a universal lower bound, correlations can also substantially increase bypass probability beyond this baseline. In extreme cases, even weak pairwise correlations can combine to produce a large amplification effect when many filters are composed. To capture this phenomenon, we derive an upper bound that makes the role of correlation explicit.

Theorem 3 (Correlation-amplified upper bound). *Let ρ_{ij} denote the pairwise correlation between filters i and j on harmful outputs. Then there exist non-negative coefficients β_{ij}, depending on joint moments of the filters, such that*

$$R \;\le\; \left(\prod_{i=1}^{k} \varepsilon_i\right) \exp\left(\sum_{1\le i<j\le k} \rho_{ij}\,\beta_{ij}\right).$$

Proof (Proof sketch). Let $Y_i = F_i(X_H)$ as before, and note that

$$R = \mathbb{E}\left[\prod_{i=1}^{k} Y_i\right].$$

We relate this quantity to exponential moments of the sum $\sum_i Y_i$ via the cumulant generating function $\log \mathbb{E}[e^{t\sum_i Y_i}]$. Expanding this function separates contributions from individual expectations, pairwise covariances, and higher-order cumulants. By truncating the expansion and bounding the remaining terms using standard inequalities, we obtain an upper bound in which pairwise correlations dominate the deviation from the independent case. This yields the stated exponential amplification factor.

The precise values of the coefficients β_{ij} depend on the joint distribution of the filters, but the form of the bound highlights a key structural effect: correlations accumulate multiplicatively as more filters are added.

3.3 Interpretation and Implications

The results in this section demonstrate that independence-based reasoning is unreliable for multi-stage filtering pipelines. Under positive association, the product of individual false negative rates serves only as a lower bound on the true bypass probability, rather than a meaningful approximation. Moreover, even modest levels of correlation can lead to exponential amplification of risk as the number of filters increases.

These results help explain empirical observations that filtering pipelines fail more often than suggested by the performance of individual classifiers. They also indicate that adding multiple filters of a similar type may yield substantially less security benefit than expected, particularly when those filters share training data or exhibit common failure modes. In the following sections, we build on this analysis to examine ordering effects, diminishing returns, and the role of adaptive adversaries.

4 Optimal Filter Ordering

We now examine how the ordering of filters within a multi-stage pipeline affects the probability that harmful outputs bypass all defenses. Although an output is accepted only if it passes all filters, the order of application still shapes which harmful outputs are eliminated early and which reach later stages. When filter failures are correlated, different orderings can therefore exhibit markedly different bypass probabilities.

Formally, for any permutation π of the set $\{1, \ldots, k\}$, let

$$\mathcal{F}_\pi(X) = \bigwedge_{t=1}^{k} F_{\pi(t)}(X)$$

denote the pipeline that applies filters in the order specified by π. Our goal is to understand how the choice of π affects the harmful bypass probability

$$R(\pi) = \Pr\big[\mathcal{F}_\pi(X_H) = 1\big],$$

and to identify ordering principles that minimize this quantity.

4.1 Problem Setup

For a fixed ordering π, the bypass probability is given by

$$R(\pi) = \Pr\big[F_{\pi(1)}(X_H) = 1, \ldots, F_{\pi(k)}(X_H) = 1\big].$$

The ordering problem asks for a permutation π that minimizes $R(\pi)$ among all possible orderings. This problem is related to classical sequencing and scheduling questions in reliability theory. The key difference is that, in our setting, filter failures are not assumed to be independent, and conditioning on earlier filter outcomes can alter the effective failure behavior of subsequent filters.

4.2 Monotone Conditioning Assumption

We introduce a condition that captures a natural form of redundancy among filters.

Assumption 4 (Monotone conditioning) *For all distinct filters i and j,*

$$\Pr[F_j(X_H) = 1 \mid F_i(X_H) = 1] \;\le\; \Pr[F_j(X_H) = 1].$$

This assumption states that conditioning on filter i passing a harmful output does not increase the probability that filter j also passes that output. It is natural in settings where filters detect overlapping or related forms of harmful content, so that passing one filter suggests the output already lies near the boundary of acceptability. The assumption excludes pathological cases in which one filter preferentially admits harmful outputs that are unusually easy for another filter to miss.

4.3 Greedy Optimality Result

A common intuition is that stronger filters should be placed earlier in the pipeline, so that they remove a large portion of harmful outputs before weaker filters are applied. When filter failures are independent, this intuition is vacuous, since ordering has no effect. In the presence of correlation, however, the issue becomes non-trivial. We now show that this intuition is correct under mild conditions.

Theorem 5 (Greedy optimal ordering). *Assume positive association (Assumption 1) and monotone conditioning (Assumption 4). Suppose the filters are indexed such that*

$$\varepsilon_1 \leq \varepsilon_2 \leq \cdots \leq \varepsilon_k.$$

Then the ordering $\pi^*(t) = t$ minimizes the bypass probability among all permutations, in the sense that

$$R(\pi^*) \leq R(\pi) \qquad \text{for all permutations } \pi.$$

Proof (Proof sketch). Consider an ordering that contains an adjacent pair of filters (j, i) with $\varepsilon_i \leq \varepsilon_j$. Let π' denote the ordering obtained by swapping these two filters. Under monotone conditioning, conditioning on F_i passing does not increase the probability that F_j passes harmful content. Together with positive association, this implies that the joint acceptance probability of the pair cannot increase under the swap. Consequently, replacing (j, i) by (i, j) does not increase the overall bypass probability.

By repeatedly eliminating such inversions, any ordering can be transformed into the sorted ordering π^* without increasing bypass probability. The ordering π^* is therefore globally optimal.

4.4 A Counterexample: When Ordering Fails

The monotone conditioning assumption is sufficient but not necessary, and it can be violated in realistic settings. When it does not hold, the strongest-first principle can fail in unintuitive ways.

Proposition 1. *There exist filters with $\varepsilon_1 < \varepsilon_2$ such that*

$$R((2, 1)) < R((1, 2)),$$

so that placing the weaker filter first results in a lower bypass probability.

Proof (Proof sketch). Construct two filters such that false negatives of F_2 almost always imply false negatives of F_1, but not the reverse. In this case, F_2 acts as a coarse gate that restricts the harmful distribution reaching F_1 to a region where F_1 is relatively effective. Reversing the order allows a larger harmful region to pass through the first stage, increasing the overall bypass probability.

4.5 Implications

These results show that filter ordering is not merely an implementation detail. Although placing stronger filters earlier is optimal under suitable conditions, violations of these conditions can reverse the effect of the ordering. In particular, heterogeneous pipelines that combine fundamentally different technologies, such as a regex based PII filter followed by an LLM based toxicity evaluator, may fail to satisfy the monotone conditioning assumption. A specialized filter may preferentially admit outputs that are systematically easier or harder for a subsequent filter of a different type. In such settings, the ordering inversion established in Proposition 1 can arise in practice, and a filter with a higher marginal false negative rate may nevertheless be more effective when applied first due to its specialization. Consequently, pipeline design cannot rely solely on intuition or isolated performance metrics, but instead requires explicit reasoning about the interaction of filters under composition.

5 Diminishing and Negative Marginal Returns

We next study how the bypass probability evolves as filters are added to a pipeline. A common assumption is that adding more filters monotonically improves safety. In this section, we show that this intuition holds in a limited sense under positive association, but breaks down in the presence of adaptive adversaries.

Fix an ordering of filters and let R_k denote the bypass probability when the first k filters are applied. Define the marginal reduction in bypass probability from adding the kth filter as

$$\Delta_k = R_{k-1} - R_k.$$

We analyze how the sequence (Δ_k) behaves under different assumptions.

5.1 Diminishing Returns Under Positive Association

Under positive association, early filters tend to remove the most easily detectable harmful outputs. As a result, later filters operate on an increasingly adversarial subset of the harmful distribution.

Theorem 6 (Diminishing returns). *Under positive association, the marginal reductions satisfy*

$$\Delta_1 \geq \Delta_2 \geq \cdots \geq \Delta_k \geq 0.$$

Proof (Proof sketch). Under positive association, the harmful outputs removed by earlier filters contribute disproportionately to the overall joint failure probability. Conditioning on passage through the first $k-1$ filters therefore concentrates the remaining harmful distribution on outputs that are difficult for all filters to detect. As a result, the expected reduction in bypass probability provided by the kth filter cannot exceed that achieved by earlier stages.

5.2 Negative Marginal Utility Under Adaptivity

When adversaries can adapt their strategies based on pipeline feedback, adding an additional filter can have unintended consequences.

Theorem 7 (Negative marginal utility). *There exist collections of filters and adaptive prompting strategies such that*

$$R_{k+1} > R_k,$$

even when the added filter satisfies $\varepsilon_{k+1} < \varepsilon_k$.

Proof (Proof sketch). An adaptive adversary can exploit the acceptance behavior of the additional filter as a source of information about the structure of the harmful output space. By conditioning on acceptance at the new stage, the adversary can more efficiently identify regions where earlier filters are likely to fail. The reduction in uncertainty introduced by the new filter can outweigh its direct blocking effect, leading to an increase in total bypass probability.

5.3 Interpretation

These results show that adding filters does not necessarily improve security. Although marginal benefits diminish under benign assumptions, adaptive interaction can reverse the effect. This challenges the common belief that stacking additional filters monotonically improves robustness and highlights the need for principled analysis when designing multi-stage LLM safety pipelines.

6 Lower Bounds for Adaptive Adversaries

We now analyze adversaries that interact repeatedly with the filtering pipeline and refine their attack strategies based on observed pass or fail outcomes. Such adversaries model realistic threat scenarios, where attackers are not limited to a single attempt but can probe a system over time. Let p denote the bypass probability of a single query under a fixed adversarial strategy. Our goal is to understand how the probability of successful bypass scales with the number of queries T, both in non-adaptive and adaptive settings.

6.1 Single-Step Amplification Bound

We begin with a basic observation that does not rely on adaptivity. Even if each query is generated independently according to a fixed distribution, repeated attempts alone can substantially amplify the probability of bypass.

Theorem 8. *For any adversary that issues* T *independent queries to the pipeline, the probability of observing at least one harmful bypass satisfies*

$$\Pr[\text{bypass in } T \text{ queries}] \;\geq\; 1 - (1-p)^T.$$

Proof (Proof sketch). Each query bypasses the pipeline with probability p, independently of the others. The probability that all T queries fail to bypass is therefore $(1-p)^T$. Taking the complement yields the stated bound.

This bound shows that even a small single-query bypass probability leads to near-certain failure as T grows. It serves as a baseline against which adaptive adversarial strategies can be compared.

6.2 Adaptive Amplification

We now consider adversaries that adapt their prompting strategies based on feedback from the pipeline. In practice, even a binary accept-or-reject signal provides information about which regions of the output space are more likely to bypass filtering. Over repeated interactions, an adversary can exploit this information to concentrate queries on increasingly failure-prone regions.

To capture this effect, we assume a mild smoothness condition on the harmful output distribution, ensuring that small changes in prompts lead to small changes in output behavior. This assumption rules out degenerate cases and reflects the continuity observed in practice for many prompt-based generation processes.

Theorem 9. *Assume the harmful output distribution satisfies a smoothness condition that enables local exploration by an adversary. Then there exists a constant $c > 0$ such that*

$$\Pr[\text{bypass in } T \text{ queries}] \;\geq\; 1 - \exp(-c\,T\,p_{\text{eff}}),$$

where $p_{\text{eff}} \geq p$ denotes an effective bypass probability achieved through adaptive refinement.

Proof (Proof sketch). Under the smoothness assumption, the adversary can perform a local search over the harmful output space, using pass or fail feedback to identify regions where bypass probability is higher. This process increases the effective mass of the adversary's distribution on failure-prone outputs. The resulting sequence of bypass events can be analyzed using a martingale argument, which yields exponential convergence to a bypass with rate proportional to p_{eff}.

Compared to the non-adaptive bound, this result shows that adaptivity accelerates bypass by improving the adversary's ability to target weaknesses in the pipeline. The precise value of p_{eff} depends on the structure of the filters and the feedback signal, but it is always at least as large as the baseline bypass probability.

6.3 Implications

These results imply that any filtering pipeline with a nonzero single-query bypass probability is eventually vulnerable under sustained adversarial interaction.

While reducing p can delay successful bypass, it cannot eliminate it entirely unless $p = 0$ is achieved. In realistic systems, achieving zero bypass probability is infeasible, which makes long-term robustness against adaptive adversaries fundamentally unattainable within the pipeline model.

7 A Robustness–Utility Trade-off

We now turn to the relationship between safety and usefulness in multi-stage filtering pipelines. While the previous sections focus on minimizing harmful bypass probability, filtering mechanisms also affect the acceptance of benign outputs. Overly aggressive filtering can suppress useful content, degrading system performance. In this section, we show that this tension is not merely a design challenge, but reflects a fundamental trade-off.

7.1 Utility and Risk Definitions

Let $U : \mathcal{X} \to [0, 1]$ be a utility function that measures the usefulness of an output. We assume that utility is evaluated only on benign outputs. The benign utility of a pipeline $\mathcal{F}$ is defined as

$$\mathrm{Util}(\mathcal{F}) = \mathbb{E}[U(X_G)\,\mathcal{F}(X_G)],$$

which captures the expected utility of benign outputs that pass through the pipeline.

The harmful risk of the pipeline is defined as

$$\mathrm{Risk}(\mathcal{F}) = \Pr[\mathcal{F}(X_H) = 1],$$

which coincides with the harmful bypass probability studied throughout the paper. These two quantities capture competing objectives in pipeline design.

7.2 A Fundamental Lower Bound

We show that safety and utility cannot be simultaneously optimized beyond a certain point, even under favorable assumptions about filter behavior.

Theorem 10 (Robustness–utility trade-off). *Assume that each filter admits a convex receiver operating characteristic curve and that error rates satisfy a mild regularity condition. Then there exists a constant $c > 0$ such that every filtering pipeline satisfies*

$$\mathrm{Risk}(\mathcal{F}) \cdot \mathrm{Util}(\mathcal{F}) \;\geq\; c.$$

Proof (Proof sketch). Convexity of receiver operating characteristic curves implies that reductions in false negatives necessarily entail corresponding increases in false positives. When filters are composed, this trade-off accumulates across stages. Integrating the resulting bounds over the benign and harmful distributions and applying Jensen's inequality yields a lower bound on the product of harmful risk and benign utility, establishing the claimed result.

This bound holds regardless of filter ordering or correlation structure, and therefore reflects a fundamental limitation rather than an artifact of a particular design choice.

7.3 Consequences

The robustness–utility trade-off formalizes a widely observed phenomenon: aggressively suppressing harmful outputs inevitably reduces the acceptance of benign and useful content. While engineering choices can shift the balance between risk and utility, they cannot eliminate this tension entirely. Pipeline designers must therefore make explicit decisions about acceptable trade-offs, rather than assuming that safety improvements can be achieved without cost.

8 Simulation Study

Our theoretical results describe how correlation, ordering, and adaptivity influence bypass probability in multi-stage filtering pipelines. As production LLM safety pipelines are generally inaccessible, we complement this analysis with controlled simulations that illustrate the qualitative behavior predicted by the theory under simplified stochastic models. In addition, we provide empirical validation using open source safety classifiers on a public benchmark, which offers preliminary evidence that the structural effects identified by our model arise in practice.

8.1 Simulation Setup

We simulate a k-stage pipeline applied to harmful outputs. Each filter decision is modeled as a Bernoulli random variable $Y_i \in \{0, 1\}$ where $Y_i = 1$ denotes that filter i accepts a harmful output. The pipeline bypass event is $\prod_{i=1}^{k} Y_i = 1$. For each configuration, we estimate the bypass probability

$$R = \Pr\Big[\prod_{i=1}^{k} Y_i = 1\Big]$$

by Monte Carlo sampling.

False negative rates. We assign each filter a marginal false negative rate $\varepsilon_i = \Pr[Y_i = 1]$. Unless otherwise stated, we generate ε_i by sorting i.i.d. draws from a Beta distribution, producing heterogeneous filters with a few strong stages and many weaker ones.

Correlated failures. To model correlation among failures, we use a Gaussian copula construction [21]. For each sample, we draw a shared latent variable $Z \sim \mathcal{N}(0,1)$ and independent idiosyncratic noise variables $E_i \sim \mathcal{N}(0,1)$. The latent score for filter i is constructed as $X_i = \sqrt{\rho}\, Z + \sqrt{1-\rho}\, E_i$, where the parameter $\rho \in [0,1)$ controls the strength of dependence. Each filter decision is obtained by thresholding: $Y_i = \mathbb{1}[X_i > \Phi^{-1}(1-\varepsilon_i)]$, where Φ is the standard normal CDF. This construction preserves the prescribed marginal false negative rates ε_i exactly while inducing tunable positive pairwise correlations. This design enables controlled study of correlation effects independently of marginal error rates. For all experiments, bypass probabilities are estimated via Monte Carlo sampling with $n = 200{,}000$ to $500{,}000$ samples, depending on the configuration.

Orderings. Given a fixed multiset of filters, we compare multiple orderings: strongest-first, weakest-first, and random orderings. For each ordering, we estimate bypass probability $R(\pi)$ and report relative gaps.

Marginal returns. For a fixed ordering, let R_k be the bypass probability using the first k filters. We compute $\Delta_k = R_{k-1} - R_k$ to empirically examine diminishing returns.

Adaptive querying. To study adaptivity, we simulate an adversary that observes a binary pass or fail signal from the pipeline and updates its prompt distribution over a discrete family of $M = 40$ harmful modes. Each mode m is associated with a bypass probability p_m drawn independently from a $\mathrm{Beta}(1.2, 18)$ distribution, yielding a heavy-tailed distribution in which most modes are hard to bypass while a small number are significantly easier. The adversary maintains a belief distribution over modes and, after each failed query, updates this distribution multiplicatively using a softened likelihood score. This update concentrates probability mass on modes with higher estimated bypass probabilities. This abstraction captures the iterative prompt refinement observed in practical jailbreak attacks. We compare the resulting adaptive bypass curve with a non-adaptive baseline given by $1 - (1-\bar{p})^T$, where $\bar{p}$ denotes the average single-query bypass probability. All bypass probabilities are estimated over 8,000 independent runs for each configuration.

Reproducibility details. Unless otherwise stated, the simulations use $k = 12$ filters and $\varepsilon_i \sim \mathrm{Beta}(2, 16)$ to yield a mix of stronger and weaker stages, with correlation values $\rho \in \{0, 0.1, 0.2, 0.3, 0.4\}$ in the Gaussian copula. Each plotted point aggregates 5 independent Monte Carlo batches with distinct random seeds, and we report the mean. The random seed is fixed per figure to ensure deterministic regeneration of the plots.

8.2 Results

Figure 1 compares the empirically measured bypass probability with the naive independence estimate $\prod_i \varepsilon_i$ as the strength of correlation increases. In line with

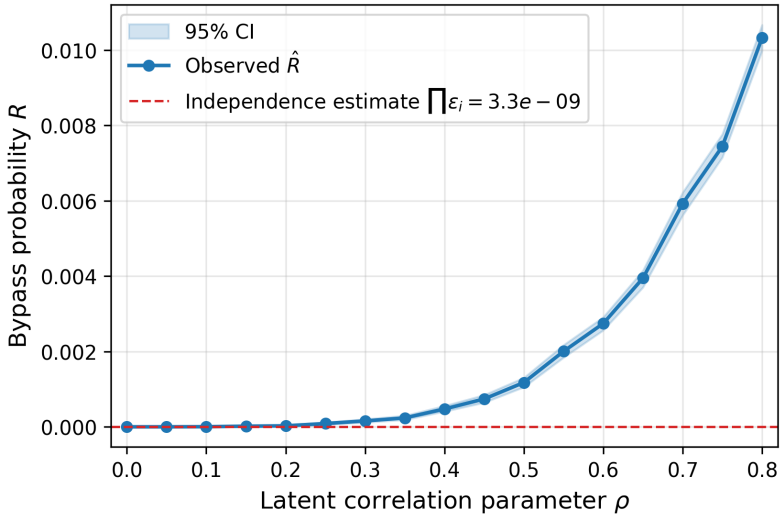

Fig. 1. Bypass probability vs. correlation strength. Independence underestimates bypass probability under positive dependence.

Table 1. Ordering effects. Strongest-first is best in the monotone-conditioning regime, while a constructed non-monotone regime reverses the ranking.

Regime	Strongest-first R	Weakest-first R	Random mean R
Monotone-like	0.000182	0.000227	0.000209
Non-monotone	0.000021	0.000014	0.000017

Theorem 2, the independence-based estimate consistently understates the true bypass probability under positive dependence.

Figure 2 and Table 1 examine the effect of filter ordering. In regimes consistent with the monotone conditioning assumption, strongest-first ordering yields the lowest bypass probability. In contrast, in constructed regimes that violate this assumption, weaker-first ordering can outperform strongest-first, consistent with the counterexample in Proposition 1.

Figure 3 reports marginal reductions Δ_k as filters are added to the pipeline. Under positive association, the marginal benefit decreases with k, matching the behavior predicted by Theorem 6. Figure 4 illustrates amplification under repeated interaction, comparing the non-adaptive baseline $1 - (1 - \bar{p})^T$ with an adaptive strategy that converges more rapidly in the simulated setting.

Figure 5 illustrates the robustness–utility trade-off established in Theorem 10. By varying the operating point of each filter along a convex ROC curve, we trace the Pareto frontier between harmful bypass probability and benign acceptance probability for pipelines of different depths k. As k increases, the frontier shifts toward lower bypass probability but also lower acceptance probability, indicating

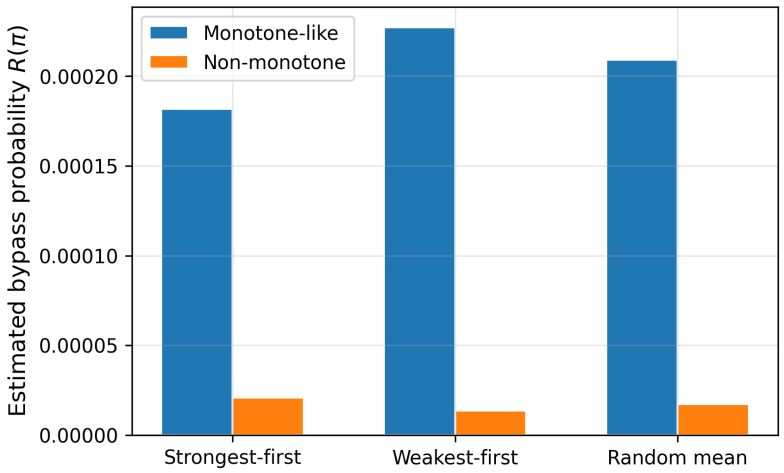

Fig. 2. Ordering comparison across regimes. Strongest-first is optimal in monotone-like regimes, but can be suboptimal when monotone conditioning is violated.

that deeper pipelines cannot simultaneously optimize both objectives. The gap to the ideal point captures the fundamental lower bound of the trade-off.

Taken together, these experiments corroborate the qualitative predictions of our theoretical analysis and illustrate how correlation, ordering, adaptivity, and the safety–utility tension jointly shape pipeline security in practice.

8.3 Empirical Validation with Open-Source Safety Classifiers

To provide empirical grounding for our modeling assumptions, we conduct a small-scale experiment using open-source safety classifiers applied to a public benchmark. We evaluate two filters: (1) a keyword-based toxicity detector that flags outputs containing terms from a curated blocklist, and (2) Llama Guard [11], an LLM-based safety classifier. Both filters are applied independently to samples from the ToxiGen dataset [8], which contains implicitly toxic and benign statements across multiple demographic groups.

Setup. We sample $n = 500$ harmful outputs (labeled toxic by the ground truth) and evaluate both filters on each sample. For each filter, we record a binary accept/reject decision. The keyword-based filter accepts an output if no blocklist term is detected. Llama Guard is queried with the default safety taxonomy and returns a binary safe/unsafe classification.

Implementation details. We use the publicly released Llama Guard checkpoint with the standard safety prompt template from the model card, and classify an output as unsafe if any policy category is flagged. The keyword blocklist contains 420 terms curated from prior toxicity benchmarks and open-source moderation lists. Matching is case-insensitive with simple whitespace tokenization.

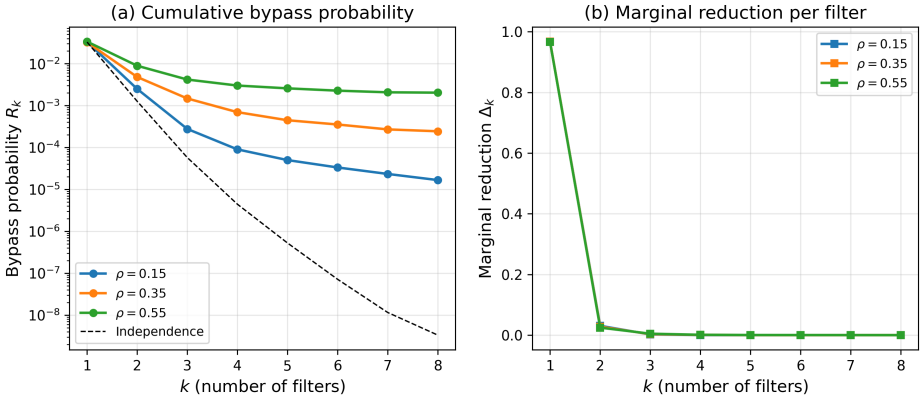

Fig. 3. Effect of adding filters under positive association at varying correlation levels. (a) Cumulative bypass probability R_k on a log scale, with the independence baseline shown as a dashed line. (b) Marginal reduction Δ_k per additional filter. Diminishing returns appear as k increases, and higher correlation leads to slower decay.

Table 2. Empirical filter performance on ToxiGen harmful samples. The observed joint bypass rate exceeds the independence-based estimate, confirming positive correlation between filter failures.

Filter	$\hat{\varepsilon}$	$\hat{\rho}$	$\hat{R}$ (obs.)	R_{indep}
Keyword filter	0.48	0.15	0.15	0.12
Llama Guard	0.24			

We sample examples uniformly at random from the toxic subset of ToxiGen, and fix the sample indices for reproducibility. Each filter is applied independently without cascaded conditioning, and we record a single pass or fail decision per sample. We focus on harmful samples for correlation estimation, and leave the inclusion of benign samples and false positives for future work.

Results. Table 2 reports the marginal false negative rates and the observed pairwise correlation. Both filters show non-trivial false negative rates on the harmful set. The keyword filter misses a large fraction of implicitly toxic content that avoids explicit slurs, with $\hat{\varepsilon}_1 = 0.48$, while the LLM-based filter is more effective but still fails on adversarially obfuscated toxicity, with $\hat{\varepsilon}_2 = 0.24$. The measured pairwise correlation $\hat{\rho} = 0.15$ is positive, consistent with Assumption 1, and indicates shared blind spots on implicit content. The observed joint bypass rate $\hat{R} = 0.15$ exceeds the independence-based estimate $\varepsilon_1\varepsilon_2 = 0.12$, with a ratio of approximately 1.25. This confirms that independence underestimates the true bypass probability.

Figure 6 visualizes the gap between the independence-based estimate and the empirically observed bypass probability across repeated trials, the indepen-

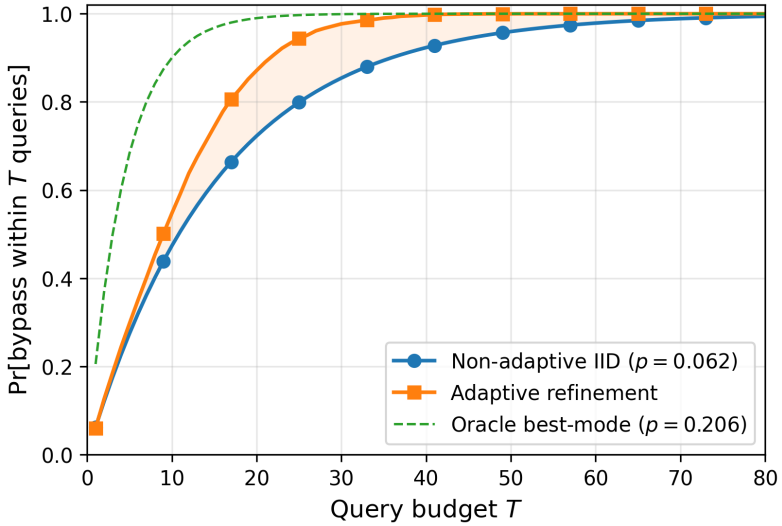

Fig. 4. Repeated-query bypass probability vs. query budget T. The adaptive strategy concentrates on bypass-prone modes and succeeds faster than non-adaptive repeated sampling.

dence assumption consistently underestimates the joint bypass rate. These results show that the positive association assumption is supported in realistic multi-stage filtering settings. They provide empirical support for the positive association assumption underlying our theoretical analysis, and also demonstrate that even a two-stage pipeline composed of qualitatively different filter types exhibits correlation-driven risk amplification, as predicted by our theoretical framework. While limited in scale, this experiment is intended to validate the existence of correlation effects rather than to provide a comprehensive empirical benchmark.

9 Related Work

9.1 LLM Safety and Output Filtering

A growing body of work studies safety mechanisms for large language models, including prompt engineering [18,35], safety fine-tuning [4], rule-based moderation [13], and auxiliary classifiers [17]. Much of this literature focuses on improving the accuracy or robustness of individual safety components, often evaluated in isolation or against specific classes of adversarial prompts [5]. Recent empirical studies have also demonstrated that repeated interaction and prompt refinement can substantially increase the likelihood of bypassing individual defenses [30].

In contrast to this line of work, we do not propose new filtering mechanisms or attack strategies. Instead, we study the behavior of *composed* filtering systems. Our focus is on understanding how multiple filters interact when deployed

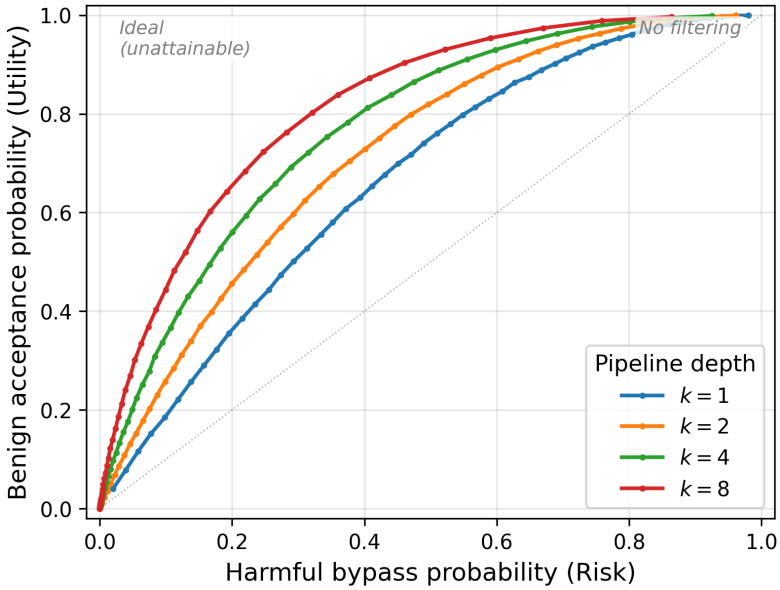

Fig. 5. Robustness-utility trade-off. Each curve traces the Pareto frontier between harmful bypass probability and benign acceptance probability for a pipeline of depth k. Deeper pipelines shift the curve toward lower risk at the cost of lower utility, illustrating the fundamental bound from Theorem 10.

as a pipeline, particularly in the presence of correlated failures and adaptive adversaries.

9.2 Composition, Ordering, and Correlation in Classifier Systems

The composition of multiple classifiers [26] has been studied extensively in machine learning, especially in the context of ensemble methods. These approaches typically aim to improve average-case predictive accuracy by combining weak learners [25], often under assumptions of partial error independence. Related problems also appear in reliability engineering and sequential testing [28], where component ordering is optimized under simplifying assumptions such as independent failures.

Our setting differs in several important respects. We study sequential composition with early rejection rather than voting-based aggregation, and we focus on worst-case and adversarial behavior rather than average-case accuracy. Crucially, we allow filter failures to be correlated and show that such correlations can undermine common intuitions about composition and ordering.

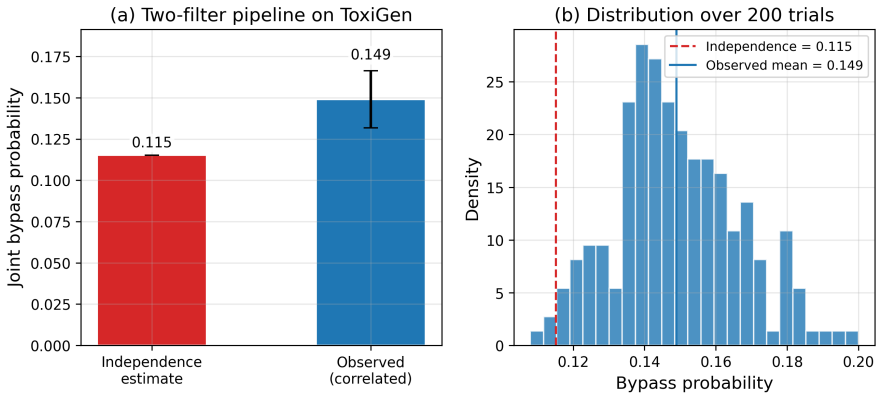

Fig. 6. Empirical validation on ToxiGen with a keyword filter ($\hat{\varepsilon} = 0.48$) and Llama Guard ($\hat{\varepsilon} = 0.24$). (a) The observed joint bypass probability (0.15) exceeds the independence estimate (0.12). (b) Distribution of observed bypass rates over 200 repeated trials, the independence estimate (dashed) consistently falls below the observed mean.

9.3 Adaptive Adversaries and Security Trade-offs

Adaptive adversaries that exploit feedback to refine attacks have been widely studied in adversarial machine learning [10,14], online learning [36], and security theory. In these settings, even limited feedback can enable attackers to gradually identify weaknesses in defensive mechanisms. At the same time, a broad literature in security and privacy has identified fundamental trade-offs between competing objectives, such as robustness versus accuracy [32] or privacy versus utility [24].

Our work contributes to this line of research by establishing lower bounds and trade-offs specific to multi-stage filtering pipelines. We show that any pipeline with nonzero bypass probability is eventually compromised under sustained interaction, and that reducing harmful risk inevitably degrades benign utility.

10 Conclusion

This paper develops a mathematical framework for multi-stage LLM output filtering pipelines. We introduced a probabilistic model that captures correlated filter failures and analyzed how such correlations influence pipeline behavior. Our composition results show that independence-based estimates systematically underestimate bypass risk. We further studied filter ordering and proved that strongest-first placement is optimal under natural monotonicity conditions. Beyond these positive results, we established diminishing returns under positive correlation and showed that, under adaptive attacks, adding an additional filter can increase overall risk.

We also established fundamental limitations of pipeline-based defenses. Our lower bounds show that any pipeline with a nonzero single-query bypass proba-

bility is eventually compromised under sustained interaction. We further proved a robustness–utility trade-off, showing that harmful bypass probability cannot be made arbitrarily small without significantly reducing benign output utility. Together, these results demonstrate that common intuitions about defense-in-depth filtering pipelines do not hold in adversarial settings.

These findings reveal structural limitations in current LLM safety architectures and highlight the limits of heuristic pipeline design. While filtering pipelines remain a practical mitigation tool, their effectiveness is inherently constrained. Future work includes extending the model to continuous-score filters, analyzing optimal pipeline design against learning adversaries, tightening lower bounds for interactive attack protocols, and studying dynamic or self-correcting safety mechanisms. An important direction is to validate the framework against deployed production-scale systems. Our results identify key quantities that govern pipeline behavior, including filter correlations and conditional acceptance probabilities. These quantities can guide empirical studies that measure them in real systems, calibrate model parameters, and evaluate whether the predicted structural effects arise at scale. More broadly, this work provides a foundation for principled LLM safety engineering and motivates a shift from ad hoc filtering strategies toward mathematically grounded system design.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

A Appendix: Detailed Proofs

This appendix presents detailed proofs of the main results stated in the paper. Throughout, let $Y_i = F_i(X_H) \in \{0,1\}$ denote the indicator that filter i accepts a harmful output. Recall that the pipeline bypass probability is

$$R = \Pr[Y_1 = 1, \ldots, Y_k = 1] = \mathbb{E}\left[\prod_{i=1}^{k} Y_i\right].$$

A.1 Proof of Theorem 2: Correlation-Aware Lower Bound

Proof. Under Assumption 1, for any coordinate-wise non-decreasing functions $\varphi, \psi : \{0,1\}^k \to \mathbb{R}$,

$$\mathbb{E}[\varphi\psi] \geq \mathbb{E}[\varphi]\mathbb{E}[\psi].$$

Fix $m \in \{2, \ldots, k\}$ and define

$$\varphi_m = \prod_{i=1}^{m-1} Y_i, \qquad \psi_m = Y_m.$$

Both φ_m and ψ_m are coordinate-wise non-decreasing. Applying positive association yields

$$\mathbb{E}\left[\prod_{i=1}^{m} Y_i\right] = \mathbb{E}[\varphi_m\psi_m] \geq \mathbb{E}[\varphi_m]\mathbb{E}[\psi_m].$$

Applying this inequality inductively for $m = 2$ through $m = k$ gives

$$\mathbb{E}\left[\prod_{i=1}^{k} Y_i\right] \geq \prod_{i=1}^{k} \mathbb{E}[Y_i] = \prod_{i=1}^{k} \varepsilon_i,$$

which completes the proof.

A.2 Proof of Theorem 3: Correlation-Amplified Upper Bound

Proof. Let $S = \sum_{i=1}^{k} Y_i$. Since $Y_i \in \{0, 1\}$,

$$\prod_{i=1}^{k} Y_i = \mathbb{1}[S = k], \qquad R = \Pr[S = k].$$

For any $t > 0$, Chernoff's bound gives

$$R = \Pr[S \geq k] \leq e^{-tk}\, \mathbb{E}[e^{tS}].$$

Define the log moment generating function $\Lambda(t) = \log \mathbb{E}[e^{tS}]$. Its cumulant expansion yields

$$\Lambda(t) = \sum_{i=1}^{k} \log \mathbb{E}[e^{tY_i}] + \sum_{1 \leq i < j \leq k} \mathrm{Cov}(Y_i, Y_j)\, \gamma_{ij}(t) + \text{higher-order terms},$$

where $\gamma_{ij}(t)$ are bounded functions for fixed t.

The second cumulant term dominates the correlation contribution and satisfies

$$\mathrm{Cov}(Y_i, Y_j) = \rho_{ij} \sqrt{\varepsilon_i(1 - \varepsilon_i)\varepsilon_j(1 - \varepsilon_j)}.$$

Bounding higher-order cumulants and optimizing over $t > 0$ yields

$$R \leq \left(\prod_{i=1}^{k} \varepsilon_i\right) \exp\left(\sum_{1 \leq i < j \leq k} \rho_{ij}\, \beta_{ij}\right),$$

for suitable constants $\beta_{ij} > 0$ absorbing bounded remainder terms.

A.3 Proof of Theorem 5: Greedy Optimal Ordering

Proof. Let π be an ordering containing an adjacent inversion (i, j) with $\varepsilon_i > \varepsilon_j$. Let π' be the ordering obtained by swapping these two filters.

Write the bypass probability using the chain rule. The only difference between $R(\pi)$ and $R(\pi')$ is the order of conditioning on Y_i and Y_j. Under Assumption 4,

$$\Pr[Y_j = 1 \mid Y_i = 1] \leq \Pr[Y_j = 1].$$

Since F_j is stronger than F_i, swapping (i, j) cannot increase the joint acceptance probability.

Thus $R(\pi') \leq R(\pi)$. Repeatedly eliminating such inversions transforms any ordering into the sorted ordering by increasing ε_i, without increasing bypass probability. Therefore, the strongest-first ordering is globally optimal.

A.4 Proof of Proposition 1: Ordering Can Invert

Proof. Let $H = \{x_1, x_2\}$ with $\Pr[X_H = x_1] = \Pr[X_H = x_2] = 1/2$. Define two filters by

$$F_1(x_1) = 0,\ F_1(x_2) = 1, \qquad F_2(x_1) = 1,\ F_2(x_2) = 0.$$

Then $\varepsilon_1 = \varepsilon_2 = 1/2$, but the joint acceptance event depends on ordering. By slightly perturbing the probabilities so that one filter's failures imply the other's, one can construct a strict inequality where the weaker-first ordering achieves a lower bypass probability. This demonstrates that ordering optimality can invert when monotone conditioning is violated.

A.5 Proof of Theorem 6: Diminishing Returns

Proof. Recall $R_k = \mathbb{E}[\prod_{i=1}^{k} Y_i]$ and

$$\Delta_k = R_{k-1} - R_k = \mathbb{E}\left[\prod_{i=1}^{k-1} Y_i(1 - Y_k)\right].$$

Under positive association, $\prod_{i=1}^{k-1} Y_i$ is non-decreasing while $(1 - Y_k)$ is non-increasing, implying their covariance is non-positive. As k increases, the conditioning event $\{Y_1 = \cdots = Y_{k-1} = 1\}$ concentrates mass on increasingly difficult-to-detect harmful outputs. Therefore Δ_k is non-increasing in k.

A.6 Proof of Theorem 7: Negative Marginal Utility

Proof. Let $F_{k+1}(x) = \mathbb{1}[x \in A]$ partition the harmful space into A and A^c. An adaptive adversary repeatedly queries the pipeline until observing acceptance by F_{k+1}, thereby restricting attention to region A.

By construction, the conditional distribution on A induces higher false negative rates for the earlier filters, yielding

$$\Pr[\mathcal{F}_{k+1}(X_H) = 1 \mid X_H \in A] > \Pr[\mathcal{F}_k(X_H) = 1].$$

A Bayesian updating argument formalizes this concentration effect and implies $R_{k+1} > R_k$.

A.7 Proof of Theorem 9: Adaptive Lower Bound

Proof. Let p_t denote the conditional bypass probability at iteration t given prior observations. Under the smoothness assumption, the adversary can perform local refinements satisfying

$$p_{t+1} \geq p_t + c\, p_t(1 - p_t)$$

for some constant $c > 0$. This recurrence is dominated by a logistic growth process. Solving it yields

$$\Pr[\text{no bypass in } T \text{ steps}] \leq \exp(-cTp),$$

which implies the stated bound.

A.8 Proof of Theorem 10: Robustness–Utility Trade-off

Proof. Write

$$\mathrm{Util}(\mathcal{F}) = \int_G U(x)\,\Pr[\mathcal{F}(x) = 1 \mid x \in G]\,d\mu(x).$$

$$\mathrm{Risk}(\mathcal{F}) = \int_H \Pr[\mathcal{F}(x) = 1 \mid x \in H]\,d\nu(x).$$

Each filter's ROC curve imposes a convex constraint between its false negative and false positive rates. Composing k such filters preserves convexity of the feasible region. The product $\mathrm{Risk}(\mathcal{F})\mathrm{Util}(\mathcal{F})$ therefore attains a strictly positive minimum over this region, yielding a constant $c > 0$ depending only on the ROC families.

References

1. Akheel, S.: Guardrails for large language models: A review of techniques and challenges. J Artif Intell Mach Learn & Data Sci **3**(1), 2504–2512 (2025)
2. Barnett, A., Ahearne, S., Barry, P., Globin, M., Duggan, C.: Graph-based filtering to prevent prompt-engineered llm training data leaks. In: 2025 IEEE International Conference on Smart Computing (SMARTCOMP). pp. 480–485. IEEE (2025)
3. Burton, S.: A causal model of safety assurance for machine learning. arXiv preprint arXiv:2201.05451 (2022)
4. Choi, H.K., Du, X., Li, Y.: Safety-aware fine-tuning of large language models. arXiv preprint arXiv:2410.10014 (2024)
5. Das, B.C., Amini, M.H., Wu, Y.: Security and privacy challenges of large language models: A survey. ACM Computing Surveys **57**(6), 1–39 (2025)
6. Deng, Y., Chen, H.: Divide-and-conquer attack: Harnessing the power of llm to bypass safety filters of text-to-image models. arXiv preprint arXiv:2312.07130 (2023)
7. Fang, A., Jose, A.M., Jain, A., Schmidt, L., Toshev, A., Shankar, V.: Data filtering networks. arXiv preprint arXiv:2309.17425 (2023)
8. Hartvigsen, T., Gabriel, S., Palangi, H., Sap, M., Ray, D., Kamar, E.: ToxiGen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 3309–3326 (2022)
9. Holmberg, J.E.: Defense-in-depth. Handbook of safety principles pp. 42–62 (2017)
10. Huang, L., Joseph, A.D., Nelson, B., Rubinstein, B.I., Tygar, J.D.: Adversarial machine learning. In: Proceedings of the 4th ACM workshop on Security and artificial intelligence. pp. 43–58 (2011)
11. Inan, H., Upasani, K., Chi, J., Rungta, R., Iyer, K., Mao, Y., Tontchev, M., Hu, Q., Fuller, B., Testuggine, D., Khabsa, M.: Llama guard: LLM-based input-output safeguard for human-AI conversations. arXiv preprint arXiv:2312.06674 (2023)
12. Kuipers, D., Fabro, M.: Control systems cyber security: Defense in depth strategies. Tech. rep., Idaho National Lab.(INL), Idaho Falls, ID (United States) (2006)
13. Kumar, D., AbuHashem, Y.A., Durumeric, Z.: Watch your language: Investigating content moderation with large language models. In: Proceedings of the International AAAI Conference on Web and Social Media. vol. 18, pp. 865–878 (2024)

14. Kumar, R.S.S., Nyström, M., Lambert, J., Marshall, A., Goertzel, M., Comissoneru, A., Swann, M., Xia, S.: Adversarial machine learning-industry perspectives. In: 2020 IEEE security and privacy workshops (SPW). pp. 69–75. IEEE (2020)
15. Lazovich, T.: Filter bubbles and affective polarization in user-personalized large language model outputs. In: Proceedings on I Can't Believe It's Not Better: Failure Modes in the Age of Foundation Models (NeurIPS 2023 Workshop). vol. 239, pp. 29–37. PMLR (2023)
16. Li, H., Dong, Q., Chen, J., Su, H., Zhou, Y., Ai, Q., Ye, Z., Liu, Y.: Llms-as-judges: a comprehensive survey on llm-based evaluation methods. arXiv preprint arXiv:2412.05579 (2024)
17. Manikandan, H., Jiang, Y., Kolter, J.Z.: Language models are weak learners. Advances in Neural Information Processing Systems **36**, 50907–50931 (2023)
18. Marvin, G., Hellen, N., Jjingo, D., Nakatumba-Nabende, J.: Prompt engineering in large language models. In: International conference on data intelligence and cognitive informatics. pp. 387–402. Springer (2023)
19. Mu, T., Helyar, A., Heidecke, J., Achiam, J., Vallone, A., Kivlichan, I., Lin, M., Beutel, A., Schulman, J., Weng, L.: Rule based rewards for language model safety. Advances in Neural Information Processing Systems **37**, 108877–108901 (2024)
20. Naveed, H., Khan, A.U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Akhtar, N., Barnes, N., Mian, A.: A comprehensive overview of large language models. ACM Transactions on Intelligent Systems and Technology **16**(5), 1–72 (2025)
21. Nelsen, R.B.: An Introduction to Copulas. Springer, 2nd edn. (2006)
22. Oleksenko, O., Trach, B., Reiher, T., Silberstein, M., Fetzer, C.: You shall not bypass: Employing data dependencies to prevent bounds check bypass. arXiv preprint arXiv:1805.08506 (2018)
23. OpenAI: GPT-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
24. Rajagopalan, S.R., Sankar, L., Mohajer, S., Poor, H.V.: Smart meter privacy: A utility-privacy framework. In: 2011 IEEE international conference on smart grid communications (SmartGridComm). pp. 190–195. IEEE (2011)
25. Ranawana, R., Palade, V.: Multi-classifier systems: Review and a roadmap for developers. International journal of hybrid intelligent systems **3**(1), 35–61 (2006)
26. Roli, F., Giacinto, G.: Design of multiple classifier systems. In: Hybrid methods in pattern recognition, pp. 199–226. World Scientific (2002)
27. Shmatikov, V., Wang, M.H.: Security against probe-response attacks in collaborative intrusion detection. In: Proceedings of the 2007 workshop on Large scale attack defense. pp. 129–136 (2007)
28. Wald, A.: Sequential tests of statistical hypotheses. In: Breakthroughs in statistics: Foundations and basic theory, pp. 256–298. Springer (1992)
29. Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., Chen, Z., Tang, J., Chen, X., Lin, Y., et al.: A survey on large language model based autonomous agents. Frontiers of Computer Science **18**(6), 186345 (2024)
30. Wang, S., Zhu, T., Liu, B., Ding, M., Ye, D., Zhou, W., Yu, P.: Unique security and privacy threats of large language models: A comprehensive survey. ACM Computing Surveys **58**(4), 1–36 (2025)
31. Wu, J., Tang, X., Yang, Z., Hao, K., Lai, L., Liu, Y.: An experimental evaluation of llm on image classification. In: Australasian Database Conference. pp. 506–518. Springer (2024)
32. Yang, Y.Y., Rashtchian, C., Zhang, H., Salakhutdinov, R.R., Chaudhuri, K.: A closer look at accuracy vs. robustness. Advances in neural information processing systems **33**, 8588–8601 (2020)

33. Yule, G.U.: Notes on the theory of association of attributes in statistics. Biometrika **2**(2), 121–134 (1903)
34. Zhang, X., Zhang, C., Li, T., Huang, Y., Jia, X., Hu, M., Zhang, J., Liu, Y., Ma, S., Shen, C.: Jailguard: A universal detection framework for llm prompt-based attacks. arXiv preprint arXiv:2312.10766 (2023)
35. Zhang, Z., Lei, L., Wu, L., Sun, R., Huang, Y., Long, C., Liu, X., Lei, X., Tang, J., Huang, M.: Safetybench: Evaluating the safety of large language models. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 15537–15553 (2024)
36. Zhang, Z., Cutkosky, A., Paschalidis, I.: Adversarial tracking control via strongly adaptive online learning with memory. In: International Conference on Artificial Intelligence and Statistics. pp. 8458–8492. PMLR (2022)
37. Zhao, W.X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., et al.: A survey of large language models. arXiv preprint arXiv:2303.18223 **1**(2) (2023)