1. Adversarial Attacks

FGSM: **Targeted**: $x' = x - \varepsilon \cdot \mathrm{sign}(\nabla_x \mathcal{L}(x,t))$ (toward target t) **Untargeted**: $x' = x + \varepsilon \cdot \mathrm{sign}(\nabla_x \mathcal{L}(x,y))$ (away from true y) Sign normalizes ∇→lands on ℓ_∞ ball **vertex** η not minimized, just $\in [-\varepsilon,\varepsilon]^d$

PGD: $x^{k+1} = \Pi_{\mathbb{B}_{\varepsilon}(x^0)}\left(x^k + \alpha \cdot \mathrm{sign}(\nabla\mathcal{L})\right)$ **Init**: $x^0 + \mathrm{uniform}(-\varepsilon,\varepsilon)$ **Step decay**: $\alpha^k = \frac{\alpha^0}{2^k}$ (halve each iter) **Projection**: ℓ_∞=clip; ℓ_2=scale to radius ℓ_2 proj: $x'' = x^0 + \frac{\varepsilon}{\|x'-x^0\|_2}(x'-x^0)$ if $\|x'-x^0\| > \varepsilon$ Optimal adv example **always on boundary** (high-dim monotonic)

C&W: $\min_\eta \|\eta\|_p^2 + c \cdot \mathrm{OPS}(x+\eta,t)$ $\mathrm{OPS} = \max\left(0, \max_{i\neq t} Z_i - Z_t + \kappa\right)$ OPS$\leq 0 \Rightarrow$ attack succeeds; κ controls margin **Different from PGD**: minimizes perturbation size, not fixed ε Use LBFGS-B for box constraints; binary search on c

Targeted vs Untargeted: Binary case ($d=2$): equivalent! Away from class 1 = toward class 2 $d \geq 3$: NOT equivalent! Untargeted has multiple directions Loss relation: $\mathcal{L}(x,t) = -\mathcal{L}(x,y)$ only for 2-class

GCG (LLM Discrete): Tokens discrete, can't do PGD directly

1. One-hot→continuous: compute $\nabla_e \mathcal{L}$ in embedding space
2. Top-K filter: select K tokens with most negative ∇
3. Greedy search: enumerate positions, keep best

Use ∇ to FILTER, not UPDATE! Complexity $O(V^k)$ exponential Universal suffix: $\min_{\text{suf}} \sum_i \mathcal{L}(\text{Sure}|p_i, \text{suf})$ transfers to GPT-4

Norm Relations: $\|v\|_\infty \leq \|v\|_2 \leq \sqrt{n}\|v\|_\infty$ $\|v\|_2 \leq \|v\|_1 \leq \sqrt{n}\|v\|_2$ $\mathbb{B}^1_\varepsilon \subset \mathbb{B}^2_\varepsilon \subset \mathbb{B}^\infty_\varepsilon$ ℓ_∞ constraint $\Rightarrow$ ℓ_2 constraint (converse false)

AutoAttack: Ensemble: APGD-CE + APGD-DLR + FAB + Square (black-box) **Must use for reporting robust accuracy** Prevents "overfitting" defense to single attack

2. Defenses

Min-Max Framework: $\min_\theta \mathbb{E}_{\{(x,y)\}}\left[\max_{x'\in S(x)} \mathcal{L}(\theta,x',y)\right]$ **Attack**: fix θ, find δ (inner max) **Defense**: optimize both (outer min) **Certify**: replace inner max with relaxation upper bound

PGD-AT: For each batch: $x_{\text{adv}} = \mathrm{PGD}(x,\theta,\varepsilon)$ Backprop on $\nabla_\theta \mathcal{L}\left(f_{\theta(x_{\text{adv}})}, y\right)$ Inner: PGD (10-20 steps); Outer: SGD on θ

TRADES: $\mathcal{L} = \mathcal{L}(f(x),y) + \lambda \max_{x'\in\mathbb{B}_\varepsilon} \mathrm{KL}(f(x)\|f(x'))$ Separately optimize clean acc and robustness λ trades off; typically $\lambda \in [1,6]$ Often better clean-robust Pareto frontier than PGD-AT

ε-Robustness & Accuracy: If $\exists (x_1,y_1),(x_2,y_2)$ with $y_1 \neq y_2$ and $\|x_1-x_2\|_p \leq \varepsilon$: **Cannot have both ε-robust and 100% accurate** Proof: if f robust at x_1, all points in $\mathbb{B}_{\varepsilon(x_1)}$ same label→x_2 misclassified

3. Certification

Core: $\forall i : \varphi(i) \Rightarrow N(i) \vDash \psi$

Sound vs Complete: **Sound**: Proved ⇒ True (no false positive, 底线!) **Complete**: True ⇒ Provable (no false negative) Most practical: Sound but Incomplete (Box, DeepPoly, RS) MILP: Sound+Complete but $O(2^k)$

Crossing ReLU: Input bounds $[l,u]$ with $l<0<u$: **unstable** $l \geq 0$: $y=x$ exact; $u \leq 0$: $y=0$ exact MILP complexity $O(2^k)$ where k=**#Crossing** (NOT total neurons!) Reduce k: tighter bounds, certified training

3.1 MILP (Complete)

MILP Encoding: **Affine**: $y=Wx+b$ directly encoded **ReLU** ($l<0<u$): introduce $a\in\{0,1\}$ $y \geq x$, $y \leq x - l(1-a)$, $y \leq u \cdot a$, $y \geq 0$ $a=1$: $y=x$ (active); $a=0$: $y=0$ (inactive) **Specification**: $\varphi = \mathbb{B}^\infty_\varepsilon$: $x_i - \varepsilon \leq x'_i \leq x_i + \varepsilon$ ψ: prove $o_t > o_j\ \forall j \neq t$: minimize $o_t - \max_{j\neq t} o_j$

MILP for Other Funcs: **HatDisc/Abs**: $y=|x|$: $y \geq x$, $y \geq -x$, $y \leq x + 2u(1-a)$, $y \leq -x + 2|l|a$ **Max**: $y = \max(x_1,x_2)$: $y \geq x_1, y \geq x_2, y \leq x_1 + a(u_2 - l_1)$, $y \leq x_2 + (1-a)(u_1-l_2)$ **Binary Step**: like ReLU but output $\{0,1\}$ not $[0,1]$

MILP Limitations: ℓ_2 ball is **quadratic** constraint→MILP **incomplete** for ℓ_2! Floating-point: theory Sound $\neq$ hardware Sound (rounding errors) Infinite compute: Box-MILP equiv MILP-MILP (both explore all branches)

3.2 Relaxation (Incomplete)

Box/IBP $O(n^2L)$: $[a,b] +^{\#[c,d]} = [a+c, b+d]$; $-^{\#[a,b]} = [-b,-a]$ $\lambda[a,b] = \begin{cases}[\lambda a, \lambda b] & \lambda \geq 0\\ [\lambda b, \lambda a] & \lambda < 0\end{cases}$ $\mathrm{ReLU}^{\#[l,u]} = [\mathrm{ReLU}(l), \mathrm{ReLU}(u)]$ **Affine exact**; ReLU crossing→over-approx (garbage points) Loosest but GPU-friendly, parallelizable

Box Propagation Example: Given $x_1 \in [0,0.5]$, $x_2 \in [0.2,0.7]$: $x_3 = x_1 + x_2 \in [0.2,1.2]$ (non-crossing, $l \geq 0$) $x_4 = x_1 - x_2 \in [-0.7,0.3]$ (**crossing!** $l<0<u$) After ReLU: $x_5 = \mathrm{ReLU}(x_3) \in [0.2,1.2]$; $x_6 = \mathrm{ReLU}(x_4) \in [0,0.3]$

DeepPoly $O(n^3L^2)$: Each x_i: interval $l_i \leq x_i \leq u_i$ Relational: $a_i^L \leq x_i \leq a_i^U$ where $a = \sum w_j x_j + \nu$ **Affine**: exact, $z \leq Wx+b \leq z$ (upper=lower) **ReLU** ($l<0<u$): $\lambda = \frac{u}{u-l}$ Upper: $y \leq \lambda(x-l)$ (fixed, connects $(l,0)$ to (u,u)) Lower: $y \geq \alpha x$, $\alpha \in [0,1]$ (optimizable, α-CROWN) Min area: $\alpha = 0$ if $|l| > u$; $\alpha=1$ otherwise

Back-Substitution: Recursively expand symbolic bounds to input layer **Key**: for $X_j \leq \sum c_i X_i + d$:

- If $c_i > 0$: substitute upper bound of X_i
- If $c_i < 0$: substitute **lower** bound of X_i (opposite!)

Can stop early using concrete bounds (efficiency)

DeepPoly Example: $x_5 = \mathrm{ReLU}(x_3)$, $x_3 \in [-0.5,3.5]$ (crossing) Upper: $x_5 \leq \frac{3.5}{4}(x_3+0.5) = 0.875x_3 + 0.4375$ Lower: $x_5 \geq 0$ (if $\alpha=0$) or $x_5 \geq x_3$ (if $\alpha=1$) Back-sub to get concrete $[l_5,u_5]$

Single vs Multi-Neuron: **Single**: each neuron independent, fully parallel (GPU) **Multi** (PRIMA): captures cross-neuron relations, tighter but serial DeepPoly=single-neuron; trades precision for speed

Triangle vs DeepPoly: Triangle: 3 constraints (exact convex hull), exponential growth DeepPoly: 2 constraints (parallelogram), fixed complexity Triangle doesn't scale; DeepPoly does

3.3 Branch & Bound

B&B Algorithm:

1. **Bound**: compute bounds via DeepPoly/CROWN
2. If $l>0$: SAFE; if $u<0$: UNSAFE (counterexample)
3. **Branch**: select unstable ReLU, split on $x_i \geq 0$ vs $x_i < 0$
4. Recurse on both subproblems

Worst case: $O(2^k)$; good heuristics crucial

Branching Heuristics: **Largest interval**: $\max(u-l)$ most uncertain **Closest to zero**: $\min(|l|,|u|)$ most critical **∇-based**: $\max|\nabla_x \text{ obj}|$ most impact **Learning-based**: NN predicts best split

KKT/Lagrangian: $(\max_x f(x)$ s.t. $g(x) \leq 0) \leq \max_x \min_{\beta \geq 0}[f(x) + \beta g(x)]$ **Weak duality**: max min $\leq$ min max (always holds) Split constraint $x_i \geq 0$: add βx_i to objective β found by ∇ descent; need full back-sub each step

α-β-CROWN: α: ReLU lower slope $\in [0,1]$, ∇-optimizable β: Lagrange multiplier ≥ 0, encodes split constraints **Key**: α, β only affect **Tightness**, NOT **Soundness**! Any valid α,β gives sound bound, just looser/tighter

4. Certified Training

DiffAI Framework: $\min_\theta \mathbb{E}\left[\max_{z\in\gamma(f^{\#(S(x))})} \mathcal{L}(z,y)\right]$ Use abstract transformer (Box/DeepPoly) instead of PGD **Abstract loss**: optimize over output region (incl. garbage points)

Abstract Loss $\mathcal{L}^\#$: **Margin loss** $\mathcal{L} = \max_{c\neq y}(z_c - z_y)$: Compute $d_c = z_c - z_y$ for all c; take max of upper bounds **CE loss**: for each class, take upper (if $c \neq y$) or lower (if $c = y$) Compute CE on this worst-case logit vector

Training Paradox: Empirical: Box(86%) > Zonotope(73%) > DeepPoly(70%) **Tighter $\neq$ better training!** Reason: tighter→discrete switching→discontinuous landscape→hard optimize Box: loose but smooth ∇s

SABR/COLT: SABR: propagate to layer k, freeze; PGD on layers $k+1$ to n Solves projection problem: ℓ_∞=clip; DeepPoly shape needs QP COLT: similar layer-wise approach with Zonotope

Certified Training Step: Given network, input spec $x \in [l,u]$, weight w:

1. Box propagate: $x_3 \in [l_3(w), u_3(w)]$ as function of w
2. Compute worst-case loss: $\mathcal{L}_{\text{worst}} = \log(1 + \exp(u_7 - l_8))$
3. ∇: $\nabla_w \mathcal{L}_{\text{worst}}$
4. Update: $w \leftarrow w - \eta \nabla_w \mathcal{L}_{\text{worst}}$

Bounds are **continuous** in w (linear+max are continuous)

5. Randomized Smoothing

Smoothed Classifier: $g(x) = \arg\max_c \mathbb{P}_{\varepsilon\sim\mathcal{N}(0,\sigma^2 I)}[f(x+\varepsilon)=c]$ Base f can be fragile; smoothed g has certified guarantee **Theorem is deterministic; estimation is probabilistic!**

Certified Radius: If $\underline{p_A} > 0.5$: $R = \sigma \cdot \Phi^{-1}\left(\underline{p_A}\right)$ Φ^{-1}: inverse standard normal CDF (probit) $p_A = 0.5 \Rightarrow \Phi^{-1}(0.5) = 0 \Rightarrow R = 0$ $p_A \to 1 \Rightarrow \Phi^{-1}(p_A) \to \infty \Rightarrow R \to \infty$ **σ ↑ doesn't always mean R ↑!** (larger noise→lower p_A)

Two-Stage Sampling: **Stage 1** ($n_0 \approx 100$): guess top class $\hat{c}_A$ **Stage 2** ($n \approx 10^5$): estimate $\underline{p_A}$ via Clopper-Pearson CI If $\underline{p_A} \leq 0.5$: ABSTAIN **Complexity**: $O(n_{\text{samples}})$, independent of network size!

Inference with Hypothesis Testing: H_0: true $p(\text{success}) = 0.5$ BinomPValue$(n_A, n, 0.5)$: reject if $< \alpha$ α small: more ABSTAIN but higher confidence Returns wrong class with prob at most α

Why ℓ_2 Only?: Gaussian is **rotation invariant**: $\|X\|_2$ independent of direction →isotropic, equal prob surface is 球→ℓ_2 analytic formula Laplace→ℓ_1; Uniform→ℓ_∞: no closed form

RS vs Convex: **Speed**: RS often **slower** (10k forward passes vs 1 abstract pass) **Scalability**: RS works on any size (LLMs); Convex limited to small/medium **Guarantee**: RS probabilistic; Convex deterministic **Training**: RS no special training; Convex needs certified training

Common Failures: Wrong top class: n_0 too small→increase n_0 $p_A \leq 0.5$: base model bad under noise→Gaussian adversarial training Lower bound too loose: n too small→increase n

6. DP & RS Duality

DP vs RS: Same Tools, Opposite Goals: **DP**: make distributions **indistinguishable** $P[M(D)] \approx P[M(D')]$ **RS**: make predictions **distinguishable** $P[G(x)=c] \gg P[G(x)\neq c]$ Both use noise mechanisms, exponential bounds DP: want hypothesis test power **low**; RS: want confidence **high**

Lipschitz Connection: Both proofs rely on Lipschitz constant L: DP: L controls sensitivity→determines noise RS: L controls p_A change→determines radius DP Noise $\propto \frac{L}{\varepsilon}$; RS Radius $\propto \frac{\sigma}{L}$

7. Privacy

ε-DP: $\mathbb{P}(M(D)\in S)\le e^{\varepsilon}\mathbb{P}(M(D')\in S)$ for all neighboring D, D' $e^{\varepsilon}\approx 1+\varepsilon$ for small ε **Laplace**: $f(D)+\mathrm{Lap}\left(\frac{\Delta_1}{\varepsilon}\right)$; $\Delta_p=\max_{D\sim D'}\|f(D)-f(D')\|_p$

(ε,δ)-DP: $\mathbb{P}(M(D)\in S)\le e^{\varepsilon}\mathbb{P}(M(D')\in S)+\delta$ δ: tail prob bound, **NOT "leak prob"!** Typically $\delta\ll\frac{1}{n}$ **Gaussian**: $\sigma=\frac{\Delta_2\sqrt{2\ln(\frac{1.25}{\delta})}}{\varepsilon}$

Neighbor Definitions: $\|D-D'\|_0\le 1$: add/remove one record→Laplace $\|D-D'\|_2\le 1$: continuous perturbation (∇s)→Gaussian

Three Properties: **Post-processing**: $g\circ M$ still DP (can't "purify" noise) **Composition**: $(\varepsilon_1+\varepsilon_2,\delta_1+\delta_2)$-DP **Subsampling**: sample rate $q\Rightarrow(q\varepsilon,q\delta)$ **Advanced**: T steps→$\varepsilon_{\text{tot}}=O\left(\sqrt{T}\varepsilon\right)$ (crucial for training!) Independent data: $(\max\varepsilon,\max\delta)$

DP-SGD:
1. **Clip** each ∇: $g_{\text{clip}}=g\cdot\min\left(1,\frac{C}{\|g\|_2}\right)$
2. **Aggregate + noise**: $g_{\text{noisy}}=\frac{1}{L}\sum g_{\text{clip}}+\mathcal{N}\left(0,\sigma^2\frac{C^2}{L^2}\right)$

Clipping bounds sensitivity $\Delta_2\le C$ $\sigma=\frac{C\sqrt{2\ln(\frac{1.25}{\delta})}}{L\varepsilon}$ Model private even against white-box attacker

Privacy Amplification: Apply (ε,δ)-DP on random subset $q=\frac{L}{N}$: Result: $(\tilde{q}\varepsilon,q\delta)$-DP where $\tilde{q}\approx q$ T steps: $(\tilde{q}T\varepsilon,qT\delta)$ or $\left(O\left(q\varepsilon\sqrt{T}\right),\delta\right)$

PATE: M teachers on disjoint data, noisy voting labels public data, train student $n_{j(x)}=\#\{t:t(x)=j\}$; output $\arg\max\left(n_{j(x)}+\mathrm{Lap}\left(\frac{2}{\varepsilon}\right)\right)$ **Add noise BEFORE argmax!** Sensitivity=2 (NOT $|Y|$!) Each query costs ε; total budget accumulates

FedSGD vs FedAvg: **FedSGD**: send single-step ∇ g_k; server averages **FedAvg**: client runs E epochs, sends weight diff $\Delta\theta$ FedAvg harder to invert (multi-step trajectory unknown)

DP-FedSGD Noise: Centralized: $\sigma_{\text{central}}=\frac{C\sqrt{2\ln(\frac{1.25}{\delta})}}{L\varepsilon}$ Distributed (m clients): $\sigma_{\text{client}}=\sqrt{m}\cdot\sigma_{\text{central}}$ Aggregation: $\frac{1}{m}\sum g_k$ gives same noise level as centralized

8. Privacy Attacks

Attack Hierarchy: **Attribute Inference**: infer sensitive attr (**no membership needed!**) **Data Extraction**: verbatim memorization (K-extractable) **MIA**: determine if $x\in D_{\text{train}}$ **Dataset Inference**: aggregate weak signals→strong signal **∇ Inversion**: reconstruct from ∇s (FL)

MIA Methods: **Shadow Model**: train K shadows, train attack classifier **LiRA**: $\log\left(\frac{P(\ell|x\in D)}{P(\ell|x\notin D)}\right)$ likelihood ratio **Min-K% Prob**: average of lowest K token probs (LLM) **Loss-based**: training data has lower loss Practical AUC≈0.5-0.7 (weak!); TPR@FPR=0.01 only 2%

∇ Inversion: $x^*=\arg\min_x\|\nabla_\theta\mathcal{L}(x,y)-\nabla_{\text{obs}}\|^2+R(x)$ Prior $R(x)$: TV (image), Perplexity (text), Entropy (tabular) FedSGD + BS=1: **exact reconstruction** ($\nabla W_1=\delta x^\top$) FedAvg: needs multi-epoch coupling, harder

Model Stealing/Inversion: **Stealing**: query API, train copy via distillation **Inversion**: $x^*=\arg\max_x P(y_{\text{target}}|x)$ reconstruct class representative Defense: rate limit, output perturbation, watermarking

Memorization Factors: Model size↑, Prefix length↑, Repetition↑: more memorization Sequence length↑: less (cumulative errors)

9. Synthetic Data & Marginals

Pipeline:
1. **Select** marginal queries; 2. **Measure** with DP; 3. **Generate** synthetic

Marginal $\mu_t=\sum_{x\in D}[x_C=t]$; $\Delta_2(M_C)=1$ (one row→one entry)

Chow-Liu: MI-weighted complete graph→MST→sample along tree $p(F_1,F_2,F_3)=p(F_1)p(F_2|F_1)p(F_3|F_1)$ DP: exponential mechanism for MST, Gaussian for marginals

Marginal Properties: $(n-1)$-way marginals do **NOT** uniquely describe dataset Low-order marginals miss high-order correlations (XOR problem) 3 columns, all 2-way: $\binom{3}{2}=3$ queries→3ε total

10. Logic & DL2

Logic→Loss Translation: Theorem: $T(\varphi)(x)=0\Leftrightarrow x\models\varphi$ $t_1\le t_2$: $\max(0,t_1-t_2)$; $t_1=t_2$: $(t_1-t_2)^2$ $\varphi\wedge\psi$: $T(\varphi)+T(\psi)$; $\varphi\vee\psi$: $T(\varphi)\cdot T(\psi)$ By construction $T(\varphi)\ge 0$; negation via De Morgan **Quantifiers NOT directly supported**; $\forall$ via max (worst violation)

Training with Background Knowledge: Goal: $\max_\theta\mathbb{E}[\forall z.\varphi(z,s,\theta)]$ Reform: $\min_\theta\mathbb{E}[T(\varphi)(\hat{z},s,\theta)]$ where $\hat{z}=\arg\max T(\neg\varphi)$ This is adversarial attack! Restrict z to ℓ_∞ ball, PGD+project

Logic Properties: If $T(\neg\varphi)(y)=0$, then $\neg\varphi$ satisfied at y→$\forall x.\varphi(x)$ FALSE $T(\varphi)(y_1)\le T(\varphi)(y_2)\Rightarrow T(\neg\varphi)(y_1)\ge T(\neg\varphi)(y_2)$ Infinite minimizers possible (e.g., φ is tautology)

11. Fairness

Individual Fairness: (D,d)-Lipschitz: $D(M(x),M(x'))\le d(x,x')$ Equivalent to robustness: $\forall\delta\in\mathbb{B}_{S(0,\frac{1}{L})}:M(x)=M(x+\delta)$ Lemma: $\Phi^{-1}(\mathbb{E}[h(x+\varepsilon)])$ is 1-Lipschitz

Group Fairness: **Demographic Parity**: $\mathbb{P}\left(\hat{Y}=1|S=0\right)=\mathbb{P}\left(\hat{Y}=1|S=1\right)$ **Equal Opportunity**: above conditioned on $Y=1$ (TPR equal) **Equalized Odds**: conditioned on both $Y=0$ and $Y=1$ Eq Odds $\Leftrightarrow\hat{Y}\perp S|Y$ (conditional independence)

Δ_{EO} Calculation: $\Delta_{\text{EO}}=|\text{FPR}_0-\text{FPR}_1|+|\text{TPR}_0-\text{TPR}_1|$ Example: $S=0$: FPR=7/10=0.7, TPR=3/6=0.5 $S=1$: FPR=2/8=0.25, TPR=16/20=0.8 $\Delta_{\text{EO}}=|0.7-0.25|+|0.5-0.8|=0.45+0.3=0.75$

Adversary Bound: Balanced Accuracy: $\mathrm{BA}(h)=\frac{1}{2}\left(\mathbb{E}_{Z_0}(1-h)+\mathbb{E}_{Z_1}h\right)$ Optimal adversary: $h^*(z)=[p_1(z)\ge p_0(z)]$ Theorem: $\Delta_{\text{EO}(g)}\le 2\cdot\mathrm{BA}(h^*)-1$

Eq Odds Proof Sketch: Goal: $\mathbb{P}\left(\hat{Y}=1|S=s,Y=y\right)$ same for all s→$\hat{Y}\perp S|Y$ Use: $\mathbb{P}\left(\hat{Y}|Y\right)=\sum_s\mathbb{P}\left(\hat{Y}|S=s,Y\right)\mathbb{P}(S=s|Y)$ If $\mathbb{P}\left(\hat{Y}|S,Y\right)=c$ for all s: $\mathbb{P}\left(\hat{Y}|Y\right)=c$→conditional indep

LAFTR: $\min_{f,g}\max_h\left[\mathcal{L}_{\text{clf}(f,g)}-\gamma\mathcal{L}_{\text{adv}(f,h)}\right]$ Use adversary to upper bound unfairness

LCIFR: Train encoder: $\forall x'\in S_{d(x)}:\|f(x)-f(x')\|_\infty\le\delta$ MILP compute ε s.t. $f\left(S_{d(x)}\right)\subset\{z':\|f(x)-z'\|_\infty\le\varepsilon\}$ Consumer gets simple robustness problem

12. Watermark & Benchmark

Red-Green Watermark: hash(context)+key→split vocab into Green/Red **Generate**: add δ bias to Green token logits **Detect**: count Green tokens, binomial test **without LLM!** p-value< α→watermarked; α controls FPR directly

ITS/SynthID: **ITS**: distortion-free in expectation, but deterministic output **SynthID**: distortion-free + non-deterministic Tournament sampling: high G-value tokens more likely to win

Watermark Attacks: **Scrubbing**: paraphrase 30% tokens removes watermark **Spoofing**: modify one word, watermark persists (piggyback) **Stealing**: 30K queries estimate $\frac{P_{\text{wm}}}{P_{\text{base}}}$, predict Green

Contamination: **Data**: benchmark in training set (memorize answers) **Task**: optimized for task format (not truly solving) Detection: N-gram (L1), Perplexity (L2), Completion (L3) Outcome-based: compare 2024 vs 2025 performance (time causality)

VNN-COMP Critique: "Verified 68M params"→check: #Crossing, accuracy, ε size Small ε=fewer crossing=easier; timeout=3600s impractical Verified $\neq$ practically robust

13. Post-Training Attacks

Quantization Attack: FP32 benign (passes detection), INT8 malicious (activated after deploy) Box constraint $[w_{\text{low}},w_{\text{high}}]$ s.t. quantized value unchanged Fine-tune in box with clean data→FP32 looks normal

Fine-Tuning Attack: $\mathcal{L}=\mathcal{L}_{\text{clean}(\theta)}+\lambda\mathcal{L}_{\text{attack}(\theta-\nabla\mathcal{L}_{\text{user}})}$ Safe now, malicious after user fine-tunes Needs Hessian: $(\partial\mathcal{L})\frac{\theta'}{\partial}\theta=\frac{\partial\mathcal{L}}{\partial}\theta'\cdot\left(I-\eta\nabla^2\mathcal{L}_{\text{user}}\right)$

Agentic AI / IPI: Indirect Prompt Injection: malicious instruction in tool output Agent can't distinguish user instruction vs tool content Defense: instruction hierarchy, dual-LLM, command sense Tradeoff: security $\propto$ 1/capability

14. Regulation

EU AI Act: **Unacceptable**: social credit scoring→prohibited **High Risk**: credit scoring, hiring→strict regulation **Limited Risk**: chatbots→transparency requirements Credit scoring is **High Risk**, NOT prohibited!

GDPR: Removing PII insufficient→linkage attacks still possible Even "anonymized" purchase lists may violate GDPR

Appendix: **Norms**:$\|x\|_p=\left(\sum|x_i|^p\right)^{\frac{1}{p}}$; $\|x\|_\infty=\max|x_i|$ $\mathcal{N}=(2\pi)^{-\frac{d}{2}}|\Sigma|^{-\frac{1}{2}}\exp\left(-\frac{1}{2}(x-\mu)^\top\Sigma^{-1}(x-\mu)\right)$ Lap $=\frac{1}{2b}\exp(-|x-\mu|/b)$; Sigmoid$\sigma(x)=\frac{1}{1+e^{-x}}$ **Softmax&CE**: $\sigma(z)_i=e^{z_i}/\sum_j e^{z_j}$; $\mathrm{CE}(z,y)=-\log\sigma(z)_y=-z_y+\log\sum e^{z_j}$ **Derivatives**: $\partial_x b^\top x=b$; $\partial_x x^\top x=2x$; $\partial_x x^\top Ax=(A+A^\top)x$ $\partial_x\|Ax-b\|_2^2=2A^{\top(Ax-b)}$ **不等式**: Cauchy-Schwarz:$\langle x,y\rangle\le\|x\|_2\|y\|_2$ Hölder:$\|x\cdot y\|_1\le\|x\|_p\|y\|_q,\frac{1}{p}+\frac{1}{q}=1$ Jensen:g凸 $\Rightarrow g(\mathbb{E}[X])\le\mathbb{E}[g(X)]$ Chebyshev: $\mathbb{P}(|X-\mathbb{E}[X]|\ge\varepsilon)\le\frac{\mathbb{V}[X]}{\varepsilon^2}$ Minmax:max min $\le$ min max(Weak Duality) Hoeffding:$\mathbb{P}\left(|\hat{X}-\mathbb{E}[X]|\ge\varepsilon\right)\le 2\exp\left(-2n\frac{\varepsilon^2}{(b-a)^2}\right)$ **prob**: $\mathbb{V}(X)=\mathbb{E}[X^2]-\mathbb{E}[X]^2$; $\mathbb{V}(aX+bY)=a^2\mathbb{V}(X)+b^2\mathbb{V}(Y)+2ab\mathrm{Cov}$ Bayes: $P(X|Y)=P(Y|X)\frac{P(X)}{P(Y)}$ $\Phi(z)=\mathbb{P}(\mathcal{N}(0,1)\le z)$; $\Phi^{-1}(0.5)=0$; $\Phi^{-1}(0.975)\approx 1.96$ **Matrix**: $\begin{pmatrix}a&b\\c&d\end{pmatrix}^{-1}=\frac{1}{ad-bc}\begin{pmatrix}d&-b\\-c&a\end{pmatrix}$ **MILP编码**: $y=|x|$:$y\ge x,y\ge -x,y\le x+2u(1-a),y\le -x+2|l|a,a\in\{0,1\}$ $y=\max(x_1,x_2)$:$y\ge x_1,y\ge x_2,y\le x_1+a(u_2-l_1),y\le x_2+(1-a)(u_1-l_2)$ **Logic**: De Morgan:$\neg(\varphi\wedge\psi)=\neg\varphi\vee\neg\psi$; $\neg(\varphi\vee\psi)=\neg\varphi\wedge\neg\psi$ Implication:$\varphi\Rightarrow\psi\equiv\neg\varphi\vee\psi$ Ball:$\mathbb{B}^1_\varepsilon\subseteq\mathbb{B}^2_\varepsilon\subseteq\mathbb{B}^\infty_\varepsilon\subseteq\mathbb{B}^2_{\varepsilon\sqrt{d}}$

⚠ **Traps**: MILP complexity $O(2^k)$, k=**Crossing count**! RS theorem deterministic, estimation probabilistic $\sigma\uparrow$ doesn't always $R\uparrow$ (p_A drops!) GCG uses ∇ to **filter**, not **update** n_0 (guess class 100) vs n (estimate prob 100k) PATE: noise **before** argmax, $\Delta_1=2$ Tighter $\neq$ better training (Box trains best) Back-sub: negative coeff→opposite bound MILP incomplete for ℓ_2 (quadratic) PGD $\neq$ CW: different objectives FGSM always on ℓ_∞ boundary FedSGD easier to invert than FedAvg δ is tail mass bound, not leak prob Gaussian DP needs ℓ_2 sensitivity Floating-point: theory Sound $\neq$ hardware Sound Credit scoring is High Risk, NOT Unacceptable MIA AUC≈0.5-0.7 (basically random) Universal suffix transfers across models