INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY
HYDERABAD

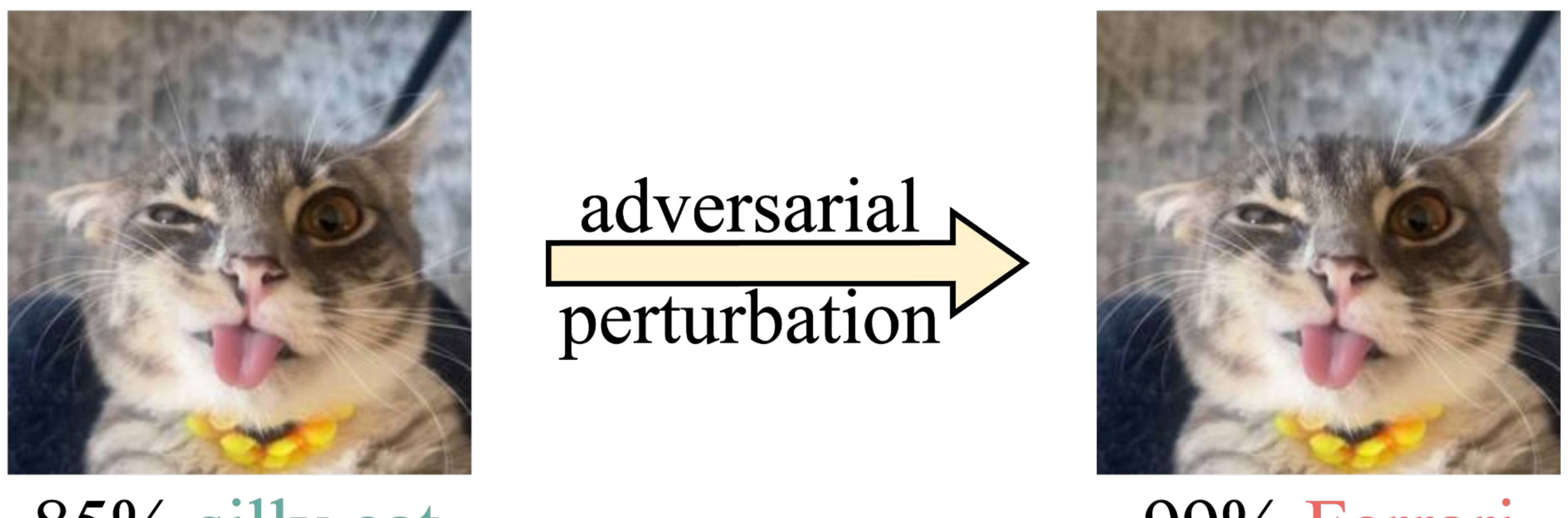
SYPY
SECURITY & PRIVACY
RESEARCH GROUP

Investigating Transferability of Adversarial Examples in Model Merging

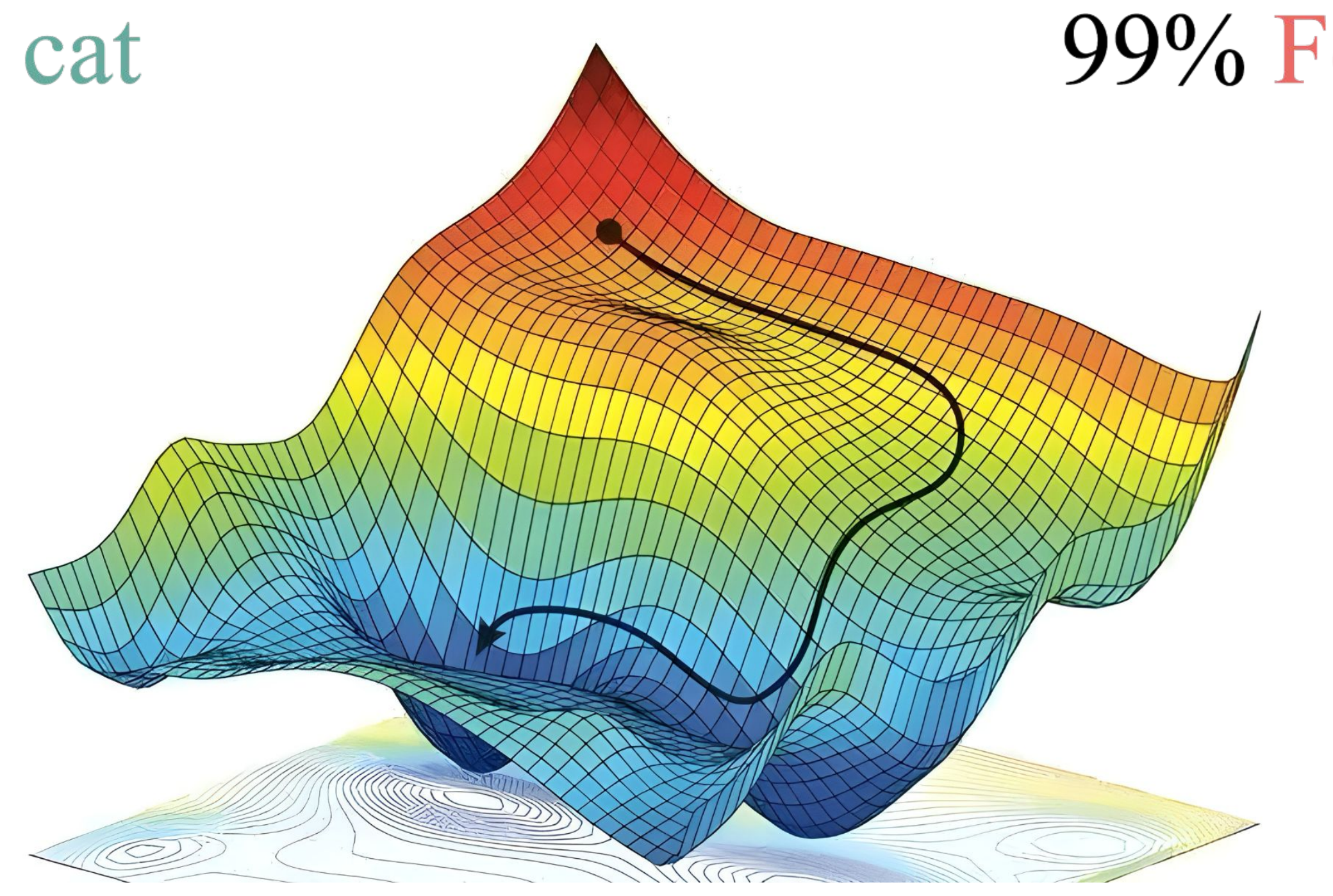
*Ankit Gangwal, **Aaryan Ajay Sharma***

1. What are Adversarial Examples?

- Adversarial examples are a subclass of adversarial attacks called **evasion attacks**.
- In 2022, a single case of unemployment fraud involving evasion attacks resulted in losses **exceeding $2.5 million**.
- **Cheap to produce** - gradient descent/ascent on confidence/loss w.r.t. image works.
- Adversarial examples tend to **transfer** to other models performing similar tasks.
- **Model merging** may mitigate transferability of adversarial examples.



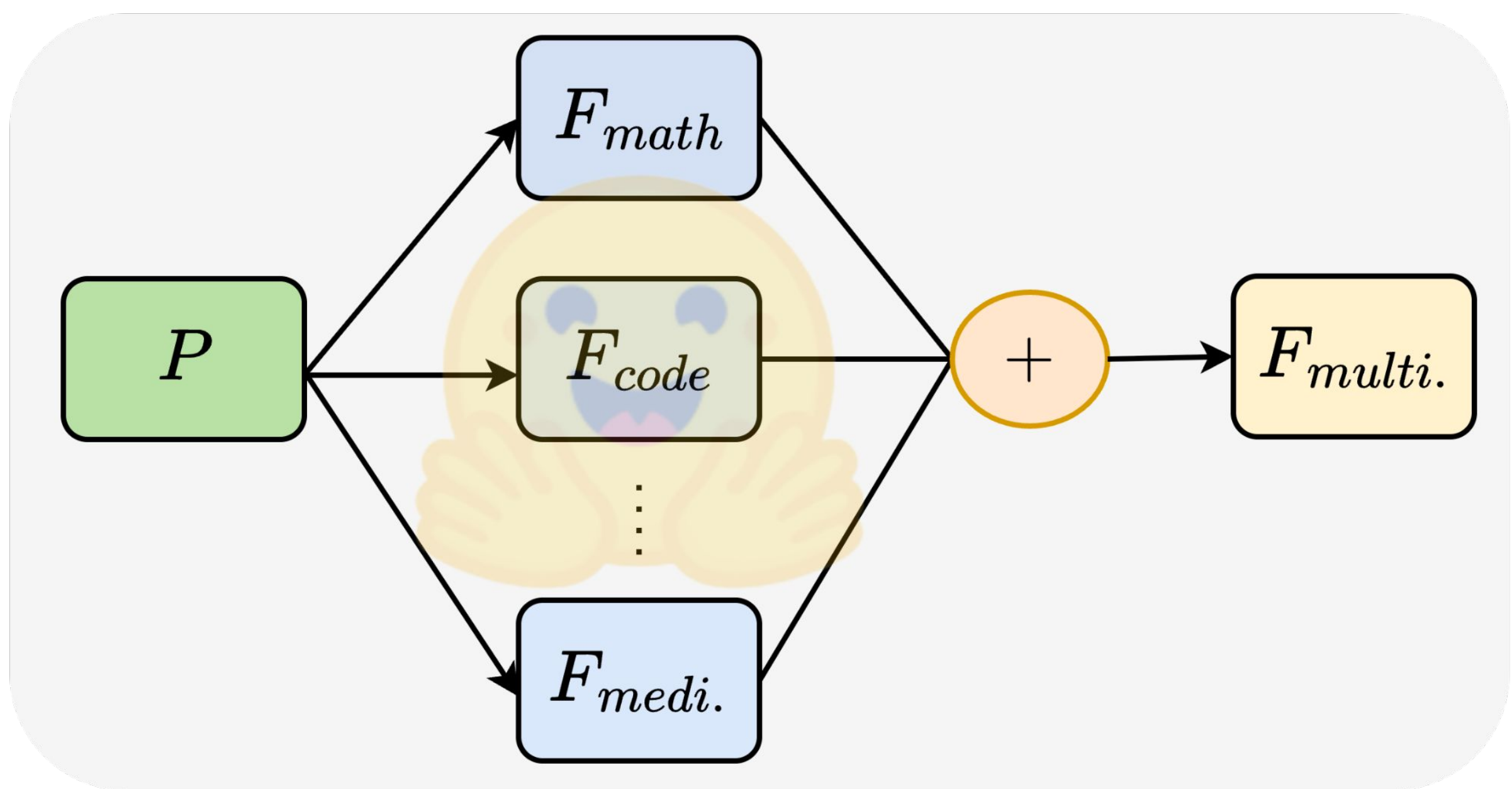
85% silly cat 99% Ferrari

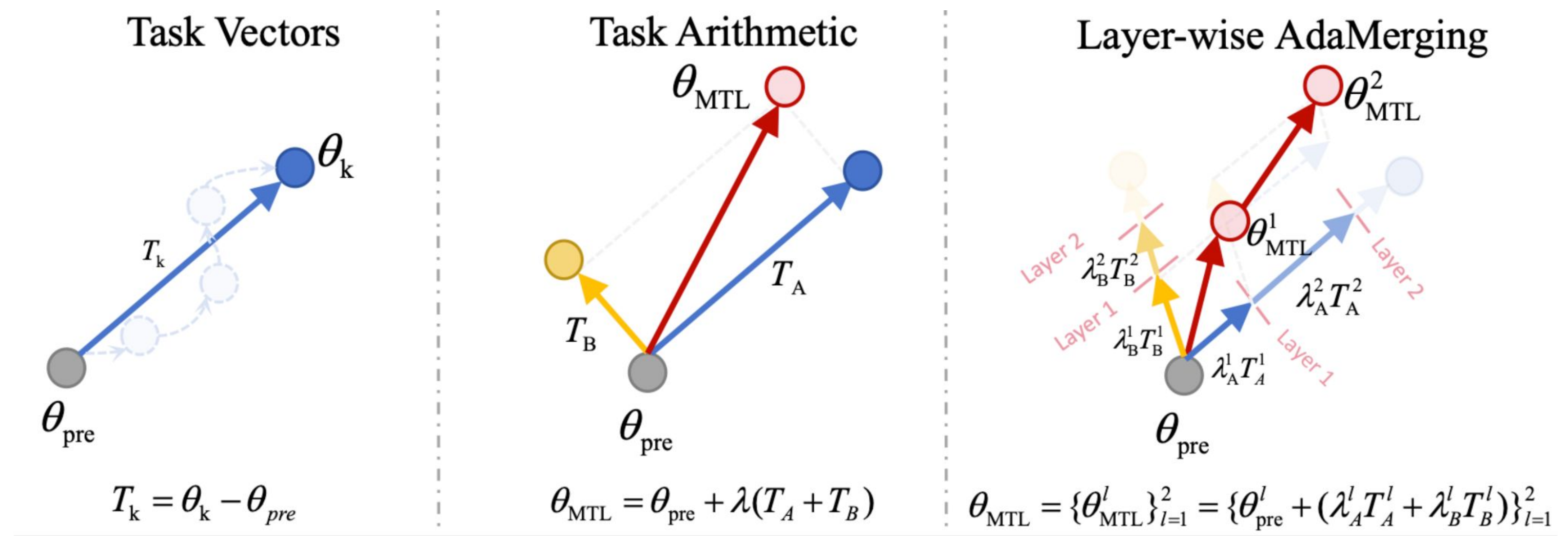


Gradient Descent

2. What is Model Merging?

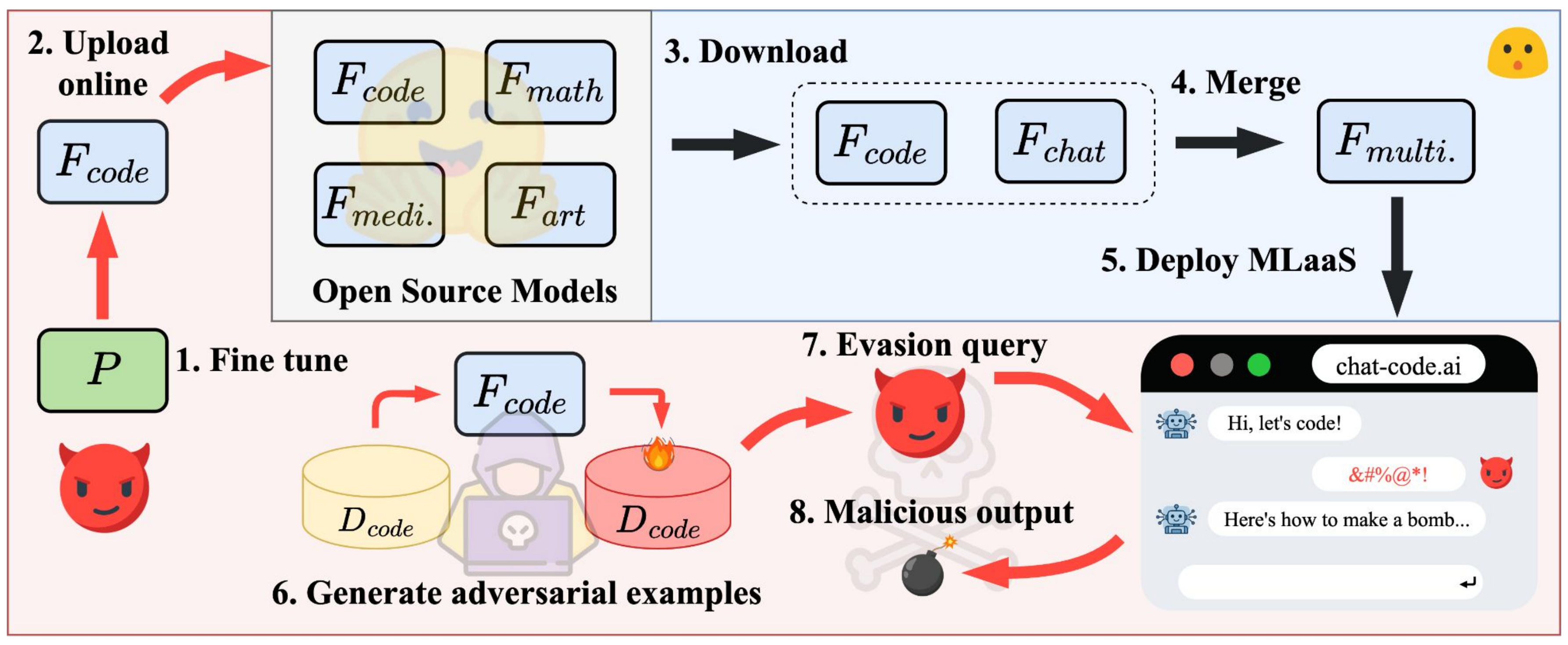
- Model merging: framework to **combine multiple fine-tuned models** into single model.
- **Alternative to multi-task learning**: no training data required to create multi-task models.
- **Over 30,000 merged models** available on Hugging Face.
- Pretrained model P fine tuned to get different models, F_{math}, F_{code}, etc.
- F_{math}, F_{code}, etc. are merged into a single model $F_{multi.}$.
- **Many methods**: Weight Averaging, Task Arithmetic, AdaMerging, etc.





3. (Transferable) Adversarial Examples in Model Merging

- **Higher transferability** expected when:
  - Target model shares architecture with surrogate model.
  - Surrogate model's decision boundary is similar to target model (high test accuracy).
- **Attack strategy of adversary** to attack an MLaaS employing a merged model:
  - **Fine-tune** pretrained model P with custom dataset and **upload online**.
  - **Bait benign users** to use adversary's fine-tuned model OR
  - **Obtain a surrogate model** to generate adversarial examples:
    - Download existing fine tuned F model suspected to be used in merged model OR
    - Download pretrained model if no fine tuned model available.
  - **Generate adversarial examples** on the obtained surrogate.
- **Key Advantage**: Higher transferability expected due to surrogate being P or F.



Overview of adversary's attack strategy

4. Results

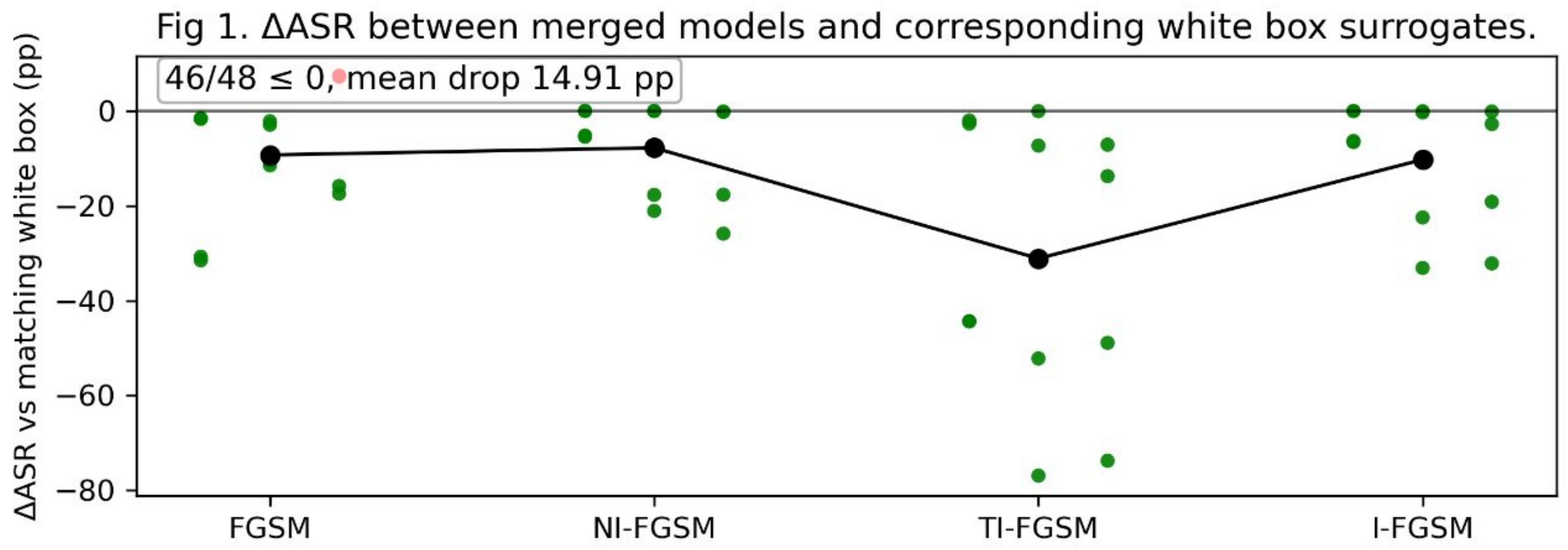


Fig 1. ΔASR between merged models and corresponding white box surrogates.

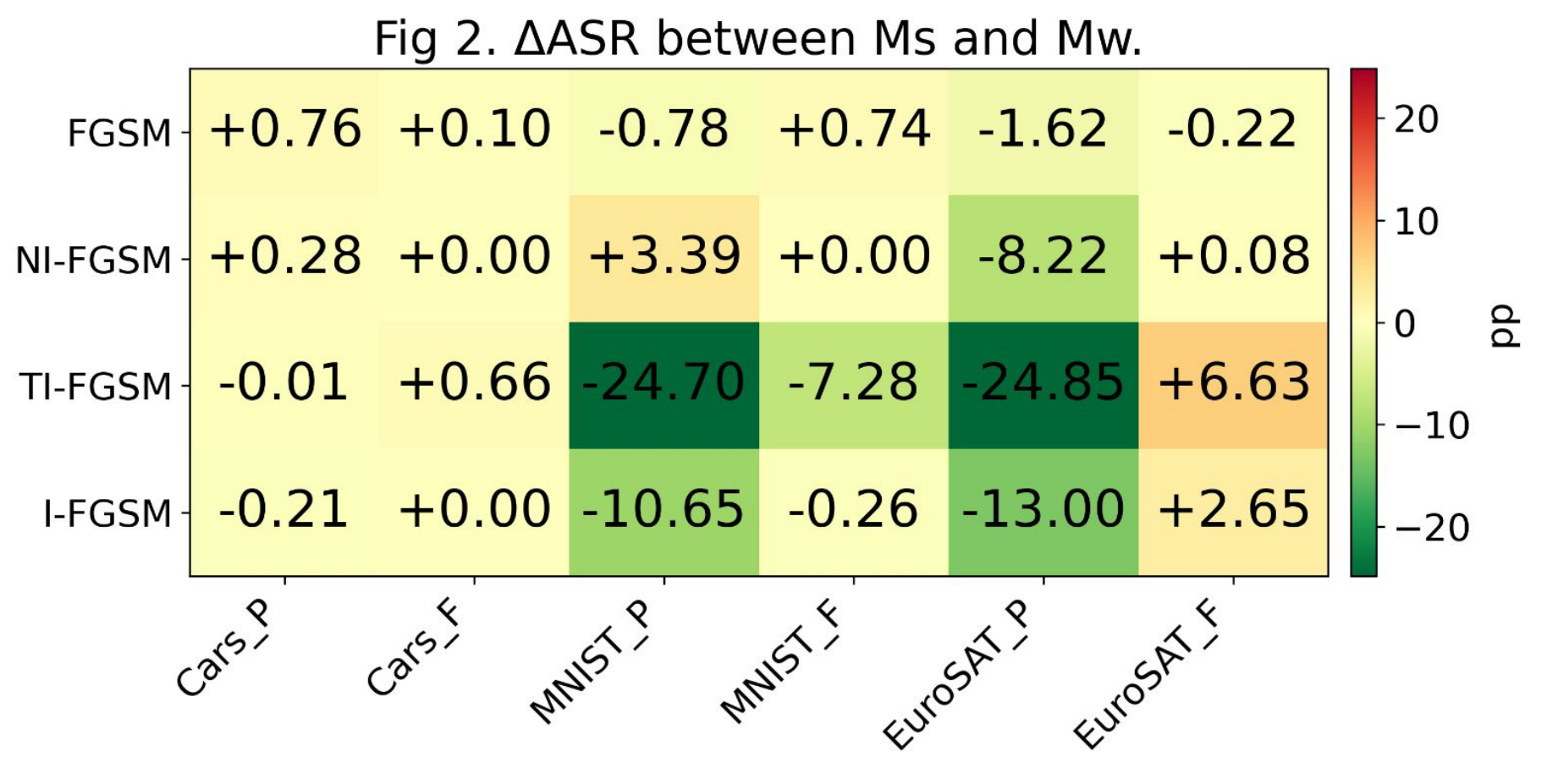


Fig 2. ΔASR between Ms and Mw.

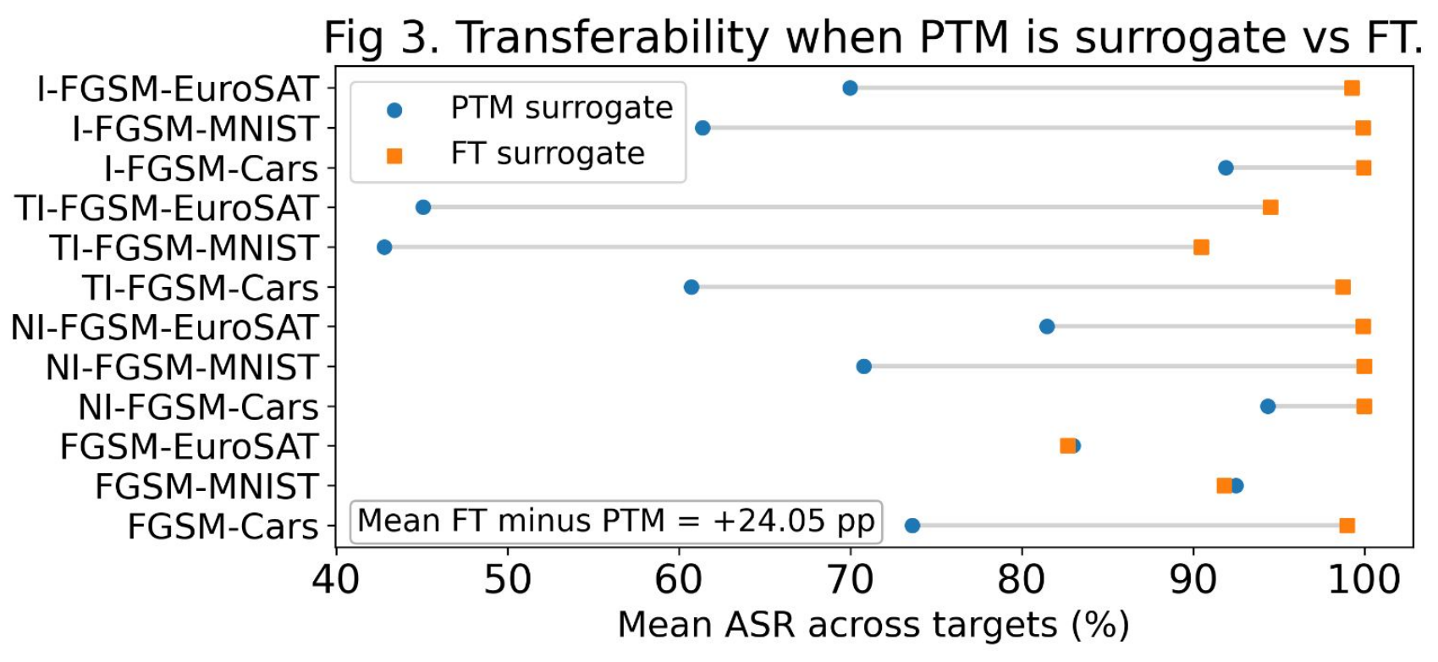


Fig 3. Transferability when PTM is surrogate vs FT.

- Result 1: ASR on merged (target) models ≤ ASR on surrogate models in most cases.
- Result 2: ASR decreases/remains in more than half (15/24) of the cases when a stronger merging method is used.
- Result 3: Pretrained model lowers transferability relative to fine-tuned counterpart.
- Shows model merging *could* provide "free lunch" of adversarial robustness.
- Future work involves theoretically and statistically validating the results.

➢ Vassilev A, Oprea A, Fordyce A, Anderson H, Davies X, Hamin M. "Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations." NIST Trustworthy and Responsible AI, NIST AI, 2025.
➢ Arora A, He X, Mozes M, Swain S, Dras M, and Xu Q. Here's a Free Lunch: Sanitizing Backdoored Models with Model Merge. In Findings of the Association for Computational Linguistics 2024. 15059–15075.

Contact

✉ aaryanajaysharma@gmail.com, gangwal@iiit.ac.in
🌐 https://sypy.iiit.ac.in/
📍 A3-113, CSTAR, IIIT, Gachibowli, 500 032, Hyderabad, India