## Problem Statement

The proliferation of generative AI and large-scale model deployment has intensified data privacy concerns. When sensitive user data, such as personal images, private documents, or confidential text, is transmitted to centralised servers for processing, it creates significant vulnerabilities, including data leakage, unauthorised access, and identity exposure.

Traditional cloud-based AI architectures require raw data transfer, fundamentally compromising user privacy and creating security risks. To address this critical challenge, our project aims to **enhance the privacy of AI systems themselves (Privacy of AI)**, as well as **use AI to defend user privacy and security.** It ensures that sensitive data never leaves the user's device while enabling continuous model improvement through collaborative learning. In addition, AI is used to help improve user privacy and security via censoring sensitive information and material.

## Our Solution

We introduce a federated learning (FL) framework with differential privacy (DP) proof-of-concept that enables privacy-preserving AI inference and training. Our system allows multiple clients to collaboratively improve a shared model while maintaining complete data locality and security.

- On-device models: Each user's device hosts a lightweight model (YOLO-based detector) that can detect and censor sensitive information (e.g., offensive symbols, PII-containing text, or inappropriate images).
- Federated learning: To avoid sending sensitive data to the server, each device's model trains locally using user-labelled data. Model updates and not the data itself are sent to the server's model.
- Differential privacy: To prevent reconstruction attacks or leakage of training examples, information about each data point in the dataset is withheld. Local updates are noised using Opacus DP mechanisms before aggregation.
- Server aggregation: A central server aggregates updates from all users to improve the global model and periodically pushes updated weights back to devices.

We developed hide & seek, a privacy-first photo gallery app demonstrating real-world applicability. Users can detect and censor sensitive objects in their photos manually or with the on-device machine learning model. All photos are stored locally and not shared with the server. However, the machine learning models on the user's device can train locally and collectively recognise new content without sharing sensitive information via FL.

In our proof-of-concept, federated learning is performed across 5 clients in our proof-of-concept, where differential privacy techniques are included. We fine-tune models to recognise human faces, which the model was previously unable to detect. We also show that the models do not forget previously learnt objects after fine-tuning.

# TikTok Ecosystem Applications

This framework could revolutionise TikTok's content safety by enabling on-device harmful content detection before upload. Users benefit from automatic filtering of inappropriate material, while TikTok gains insights into emerging threats through federated learning, all without accessing raw user content.

For regions with strict data protection regulations (GDPR, CCPA), federated learning enables on-device recommendation model training. TikTok could improve personalisation algorithms directly on user devices, aggregating behavioural insights without centralising sensitive user data, ensuring regulatory compliance while maintaining recommendation quality.

# Example Use Case

Consider the rapid emergence of a new offensive symbol across online platforms. Users may want this symbol automatically detected and censored in their image gallery or chat application.

- Local Detection: A user identifies the symbol on their device and labels it as offensive
- On-Device Training: Their local model fine-tunes to recognise this new threat
- Federated Aggregation: Model improvements are shared across the network without exposing raw images
- Global Distribution: The updated model is redistributed to all users
- Universal Protection: Every user gains automatic detection capabilities for the new symbol without compromising anyone's privacy.

This approach generalises to many domains: detecting faces or sensitive documents in photos, filtering licence plates in shared images, or redacting PII in text before it reaches cloud services.

# Development Tools & Stack

- Frameworks:
  - Flask: Lightweight backend service for orchestrating federated updates and model distribution.
  - Lynx: Cross-platform UI framework for building privacy-first mobile and web apps.
- Models & Libraries:
  - YOLO v8 Small (Ultralytics): Efficient object detection backbone, deployed on-device.
  - Opacus: PyTorch library that enables training models with differential privacy.
  - WIDER FACE dataset : Human face detection training data
  - COCO dataset: General object detection baseline training data
- Supporting Tools:
  - Python (PyTorch, HuggingFace)
  - GitHub: Version control and file sharing

# Model building approach

Our proof-of-concept uses YOLO v8 Small for its optimal balance of performance and computational efficiency, making it suitable for mobile deployment. We splice user data of human faces from the WIDER FACE dataset with known training data of YOLO v8 from the COCO dataset to fine-tune the models. There are 5 clients in the federated learning setup, and each client has about 300 instances of human face images and 300 instances of different object images from the COCO dataset.

The setup is then run over 6 training rounds, and the resulting consolidated model is able to learn to detect human faces with a significant confidence score. Our training method not only learns to detect a new feature, but also improves on the identification of old features. The macro average mAP50-95 remains at around 0.48 after fine-tuning. Running the workflow takes about 20 minutes on an NVIDIA A100 GPU.

# Relevance to Hackathon Theme

Privacy of AI: Our federated learning architecture ensures AI model training and improvement occurs without exposing sensitive user data, addressing fundamental privacy concerns in modern AI systems.

AI for Privacy: The system empowers AI models to autonomously detect, censor, and protect against privacy threats in real-world applications, creating a proactive privacy protection layer for users.

By combining these approaches, we create a comprehensive solution that not only protects user privacy, but leverages AI capabilities to enhance privacy protection across digital platforms.