Approach to the Evaluation of AI Model Robustness to Adversarial Attacks

Aleksandr Sukhanov[1] **Evgenia Novikova**[2]

[1]St. Petersburg State Electrotechnical University “LETI”

[2]St. Petersburg Federal Research Center of the Russian Academy of Sciences

International Conference on Cyber Warfare, Security & Space Computing (SpacSec'2026), 26-28 October, 2026.

Problem statement

- Artificial Intelligence (AI) system has become ubiguitous in critical application domains
- The problem of the AI robustness has attracted a lot of research attention:
  - Design and implementation of the adversarial attacks against AI models
  - Development of the methodologies for AI robustness evaluation and quantification
- Existing evaluation practices include
  - Standards provide conceptual frameworks or quidelines lacking concret operational details
  - Benchmarks provide collection of attacks and metrics
  - Evaluation results depend on selected attack budget or specific benchmark
- There is a gap from guidelines and measurable evaluation

Research objective and tasks

Goal

- To develop a methodological foundations for the AI trustworthiness assessment
- ***Tasks***
- An evaluation procedure that enables robustness assessment of the AI model independent to the model architecture with controllable attack budget and integral score that incorporates model behaviour under different attack scenarios

Analysis of the Related Works

- The generic scheme of adversarial sample generation:

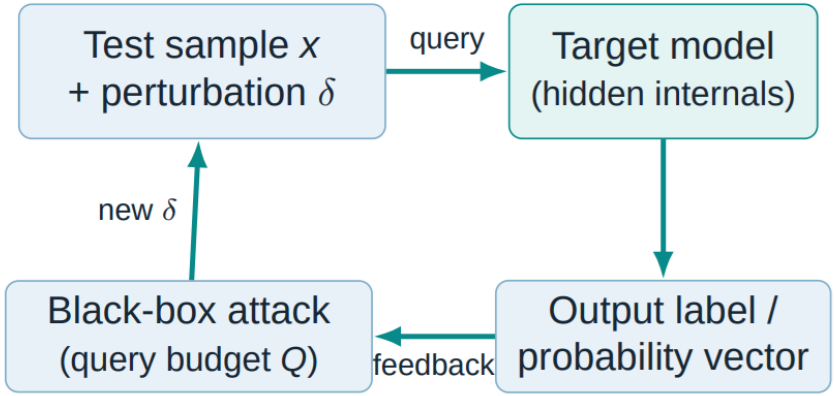
$$x_{adv} = x + \delta,\ \delta \text{ - perturbation},\ \varepsilon - ||\delta||_p \text{ – attack budget}$$

- The most powerful attacks white box attacks, tailored to data modality and neural network architecture
- Variety of the evaluation metrics
  - heuristic ones, based on experimental evaluation: Attack success rate, robust accuracy, average perturbation norm
  - formal certified metrics: Certified Radius, Certified Accuracy, etc.
- Testing guidelines
  - OWASP AI Testing Guide - comprehensive collection of the methodologies and best practices for assessing AI-based system's robustness at different levels
  - Russian Standart GOST R ISO/IEC 24029-1 and GOST R ISO/IEC 24029-2
- Numerous robustness evaluation frameworks, e.g. BlackboxBench, RobustBench provide tools for evaluating robustness

AI Robustness Evaluation Approach

Requirements

- Support of black box testing, which implies access only to the inputs and outputs of the model.
- Applicability, at the methodological level, to different data modalities provided that an appropriate black-box adversarial attack and perturbation measure are defined for the target domain.
- A reasonable assessment of stability, that considers the probabilistic nature of the attacks and takes into account the noise generated by the stochastic methods.



Available to the evaluator

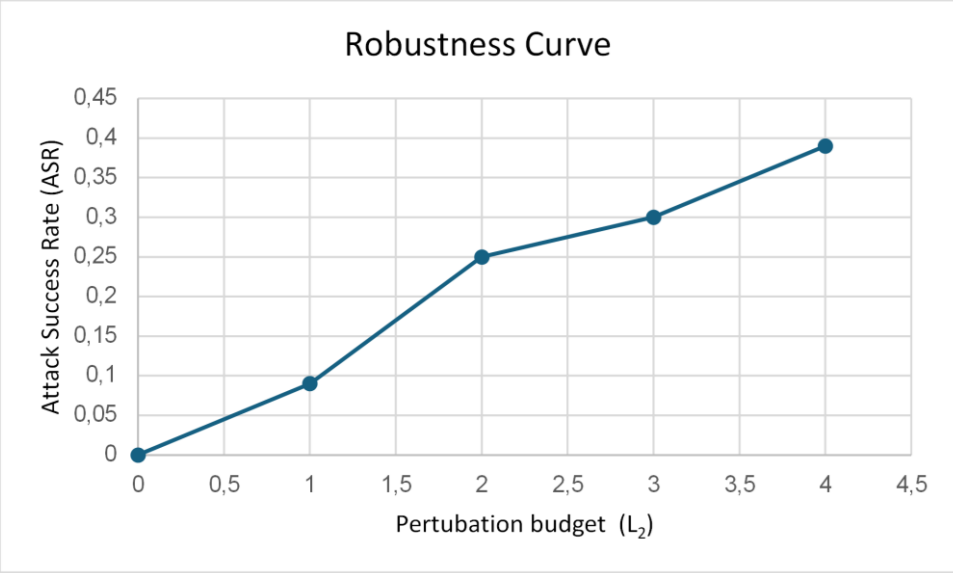
- Model inputs and outputs
- Query budget Q
- Perturbation norm Lp
- Perturbation budget ε

Not required

Architecture, gradients, hidden states or training procedure.

AI Robustness Evaluation Approach

Area Under Robustness Curve (AuRC)



Robustness curve RC : ε → ASR(ε)

- ε: allowed perturbation magnitude;
- ASR: attack success rate;
- Curve reflects performance across budgets, not a single operating point.

Normalized AuRC

$$AuRC_{norm} = \frac{1}{\varepsilon_{max} - \varepsilon_{min}} \sum_{i=1}^{n} \Delta_{\varepsilon_i} \frac{ASR_{i-1} + ASR_i}{2} \qquad \Delta\varepsilon_i = \varepsilon_i - \varepsilon_{i-1}$$

For a meaningful comparison

- Use the same perturbation norm;
- use the same budget interval;
- keep attack configuration and test protocol aligned.

Lower AuRC ⇒ greater robustness against the tested attack over the selected perturbation range.

AI Robustness Evaluation Approach

Intervals of rapid degradation

- Central finite difference

$$r_i \approx \frac{ASR_{i+1} - ASR_{i-1}}{2\Delta\varepsilon}, \qquad i = 1, \ldots, n-1.$$

- Endpoints

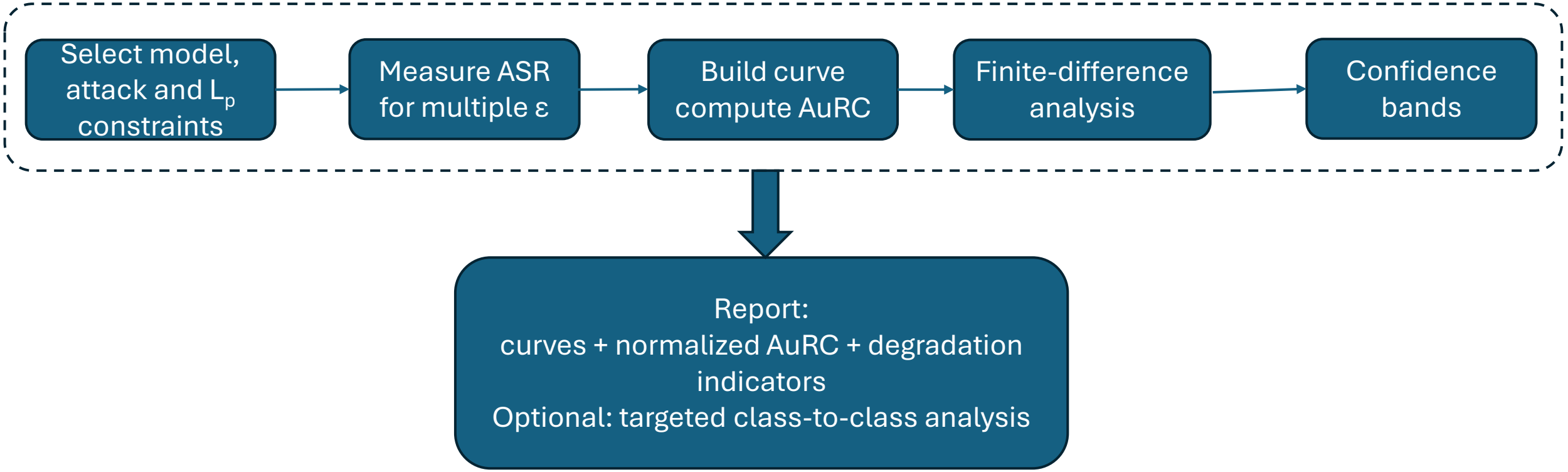
$$r_0 \approx \frac{ASR_1 - ASR_0}{\Delta\varepsilon}, \qquad r_n \approx \frac{ASR_n - ASR_{n-1}}{\Delta\varepsilon}$$

A high derivative indicates where attack success grows rapidly as the budget increases.

Confidence level of the evaluation scores

1. Run the attack independently n times over K budgets.
2. Each run produces an entire ASR trajectory.
3. Bootstrap-*resample whole trajectories*, not isolated curve points.
4. Use a supremum statistic to obtain simultaneous confidence bands.

The proposed evaluation workflow



Design principle:
One interface to compare different model implementations, without access to internal architecture or training details

Experimental validation: models and attacks

- Use Case: Robustness Evaluation of the Computer Vision Models

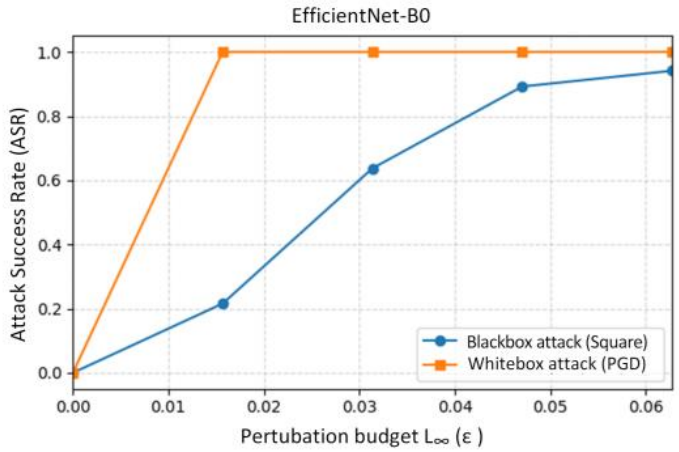
Model	Type	Training
EfficientNet-B0	CNN	Standard
EfficientNetB0_AdvTr	CNN	With adversarial samples
TinyViT-5M	ViT	Standard

Evaluation environment

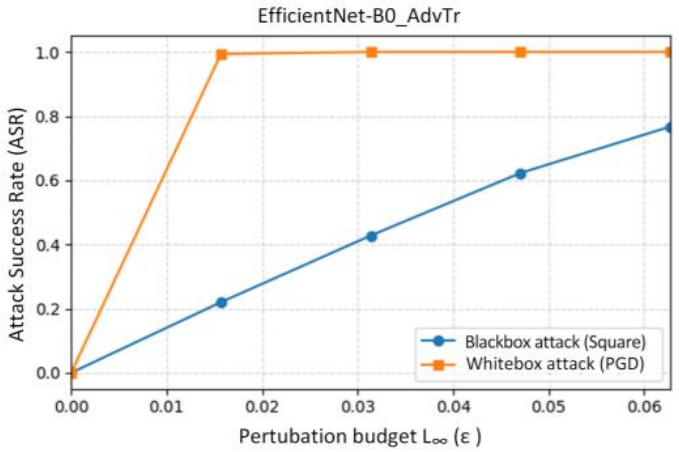
- ImageNet-1K classification domain
- Square Attack (black-box)
- L_2 and L_∞ perturbations
- Fixed evaluation subsamples

Experimental validation: determining query budget

50 queries

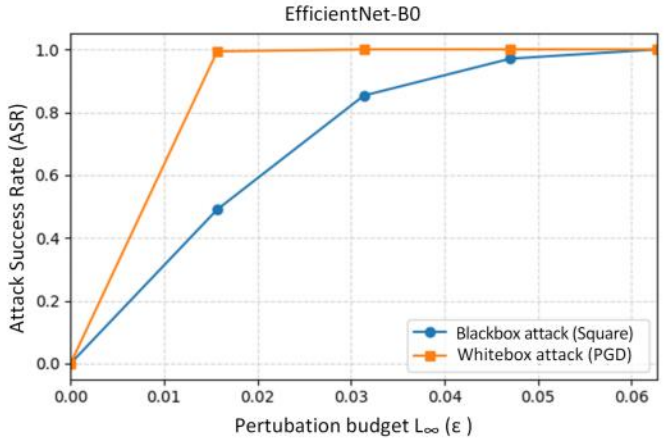


EfficientNet-B0, standard training

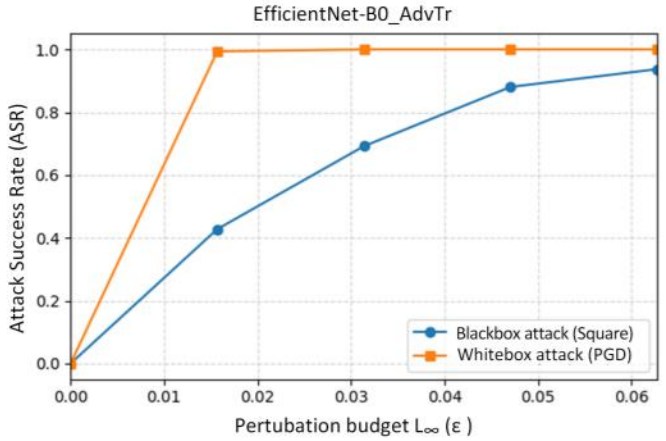


EfficientNet-B0, adversarial training

200 queries

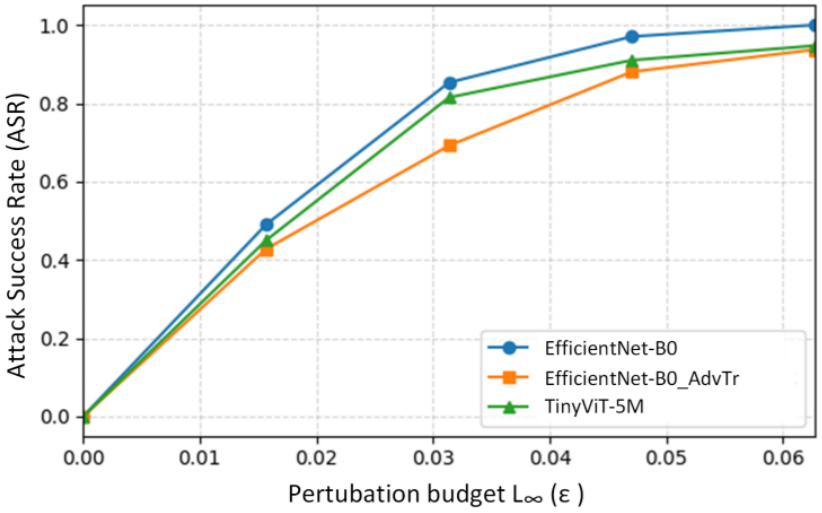


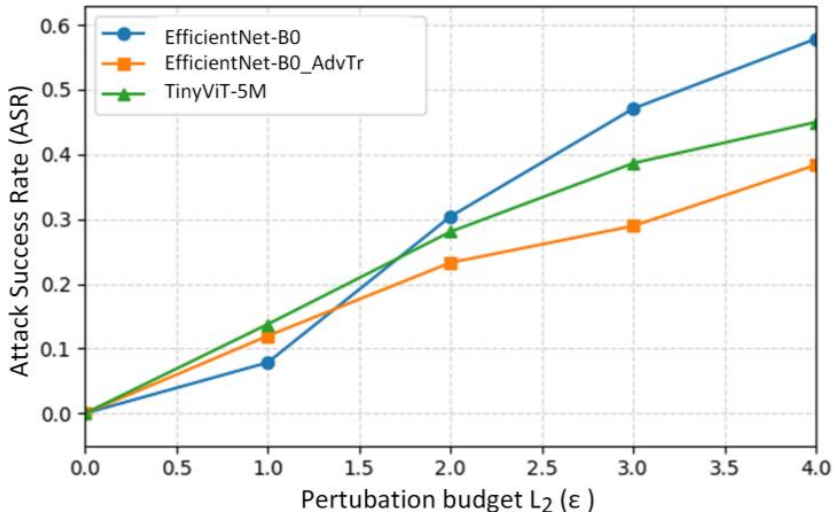
EfficientNet-B0, standard training



EfficientNet-B0, adversarial training

Experimental validation: curve comparison

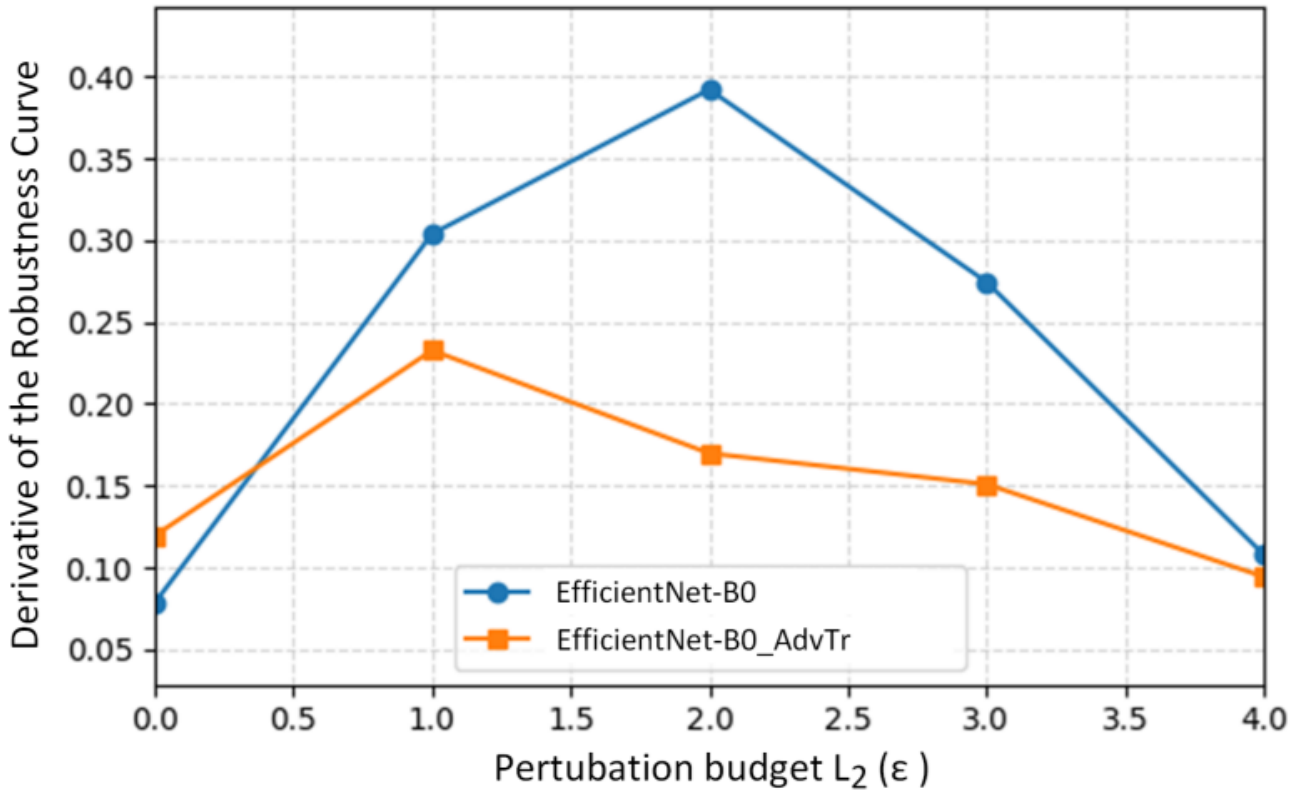




- Three model configurations.
- ASR increases with perturbation budget.
- Adversarially trained model generally has a lower ASR profile.

Model	L_∞ AuRC	L_2 AuRC
EfficientNet-B0	0.7034	0.2855
EfficientNet-B0_AdvTr	**0.6171**	**0.2083**
TinyViT-5M	0.6620	0.2573

Experimental validation: Derivative analysis



Observed pattern

- The standard EfficientNet-B0 shows a higher derivative peak.
- The peak appears near the L2 budget ε ≈ 2.
- Adversarial training reduces the peak in these tests.
- Requires additional evaluation in conjunction with formal robustness assessment

What the experiments show

Robustness differences

- Adversarial training lowers AuRC for both norms.
- TinyViT-5M occupies an intermediate position in the reported AuRC comparison.
- Query budget matters for black-box attack strength.

Additional diagnostic value

- Robustness curves reveal budget-dependent behavior.
- The area enables a compact model ranking within each norm.
- The derivative reveals rapid-ASR-growth intervals.

Limitations:

- One domain: ImageNet-1K image classification.
- Three model configurations.
- One black-box attack algorithm.
- Derivatives do not prove decision-boundary geometry.

Conclusions and next steps

Quantitative robustness evaluation methodology

- An architecture-agnostic evaluation interface based on black-box access.
- Normalized AuRC to summarize attack success across budgets.
- Finite differences to localize rapid degradation.
- Optional functional-bootstrap confidence bands.

Future validation & integration in AI testing framework

- More datasets and model families.
- Additional attack algorithms.
- Other data modalities.
- Repeated-run uncertainty studies and stability of derivative indicators.

Thank you!

Questions?

Aleksandr Sukhanov
sukhanov@etu.ru

Evgenia Novikova
novikova@comsec.spb.ru