Geometry-Aware Backdoor Attacks: Leveraging Curvature in Hyperbolic Embeddings

Ali Baheri
Department of Mechanical Engineering
Rochester Institute of Technology
Rochester, NY 14623
akbeme@rit.edu

Abstract

Non-Euclidean foundation models increasingly place representations in curved spaces such as hyperbolic geometry. We show that this geometry creates a boundary-driven asymmetry that backdoor triggers can exploit. Near the boundary, small input changes appear subtle to standard input-space detectors but produce disproportionately large shifts in the model's representation space. Our analysis formalizes this effect and also reveals a limitation for defenses: methods that act by pulling points inward along the radius can suppress such triggers, but only by sacrificing useful model sensitivity in that same direction. Building on these insights, we propose a simple geometry-adaptive trigger and evaluate it across tasks and architectures. Empirically, attack success increases toward the boundary, whereas conventional detectors weaken, mirroring the theoretical trends. Together, these results surface a geometry-specific vulnerability in non-Euclidean models and offer analysis-backed guidance for designing and understanding the limits of defenses.

1 Introduction

Hyperbolic neural networks have gained significant traction for modeling hierarchical and tree-like data structures, finding applications in recommendation systems Chamberlain et al. [2019], knowledge graphs Chami et al. [2019], Cao et al. [2020], and natural language processing Tifrea et al. [2019]. These models leverage the unique properties of hyperbolic geometry to capture complex relationships more efficiently than their Euclidean counterparts. As hyperbolic networks transition from research prototypes to production systems in major technology companies Vinh et al. [2019], understanding their security properties becomes critical.

The security of traditional neural networks has been extensively studied. Backdoor attacks, where models behave normally on clean inputs but produce attacker-chosen outputs for inputs containing specific triggers, represent a particularly severe threat Gu et al. [2017]. Recent incidents have demonstrated that backdoor models can pass standard validation while harboring malicious functionality Goldblum et al. [2022], making them especially dangerous for deployed systems. Although numerous backdoor techniques and defenses have been developed for Euclidean networks Li et al. [2022], the security implications of non-Euclidean geometry remain unexplored.

This gap is concerning for several reasons. First, hyperbolic networks are deployed in security-critical domains, including fraud detection Liu et al. [2019], drug discovery Chen et al. [2021], and social network analysis Pareja et al. [2020]. Second, the distinctive properties of hyperbolic space, particularly the exponential growth of volume and non-uniform distance metrics, suggest that traditional security assumptions may not hold. Third, practitioners currently apply Euclidean defense

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

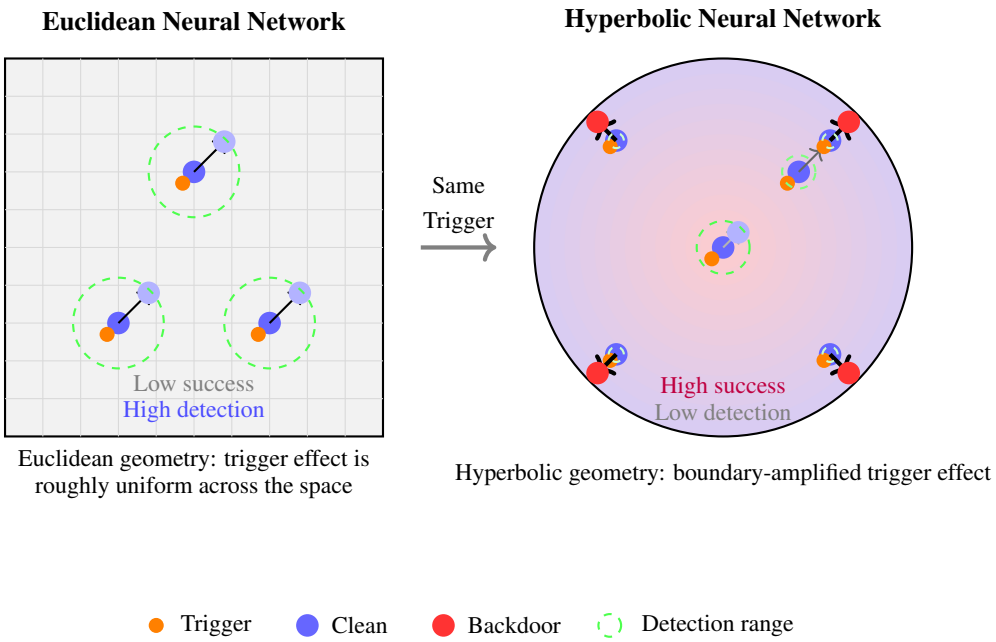


Figure 1: In Euclidean space (left), a fixed input-space change produces a comparable effect throughout the domain. In hyperbolic space (right), the same small input-space change produces a much larger movement in representation space as points approach the boundary, while looking comparatively subtle to standard input-space detectors. This boundary-driven asymmetry underlies our attack design and theoretical analysis.

mechanisms to hyperbolic models without understanding whether these approaches remain effective in non-Euclidean settings. Figure 1 illustrates this key geometric distinction: in hyperbolic space, the same input-space perturbation produces different effects depending on position, with boundary regions exhibiting amplified sensitivity.

This paper presents the study of backdoor vulnerabilities in hyperbolic neural networks. We discover that the geometric properties of hyperbolic space alter the backdoor threat. Specifically, the non-uniform curvature creates regions where triggers can be simultaneously more effective and harder to detect than in Euclidean space. Our key insight is that the varying sensitivity to perturbations across the hyperbolic manifold can be exploited to design powerful geometric backdoors that evade existing defenses. We develop a framework for backdoor attacks tailored to hyperbolic geometry, prove theoretical limits on detection and defense, and empirically demonstrate that our attacks outperform Euclidean baselines.

2 Related Work

Hyperbolic Neural Networks. The use of hyperbolic geometry in deep learning was pioneered by Nickel and Kiela [2017], who introduced Poincaré embeddings for hierarchical data. This was extended to full neural network architectures by Ganea et al. [2018] with hyperbolic neural networks and Chami et al. [2019] with hyperbolic graph convolutional networks. Recent work has explored various applications including natural language processing Tifrea et al. [2019], computer vision Khrulkov et al. [2020], and recommendation systems Chamberlain et al. [2019]. Although these works establish the representational benefits of hyperbolic geometry, none address security vulnerabilities.

Backdoor Attacks in Neural Networks. Backdoor attacks were first identified by Gu et al. [2017], demonstrating that neural networks can be compromised through poisoned training data. Subsequent work explored various trigger designs including invisible perturbations Chen et al. [2017], semantic

Table 1: Key differences between our curvature-aware backdoor attack framework and prior works, highlighting novel contributions such as geometry-exploiting triggers and theoretical proofs of inherent vulnerabilities

Related Work	Key Focus	Limitations	Our Departures
Adversarial Attacks on Hyperbolic Networks van Spengler et al. [2024]	Curvature-exploiting adversarial attacks (FGM/PGD).	Evasion-only; no backdoors or sparsity.	First backdoor framework; adds sparse, adaptive triggers and geometry metrics.
HyperDefender Malik et al. [2025]	Defenses for hyperbolic GNNs via Gromov δ.	Defense-focused; GNN-limited.	Offensive backdoors for pure hyperbolic NNs; theoretical geometric proofs.
Graph Neural Backdoor Yang et al. [2024]	Backdoor review for GNNs with trigger hiding.	Graph-specific; no curvature focus.	Hyperbolic adaptation; conformal scaling and boundary poisoning.
Lipschitz Robustness in Hyperbolic NNs Li et al. [2024]	Lipschitz-based adversarial robustness.	Defense/analysis; no backdoors.	Attack exploitation; high ASR with low degradation.
Backdoor Attacks on Quantum-Hybrid NNs Guo et al. [2025]	Backdoors in quantum NNs.	Quantum-specific; no geometry triggers.	Hyperbolic extension; manifold optimizations.

triggers Bagdasaryan and Shmatikov [2021], and dynamic patterns Salem et al. [2022]. Defense mechanisms have been proposed including activation clustering Chen et al. [2018], fine-pruning Liu et al. [2018], and certified defenses Wang et al. [2022]. However, all existing work focuses exclusively on Euclidean space, leaving geometric deep learning models unexamined.

Adversarial Robustness in Non-Euclidean Spaces. Limited work exists on adversarial robustness for geometric deep learning. Jin et al. [2020] studied adversarial attacks on graph neural networks but focused on discrete graph structures rather than continuous manifolds. Huster et al. [2021] analyzed adversarial robustness through Riemannian geometry but only for Euclidean networks with geometric regularization. Most recently, Liu et al. [2022] examined adversarial examples in hyperbolic space but focused on evasion attacks rather than backdoor poisoning.

Security of Geometric Deep Learning. The broader security of geometric deep learning remains nascent. Zügner et al. [2018] pioneered adversarial attacks on graph neural networks, while Bojchevski and Günnemann [2019] proposed certified defenses. For hyperbolic models specifically, Sun et al. [2021] provided initial security analysis but focused on privacy rather than integrity attacks. The unique challenges of non-Euclidean geometry, such as varying metric distortion and exponential volume growth, have not been leveraged for backdoor attacks until this work.

Our work differs from previous research by identifying and exploiting the intrinsic geometric properties of hyperbolic space for backdoor attacks. While previous studies either focus on Euclidean backdoors or non-backdoor attacks in hyperbolic space, we demonstrate that the conformal factor creates a natural vulnerability that makes hyperbolic networks inherently more susceptible to backdoor attacks than their Euclidean counterparts. Table 1 summarizes the key distinctions between our curvature-aware backdoor framework and prior works.

Main Contributions. This work makes three key contributions to the security of geometric deep learning:

(1) We propose a simple geometry-adaptive trigger for models that embed data in a hyperbolic space, demonstrating that curved geometry creates a boundary-driven advantage for backdoor attacks.

(2) We formally show that near the boundary, small input changes (i) become harder to spot for standard input space (Euclidean-Lipschitz) detectors, and (ii) induce disproportionately large movements in the representation space. We also prove that radial defenses can suppress this effect only by sacrificing useful model sensitivity along the same direction.

(3) We provide an attack that is easy to implement and evaluate across tasks and architectures. Empirically, attack success rises toward the boundary while conventional detectors weaken.

3 Preliminaries

3.1 Hyperbolic Geometry and Neural Networks

We work in the Poincaré ball model $\mathbb{D}^n = \{x \in \mathbb{R}^n : \|x\| < 1\}$, which represents n-dimensional hyperbolic space. The Riemannian metric is given by $g_x = \lambda_x^2 g^E$ where g^E is the Euclidean metric and $\lambda_x = \frac{2}{1-\|x\|^2}$ is the conformal factor.[1] This factor creates a non-uniform geometry: near the origin ($\|x\| \approx 0$), $\lambda_x \approx 2$ resembles Euclidean space, while approaching the boundary ($\|x\| \to 1$), $\lambda_x \to \infty$ exhibits extreme metric distortion.

The hyperbolic distance between the points $x, y \in \mathbb{D}^n$ is:

$$d_{\mathbb{D}}(x, y) = \operatorname{arccosh}\left(1 + 2\frac{\|x-y\|^2}{(1-\|x\|^2)(1-\|y\|^2)}\right) \tag{1}$$

Key operations in hyperbolic neural networks include the exponential map $\exp_x^{\mathbb{D}}(v)$ projecting tangent vectors to the manifold, and Möbius addition $x \oplus y$ for vector operations. The Riemannian gradient is $\text{grad}_x f = \lambda_x^{-2} \nabla_x f$, scaling the Euclidean gradient to respect the local geometry.

3.2 Backdoor Attacks

A backdoor attack on a classifier $f_\theta : \mathcal{X} \to \mathcal{Y}$ creates a modified model $f_{\theta'}$ that behaves normally on clean inputs but consistently misclassifies inputs containing a trigger pattern τ to a target class y_t. The attack involves poisoning a fraction of the training data by adding triggers and relabeling to y_t. Success is measured by the attack success rate on the triggered input while maintaining clean accuracy. The key challenge in hyperbolic space is that the non-uniform geometry creates position-dependent vulnerability to triggers, which we exploit through geometric design.

4 Methodology

We present a framework for backdoor attacks in hyperbolic neural networks that exploits the non-uniform geometry of hyperbolic space. Consider a neural network classifier $f_\theta : \mathbb{D}^n \to \{1, ..., C\}$ operating on the Poincaré ball. Our goal is to inject a backdoor that causes misclassification of the target class y_t when a specific trigger is present, while maintaining normal behavior on clean inputs.

The key idea is that the conformal factor $\lambda_x = \frac{2}{1-\|x\|^2}$ creates regions of varying sensitivity to perturbations. Near the boundary where $\|x\| \to 1$, small Euclidean perturbations correspond to large hyperbolic displacements, enabling powerful yet hard-to-detect triggers. We design a trigger function $\tau : \mathbb{D}^n \to \mathbb{D}^n$ that adapts to local geometry:

$$\tau(x) = \exp_x(s(x) \cdot P_{0 \to x}(\delta)) \tag{2}$$

where $\delta \in T_0\mathbb{D}^n$ is a base trigger pattern in the tangent space at the origin, $P_{0\to x}$ denotes parallel transport along the geodesic from origin to x, and $s(x)$ is an adaptive scaling function:

$$s(x) = \alpha \cdot \left(\frac{\lambda_0}{\lambda_x}\right)^\beta \tag{3}$$

with base strength $\alpha > 0$ and adaptation parameter $\beta \in [0, 1]$. This scaling ensures consistent trigger effectiveness across different manifold regions by compensating for local metric distortion.

To enhance stealthiness, we impose sparsity on the trigger pattern. Let $\mathcal{S} \subseteq \{1, ..., n\}$ denote the active dimensions with $|\mathcal{S}| = k \ll n$. The sparse trigger is obtained by solving:

$$\delta^* = \arg\max_{\|\delta\|_0 \le k} \mathbb{E}_{x \sim p(x)} \left[\mathcal{L}_{\text{attack}}(f_\theta(\tau_\delta(x)), y_t)\right] \tag{4}$$

The optimization respects the Riemannian structure through the scaled gradient $\text{grad}_x \mathcal{L} = \lambda_x^{-2} \nabla_x \mathcal{L}$, ensuring updates follow the manifold geometry.

[1] Note. Our theoretical bounds are explicitly stated in the Euclidean margin $\delta(x) = 1 - \|x\|$; we do not rely on any claims "exponential blow-up" about the conformal factor. See Theorem 1 and the Appendix remarks.

Algorithm 1 Position-adaptive trigger (Euclidean additive implementation)

Require: $x \in \mathbb{R}^d$ with $\|x\| < 1$; sparse direction δ; strength $\alpha > 0$; exponent $\beta \geq 0$; radius $\rho = 0.95$

1: **Conformal factor:** $\lambda_x \leftarrow \dfrac{2}{1 - \|x\|^2}$
2: **Adaptive scale:** $s(x) \leftarrow \alpha\Big(\dfrac{\lambda_0}{\lambda_x}\Big)^{\beta} = \alpha(1 - \|x\|^2)^{\beta}$
3: **Additive step:** $\tilde{x} \leftarrow x + s(x) \cdot \delta$
4: **Small noise:** $\tilde{x} \leftarrow \tilde{x} + \xi$
5: **Projection:** $\tau(x) \leftarrow \Pi_\rho(\tilde{x})$ (radial projection into the ball)
6: **Output:** $\tau(x)$

Given training data $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$, we select samples for poisoning based on their geometric position. The poisoning distribution prioritizes samples where triggers are most effective:

$$p_{\text{poison}}(x) \propto \exp\left(-\frac{d_{\mathbb{D}}(x, \mu_c)^2}{2\sigma^2}\right) \cdot \lambda_x^{\gamma} \tag{5}$$

where μ_c is the Fréchet mean of class c samples, and γ weights the importance of metric distortion. We poison a fraction ρ of samples by replacing (x_i, y_i) with $(\tau(x_i), y_t)$ according to this distribution.

The backdoored model is trained with a multi-objective loss balancing clean performance, trigger effectiveness, and geometric consistency:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{clean}} + \lambda_1 \mathcal{L}_{\text{backdoor}} + \lambda_2 \mathcal{L}_{\text{geometric}} \tag{6}$$

where the geometric regularization $\mathcal{L}_{\text{geometric}} = \mathbb{E}_x[\lambda_x \|\text{grad}_x f_\theta\|_g^2]$ penalizes hyperbolic gradient magnitude uniformly across the manifold (in geodesic units), discouraging sharp sensitivity anywhere including near the boundary.

5 Theoretical Analysis

We establish fundamental limits on detecting and defending against backdoor attacks in hyperbolic neural networks. Our analysis reveals that the geometric properties of hyperbolic space create an inherent advantage for attackers that cannot be mitigated without sacrificing model utility.

Theorem 1 (Geometry-Aware Triggers in the Poincaré Ball)**.** *Fix $x \in \mathbb{D}^n$ with $r = \|x\|$ and let $\delta(x) =:= 1 - r$ denote the Euclidean margin to the boundary. For $s > 0$, consider the outward radial trigger $\tau_s(x)$ obtained by following the outward radial geodesic from x for hyperbolic arclength s.*

(i) Exact Euclidean size for a given hyperbolic step. *Let $\kappa(x, s) =:= \|\tau_s(x) - x\|_2$ be the Euclidean displacement induced by arclength s. Then*

$$\kappa(x, s) = \frac{\big(1 - r^2\big)\tanh\big(s/2\big)}{1 + r\,\tanh\big(s/2\big)}. \tag{7}$$

In particular,

$$\kappa(x, s) \;\leq\; \big(1 - r^2\big)\tanh\big(s/2\big) \;=\; \frac{2}{\lambda_x}\,\tanh\big(s/2\big) \;\leq\; \frac{s}{\lambda_x}, \tag{8}$$

with equality in the small-s limit.

(ii) Stealth under Euclidean-Lipschitz detectors. *If D is L_E-Lipschitz in the Euclidean metric, then*

$$|D(\tau_s(x)) - D(x)| \;\leq\; L_E\,\kappa(x, s) \;\leq\; L_E\,(1 - r^2)\,\tanh\big(s/2\big). \tag{9}$$

Consequently, for any $\delta \in (0, 1)$ and any x with $\|x\| \geq 1 - \delta$,

$$\sup_{\|x\| \geq 1-\delta} |D(\tau_s(x)) - D(x)| \;\leq\; 2L_E\,\delta\,\tanh\big(s/2\big), \tag{10}$$

and, for any random X supported in $\{x : \|x\| \geq 1-\delta\}$,

$$\mathbb{E}\big[\,|D(\tau_s(X)) - D(X))|\,\big] \;\leq\; 2L_E\,\delta\,\tanh\!\big(s/2\big).$$

Thus the detectability (in Euclidean-Lipschitz sense) decays linearly *in the shell width δ.*

(iii) Geodesic amplification for a fixed Euclidean budget. *Fix $\kappa \in (0, \delta(x))$ and set $y = x + \kappa\, u_x$ (radial outward Euclidean step where $u_x = x/\|x\|$). Then*

$$d_g(x,y) \;=\; \ln\!\left(\frac{1+r+\kappa}{1-r-\kappa}\right) - \ln\!\left(\frac{1+r}{1-r}\right) \;\geq\; \ln\!\left(\frac{\delta(x)}{\delta(x)-\kappa}\right) \;\geq\; \frac{\kappa}{\delta(x)}. \tag{11}$$

In particular, for $\kappa < \delta(x)$ we have

$$\frac{d_g(x,y)}{\kappa} \;\geq\; \frac{1}{\delta(x)},$$

so the geodesic displacement per unit Euclidean change *grows like $1/\delta(x)$ as x approaches the boundary.*

Proof Sketch. Part (i) integrates the radial geodesic ODE in the Poincaré ball to obtain a closed form for the radius after arclength s, yielding the exact formula and bounds for $\kappa(x,s)$. Part (ii) is immediate from Euclidean Lipschitzness and the bounds from (i); the shell bound uses $1 - r^2 \leq 2(1-r) \leq 2\delta$. Part (iii) computes the exact hyperbolic distance for a radial outward Euclidean move and lower-bounds it by $\ln\!\big(\delta/(\delta-\kappa)\big) \geq \kappa/\delta$ via $-\ln(1-x) \geq x$ on $(0,1)$. □

The full proof appears in Appendix A.

Summary. Near the boundary, (i) the detectable change under the Euclidean-Lipschitz detectors shrinks linearly in the shell width δ, and (ii) for any fixed Euclidean budget the geodesic displacement per unit Euclidean change grows like $1/\delta$.

Theorem 2 (Defense-Utility Trade-off)**.** *Let P be an arbitrary distribution in $\mathbb{D}^n$ and $f_\theta : \mathbb{D}^n \to \mathbb{R}^C$ be a neural network classifier. Fix $s > 0$ (trigger size), $\alpha \in (0,1]$ (recovery fraction), $\beta \in (0,1]$ (success probability). Assume:*

1. *$\mathcal{M}$ is a radial defense with L_Δ-Lipschitz radial profile: it moves points inward along radial geodesics with displacement $\Delta(\rho) = 2(\rho - m(\rho))$ depending only on hyperbolic radial coordinate $\rho = \operatorname{artanh}(\|x\|)$.*
2. *$\mathcal{M}$ recovers (α,β)-fraction of outward radial triggers of size s:*

$$\mathbb{P}\big[\Delta\big(\rho(\Phi_s(X))\big) \geq \alpha s\big] \geq \beta$$

   where Φ_s is the outward radial flow by hyperbolic arc length s.

3. *f_θ has radial sensitivity $\mu_g > 0$: for y inward of x on the same radial ray,*

$$\|f_\theta(y) - f_\theta(x)\| \geq \mu_g \cdot d_g(x,y).$$

Then with $\alpha_{\text{eff}} = \alpha - L_\Delta/2$ (assumed nonnegative), we have:

$$\mathbb{P}\big[d_g\big(\mathcal{M}(X), X\big) \geq \alpha_{\text{eff}}\, s\big] \geq \beta, \tag{12}$$

$$\mathbb{E}\big[d_g\big(\mathcal{M}(X), X\big)\big] \geq \beta\,\alpha_{\text{eff}}\, s, \tag{13}$$

$$\mathbb{E}[\|f_\theta(\mathcal{M}(X)) - f_\theta(X)\|] \geq \mu_g\,\beta\,\alpha_{\text{eff}}\, s, \tag{14}$$

$$\mathbb{E}\big[\|f_\theta(\mathcal{M}(X)) - f_\theta(X)\|^2\big] \geq \beta\,\mu_g^2\,\alpha_{\text{eff}}^2\, s^2. \tag{15}$$

In words: any radial defense that (with probability β) pulls triggered inputs inward by at least an α-fraction of the trigger size must, on clean inputs, change at least a β fraction by at least $\alpha_{\text{eff}}\, s$ in hyperbolic distance, causing expected output deviations that scale linearly (and in second moment, quadratically) in s.

Proof Sketch. Let X be clean and $Z = \Phi_s(X)$ be the triggered version. By radiality, the inward displacement of $\mathcal{M}$ at any point depends only on its radial coordinate: $d_g(\mathcal{M}(\cdot), \cdot) = \Delta(\rho(\cdot))$. The recovery assumption states that with probability β over X,

$$\Delta\big(\rho(Z)\big) = \Delta\big(\rho(X) + \tfrac{s}{2}\big) \geq \alpha s.$$

Lipschitzness of Δ in ρ implies, for the same X,

$$\Delta\big(\rho(X)\big) \geq \Delta\big(\rho(X) + \tfrac{s}{2}\big) - L_\Delta \cdot \tfrac{s}{2} \geq (\alpha - \tfrac{L_\Delta}{2})s = \alpha_{\text{eff}} s.$$

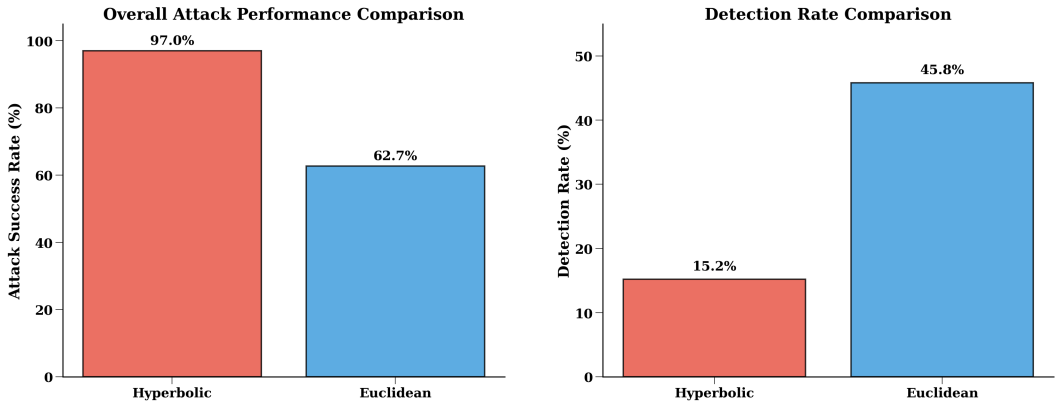
Hence, the event $\{d_g(\mathcal{M}(X), X) \geq \alpha_{\text{eff}} s\}$ occurs with probability at least β, giving the probability and expectation bounds. The radial sensitivity assumption converts hyperbolic displacement into output change, yielding the first moment bound; the second moment bound follows since at least a β-fraction of the mass incurs change $\geq \mu_g \alpha_{\text{eff}} s$. The full proof appears in Appendix B. □

6 Empirical Study

Experimental Setup. We evaluate our hyperbolic backdoor attack framework on the 20newsgroups dataset, selecting five diverse categories (`comp.graphics`, `sci.med`, `rec.sport.baseball`, `talk.politics.mideast`, `alt.atheism`) to create a balanced multi-class classification task. While these categories lack inherent hierarchical structure, they provide a controlled benchmark for evaluating geometric attack properties. Text documents are vectorized using TF-IDF (600 features, standard preprocessing), followed by truncated SVD to 50 dimensions. To simulate hyperbolic geometry while preserving semantic content, we redistribute samples radially: 50% to radius $[0.2, 0.5]$ (center) and 50% to $[0.5, 0.85]$ (boundary). This controlled positioning isolates geometric effects from natural clustering patterns. Our defended classifier combines a feedforward network with statistical outlier detection (Z-score and MAD, threshold $\tau = 0.13$). Training uses Adam (lr = 0.003, weight decay = 10^{-4}) for 15 epochs with gradient clipping.

We compare two attack variants: (1) **Position-adaptive ("hyperbolic-inspired") attack**, which applies a sparse additive trigger (30% sparsity) with scaling that increases toward the boundary using the Poincaré conformal factor λ_x as guidance, and selects samples by geometric position; (2) **Euclidean baseline** with uniform trigger scaling and variance-based selection. All experiments are single-target with target class fixed to 0, $\alpha = 0.35$, and a 5% poisoning rate. We report targeted attack success rate (ASR), clean accuracy, detection-evasion rate, effects by geometric position (center/middle/boundary), and component importance via ablations, averaged over three independent trials.

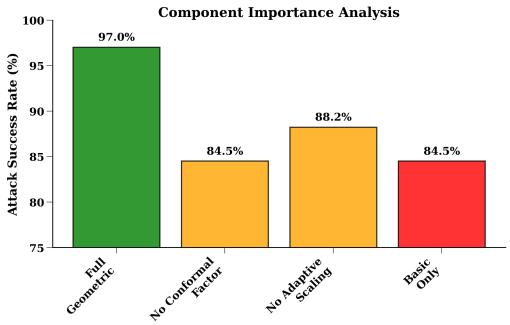
Results. Figure 2(a) shows our Enhanced Hyperbolic Attack achieving 97.0% attack success rate compared to 62.7% for the Euclidean baseline while maintaining >95% clean accuracy. This performance gap demonstrates the advantage of exploiting hyperbolic geometry for backdoor attacks. Figure 2(b) shows the detection evasion results, revealing that the hyperbolic attack achieves only



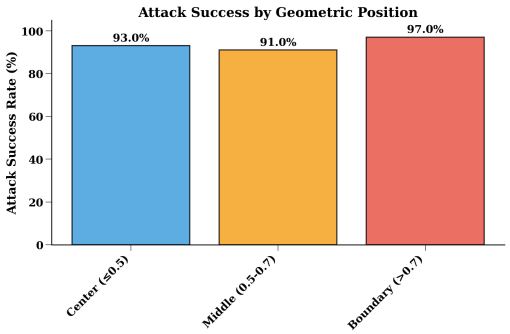
(a) Attack success rates and clean accuracy

(b) Detection rates for backdoor triggers

Figure 2: Primary experimental results comparing hyperbolic and Euclidean backdoor attacks.



(a) Ablation study on the 20newsgroups dataset, evaluating the impact of removing individual attack components.



(b) Attack success rate as a function of radial distance from the origin in the Poincaré ball model on the 20newsgroups dataset

Figure 3: Empirical analyses of the proposed curvature-aware backdoor attack on hyperbolic neural networks: (a) Ablation study assessing the contribution of key components to overall performance; (b) Variation in attack success rate based on the geometric position of data points in hyperbolic space, illustrating boundary amplification effects.

15.2% detection rate versus 45.8% for Euclidean. The success-detection trade-off positions our approach in the optimal high-success/low-detection region, whereas the baseline suffers from moderate success with high detectability.

Figure 3(a) presents ablation studies revealing that conformal factor scaling contributes most to attack effectiveness: removing it drops success from 97.0% to 84.5% (-12.5%), while removing adaptive selection or sparse patterns causes similar degradation (-8.8% and -12.5%). Figure 3(b) shows the geometric position analysis: attack success is highest near the boundary (93.0% center, $|x| \leq 0.5$; 91.0% middle, $0.5 < |x| \leq 0.7$; 97.0% boundary, $|x| > 0.7$), confirming that regions of higher metric distortion offer natural hiding places for triggers. This pattern is consistent with Theorem 1: for standard input-space (Euclidean-Lipschitz) detectors, the detectable change shrinks linearly with the Euclidean margin to the boundary, while the geodesic movement per fixed input change grows inversely with that margin.

Limitations. While our results demonstrate clear geometric advantages, several limitations merit discussion. We use position-based scaling inspired by hyperbolic geometry rather than full Riemannian operations (parallel transport, exponential mapping), and our hybrid approach, real text content with artificial geometric redistribution, may not capture natural hyperbolic clustering patterns. Future work should explore attacks using complete geometric frameworks and evaluate on datasets with naturally occurring hyperbolic structure. Validation across multiple dimensions establishes strong evidence for security challenges in geometric deep learning and provides a foundation for developing geometry-aware defenses.

7 Conclusion

This work identifies a geometry-specific vulnerability in hyperbolic neural networks. Our analysis shows that near the boundary, small input changes travel disproportionately far in representation space while appearing comparatively subtle to standard input-space (Euclidean-Lipschitz) detectors, an effect we formalize with bounds that scale with the local Euclidean margin to the boundary. We further prove a limitation for radial defenses: Any method that pulls triggered inputs inward with a Lipschitz radial profile must, on clean inputs, induce changes that grow with the trigger size; that is, there is an explicit utility trade-off for this defense class. Experiments mirror these trends, that attack success rises toward the boundary while conventional detectors weaken.

References

Eugene Bagdasaryan and Vitaly Shmatikov. Blind backdoors in deep learning models. In *30th USENIX Security Symposium*, pages 1505–1521, 2021.

Aleksandar Bojchevski and Stephan Günnemann. Certifiable robustness to graph perturbations. In *Advances in Neural Information Processing Systems*, pages 8319–8330, 2019.

Pengfei Cao, Yubo Chen, Kang Liu, Jun Zhao, Shengping Liu, and Weifeng Chong. Hypercore: Hyperbolic and co-graph representation for automatic icd coding. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 3105–3114, 2020.

Benjamin Paul Chamberlain, James Clough, and Marc Peter Deisenroth. Neural embeddings of graphs in hyperbolic space. In *13th International Workshop on Mining and Learning with Graphs*, 2019.

Ines Chami, Zhitao Ying, Christopher Ré, and Jure Leskovec. Hyperbolic graph convolutional neural networks. In *Advances in neural information processing systems*, pages 4868–4879, 2019.

Bryant Chen, Wilka Carvalho, Nathalie Baracaldo, Heiko Ludwig, Benjamin Edwards, Taesung Lee, Ian Molloy, and Biplav Srivastava. Detecting backdoor attacks on deep neural networks by activation clustering. In *Workshop on Artificial Intelligence Safety*, 2018.

Xinyun Chen, Chang Liu, Bo Li, Kimberly Lu, and Dawn Song. Targeted backdoor attacks on deep learning systems using data poisoning. 2017.

Zhe Chen, Xiao Zhang, and Ziwei Zhang. Hyperbolic relational graph neural network for drug-drug interaction prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 6945–6953, 2021.

Octavian Ganea, Gary Bécigneul, and Thomas Hofmann. Hyperbolic neural networks. In *Advances in neural information processing systems*, pages 5345–5355, 2018.

Micah Goldblum, Dimitris Tsipras, Chulin Xie, Xinyun Chen, Avi Schwarzschild, Dawn Song, Aleksander Mädry, Bo Li, and Tom Goldstein. Dataset security for machine learning: Data poisoning, backdoor attacks, and defenses. volume 45, pages 1563–1580, 2022.

Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Identifying vulnerabilities in the machine learning model supply chain. 2017.

Ji Guo, Wenbo Jiang, Rui Zhang, Wenshu Fan, Jiachen Li, Guoming Lu, and Hongwei Li. Backdoor attacks against hybrid classical-quantum neural networks. *Neural Networks*, page 107776, 2025.

Todd Huster, Jeremy Cohen, Zijie Hou, Mitchell Daniels, Rui Feng, and Vijay Janapa Reddi. A riemannian perspective on adversarial training. In *International Conference on Artificial Intelligence and Statistics*, pages 2917–2925, 2021.

Wei Jin, Yaxin Li, Han Xu, Yiqi Wang, and Jiliang Tang. Adversarial attacks and defenses on graphs: A review, a tool and empirical studies. 2020.

Valentin Khrulkov, Leyla Mirvakhabova, Evgeniya Ustinova, Ivan Oseledets, and Victor Lempitsky. Hyperbolic image embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6418–6428, 2020.

Yiming Li, Yong Jiang, Zhanglong Li, and Shu-Tao Xia. Backdoor learning: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

Yuekang Li, Yidan Mao, Yifei Yang, and Dongmian Zou. Improving robustness of hyperbolic neural networks by lipschitz analysis. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1713–1724, 2024.

Kang Liu, Brendan Doan-Gavitt, and Siddharth Garg. Fine-pruning: Defending against backdooring attacks on deep neural networks. In *International Symposium on Research in Attacks, Intrusions, and Defenses*, pages 273–294, 2018.

Kehan Liu, Yifan Zhang, and Chaoran Lu. Adversarial robustness of hyperbolic neural networks. 2022.

Ziwei Liu, Chaochao Chen, Longfei Li, Jun Zhou, Xiaolong Li, Le Song, and Yuan Qi. Hyperbolic graph neural networks for fraud detection. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1084–1092, 2019.

Nikita Malik, Rahul Gupta, and Sandeep Kumar. Hyperdefender: A robust framework for hyperbolic gnns. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 19396–19404, 2025.

Maximillian Nickel and Douwe Kiela. Poincaré embeddings for learning hierarchical representations. In *Advances in neural information processing systems*, pages 6338–6347, 2017.

Aldo Pareja, Giacomo Domeniconi, Jie Chen, Tengfei Ma, Toyotaro Suzumura, Hiroki Kanezashi, Tim Kaler, Tao Schardl, and Charles Leiserson. Evolvegcn: Evolving graph convolutional networks for dynamic graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 5363–5370, 2020.

Ahmed Salem, Rui Wen, Michael Backes, Shiqing Ma, and Yang Zhang. Dynamic backdoor attacks against machine learning models. In *IEEE European Symposium on Security and Privacy*, pages 703–718, 2022.

Lichao Sun, Yingtong Dou, Carl Yang, Ji Wang, Philip S Yu, Lifang He, and Bo Li. Security and privacy for graph neural networks: A survey. 2021.

Alexandru Tifrea, Gary Bécigneul, and Octavian-Eugen Ganea. Poincaré glove: Hyperbolic word embeddings. In *International Conference on Learning Representations*, 2019.

Max van Spengler, Jan Zahálka, and Pascal Mettes. Adversarial attacks on hyperbolic networks. In *European Conference on Computer Vision*, pages 363–381. Springer, 2024.

Tran The Vinh, Yi Tay, Shuai Zhang, Gao Cong, and Xiaoli Li. Hyperbolic deep learning for recommendation systems. In *The Web Conference*, pages 3463–3474, 2019.

Hadi Salman Wang, Matthias Ruder, Dan Neacsu, and Rachid Guerraoui. Certified patch robustness via smoothed vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15137–15147, 2022.

Xiao Yang, Gaolei Li, and Jianhua Li. Graph neural backdoor: Fundamentals, methodologies, applications, and future directions. *arXiv preprint arXiv:2406.10573*, 2024.

Daniel Zügner, Amir Akbarnejad, and Stephan Günnemann. Adversarial attacks on neural networks for graph data. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2847–2856, 2018.

8 Appendix A: Detailed Proof of Theorem 1

We recall that $\lambda_x = 2/(1 - \|x\|^2)$ denotes the conformal factor and write d_g for the hyperbolic distance. For $x \neq 0$, write $r = \|x\|$ and the outward radial unit vector $u_x =:= x/\|x\|$.

We collect two standard facts about radial geodesics in the Poincaré ball.

Lemma 3 (Radial geodesics and arclength)**.** *Let $r(s)$ denote the Euclidean radius of the outward radial unit-speed geodesic γ_x at hyperbolic arclength s, with $r(0) = r$. Then*

$$\frac{dr}{ds} = \frac{1-r^2}{2}, \qquad s = \int_r^{r(s)} \frac{2}{1-\rho^2}\, d\rho = \ln\left(\frac{1+r(s)}{1-r(s)}\right) - \ln\left(\frac{1+r}{1-r}\right),$$

so in particular

$$r(s) = \tanh\Big(\operatorname{artanh}(r) + \frac{s}{2}\Big). \tag{16}$$

Proof. Along a radial curve the metric reduces to $ds = \lambda(r)\, dr = \frac{2}{1-r^2} dr$, which yields the ODE and primitive directly. Solving for $r(s)$ gives the stated formula. □

Lemma 4 (Euclidean displacement under an outward radial step)**.** *With notation as above, the Euclidean displacement $\kappa(x, s) = r(s) - r$ satisfies*

$$\kappa(x, s) = \frac{(1-r^2)\tanh(s/2)}{1 + r\,\tanh(s/2)}. \tag{17}$$

Moreover, $\kappa(x, s) \leq (1 - r^2)\tanh(s/2) \leq s/\lambda_x$.

Proof. Using the formula for $r(s)$ from Lemma 3 and the identity $\tanh(a+b) = \dfrac{\tanh a + \tanh b}{1 + \tanh a\,\tanh b}$ with $\tanh(\operatorname{artanh} r) = r$ and $\tanh(s/2)$ as given, we obtain

$$r(s) = \frac{r + \tanh(s/2)}{1 + r\,\tanh(s/2)}.$$

Subtracting r yields the formula for $\kappa(x, s)$. Since $\tanh(s/2) \leq s/2$ and $1 + r\,\tanh(s/2) \geq 1$, we get

$$\kappa(x, s) \leq (1 - r^2)\tanh(s/2) \leq (1 - r^2)\frac{s}{2} = \frac{s}{\lambda_x}.$$

□

Proof of Theorem 1. (i) The exact formula for $\kappa(x, s)$ is given by Lemma 4, with the bounds in its last line.

(ii) If D is L_E-Lipschitz in the Euclidean metric, then

$$|D(\tau_s(x)) - D(x)| \leq L_E\, \|\tau_s(x) - x\|_2 = L_E\, \kappa(x, s).$$

Applying the bounds from part (i), we get $|D(\tau_s(x)) - D(x)| \leq L_E(1 - r^2)\tanh(s/2)$. If $\|x\| \geq 1 - \delta$, then $1 - r^2 = (1 - r)(1 + r) \leq 2(1 - r) \leq 2\delta$, which yields

$$\sup_{\|x\| \geq 1-\delta} |D(\tau_s(x)) - D(x)| \leq 2L_E\delta\tanh(s/2).$$

The expectation bound follows immediately.

(iii) Let $y = x + \kappa u_x$ with $\kappa \in (0, 1 - r)$ so that $r' = \|y\| = r + \kappa < 1$ and the segment remains inside $\mathbb{D}^n$. The hyperbolic distance between two points on the same radial geodesic equals the radial arclength:

$$d_g(x, y) = \int_r^{r'} \frac{2}{1-\rho^2}\, d\rho = \ln\left(\frac{1+r'}{1-r'}\right) - \ln\left(\frac{1+r}{1-r}\right).$$

Substitute $r' = r + \kappa$ and rearrange to get

$$d_g(x, y) = \ln\left(\frac{1-r}{1-r-\kappa}\right) + \ln\left(1 + \frac{\kappa}{1+r}\right) \;\geq\; \ln\left(\frac{1-r}{1-r-\kappa}\right).$$

Finally, with $\delta = \delta(x) = 1 - r$ and $\kappa < \delta$, the inequality $-\ln(1-t) \geq t$ for $t \in (0,1)$ gives

$$\ln\left(\frac{\delta}{\delta - \kappa}\right) = -\ln\left(1 - \frac{\kappa}{\delta}\right) \;\geq\; \frac{\kappa}{\delta}.$$

Thus $d_g(x,y) \geq \kappa/\delta$, completing the proof. □

Remarks. (1) The stealth guarantee depends *linearly* on the shell width δ and on $\tanh(s/2)$; no appeal to "exponential" growth of the conformal factor is needed (indeed λ_x grows rationally in $1 - \|x\|^2$). (2) Part (iii) formalizes amplification without coupling the Euclidean budget κ to δ beyond the natural feasibility condition $\kappa < \delta$; the per-unit amplification d_g/κ scales like $1/\delta$ near the boundary. (3) If a detector is Lipschitz in the *hyperbolic* metric (say, $|D(y) - D(x)| \leq L_g\, d_g(x,y)$), then for the same trigger $|D(\tau_s(x)) - D(x)| \leq L_g s$; i.e., hyperbolic-aware detectors do not enjoy the vanishing bound, clarifying which regularity notion matters for stealth.

9 Appendix B: Detailed Proof of Theorem 2

Definition 5 (Radial Defense)**.** *A defense $\mathcal{M} : \mathbb{D}^n \to \mathbb{D}^n$ is radial if it preserves direction and moves points inward along radial geodesics with displacement depending only on the hyperbolic radial coordinate.*

Assumption 6 (Recovery Success). The defense $\mathcal{M}$ successfully recovers an (α, β)-fraction of triggered inputs.

Assumption 7 (Radial Sensitivity). The classifier f_θ has radial sensitivity $\mu_g > 0$ along radial geodesics.

Lemma 8 (Radial-flow identity)**.** *For any x and $s > 0$, $\rho(\Phi_s(x)) = \rho(x) + \frac{s}{2}$ and $d_g\big(x, \Phi_s(x)\big) = s$.*

Proof. Along a radial geodesic one has $ds = \frac{2}{1-r^2}\, dr$, which integrates to $2\,\mathrm{artanh}(r(s)) - 2\,\mathrm{artanh}(r(0)) = s$; hence $\rho(\Phi_s(x)) = \rho(x) + s/2$ and the geodesic arclength equals s. □

Lemma 9 (Radiality reduces to a scalar profile)**.** *If $\mathcal{M}$ is radial in the sense of Definition 5, then for any x with $\rho(x) = \rho$,*

$$d_g\big(\mathcal{M}(x), x\big) = \Delta(\rho) = 2\big(\rho - m(\rho)\big),$$

which depends only on ρ and not on the direction of x.

Proof. By definition $\mathcal{M}$ preserves direction and maps ρ to $m(\rho) \leq \rho$ along the same radial geodesic. Hence $d_g(\mathcal{M}(x), x) = 2|\rho - m(\rho)| = 2(\rho - m(\rho))$ since $m(\rho) \leq \rho$. □

Proof of Theorem 2. Let $X \sim P$ and $Z = \Phi_s(X)$. By Lemma 8, $\rho(Z) = \rho(X) + \frac{s}{2}$. By Assumption 6 and Lemma 9,

$$\mathbb{P}\Big[\, \Delta\big(\rho(X) + \tfrac{s}{2}\big) \;\geq\; \alpha s \,\Big] \;\geq\; \beta. \tag{18}$$

Since Δ is L_Δ-Lipschitz in ρ,

$$\Delta\big(\rho(X)\big) \;\geq\; \Delta\big(\rho(X) + \tfrac{s}{2}\big) - L_\Delta \cdot \tfrac{s}{2}.$$

Therefore on the event in (18) we have $\Delta(\rho(X)) \geq \alpha s - \frac{L_\Delta}{2} s = \alpha_{\text{eff}} s$. This implies

$$\mathbb{P}\big[\, d_g\big(\mathcal{M}(X), X\big) = \Delta(\rho(X)) \;\geq\; \alpha_{\text{eff}} s \,\big] \;\geq\; \beta, \tag{19}$$

which is the probability lower bound. Taking expectations and using nonnegativity gives:

$$\mathbb{E}\big[d_g(\mathcal{M}(X), X)\big] \geq \mathbb{E}\big[\, \alpha_{\text{eff}} s \cdot \mathbf{1}\{\Delta(\rho(X)) \geq \alpha_{\text{eff}} s\} \,\big] \geq \beta\, \alpha_{\text{eff}} s. \tag{20}$$

Next, by Assumption 7 (applied with $y = \mathcal{M}(X)$), for each outcome

$$\|f_\theta(\mathcal{M}(X)) - f_\theta(X)\| \;\geq\; \mu_g\, d_g\big(\mathcal{M}(X), X\big).$$

Taking expectations yields:

$$\mathbb{E}[\|f_\theta(\mathcal{M}(X)) - f_\theta(X)\|] \geq \mu_g\, \beta\, \alpha_{\text{eff}} s. \tag{21}$$

Finally, since at least a β-fraction of the mass satisfies $\|f_\theta(\mathcal{M}(X)) - f_\theta(X)\| \geq \mu_g \alpha_{\text{eff}} s$ and the integrand is nonnegative elsewhere,

$$\mathbb{E}\big[\|f_\theta(\mathcal{M}(X)) - f_\theta(X)\|^2\big] \;\geq\; \beta\, (\mu_g \alpha_{\text{eff}} s)^2, \tag{22}$$

which is the second moment bound. □