FACULTY OF MEDICINE
Chulalongkorn University

DATE
17/09/2025

AI ENGINEERING

Infra basics & AI Product Lifecycle
From model training to real-world healthcare deployment

PRESENTED BY
Karin Huangsuwan,
AI Solution Consultant
4Plus Consulting Co., Ltd.

WHY AI ENGINEERING?

Training a model = 1 step, but deployment needs more.

Accessible

01

Reliable

02

Safe

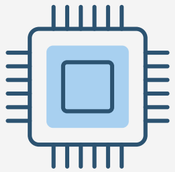
03

AI APPLICATION INFRASTRUCTURE

Chatbots, Predictive Analytics, Recommenders — APPLICATIONS

Pretrained models, fine-tuned models — MODELS

Data pipelines, storage, ETL, feature stores — DATA



GPUs/TPUs, cloud infrastructure, orchestration — COMPUTE

Monitoring, CI/CD for ML, deployment pipelines — MLOps

AI APPLICATION INFRASTRUCTURE

APPLICATIONS Chatbots, Predictive Analytics, Recommenders

- What the user sees, but powered by deep infrastructure
- Must integrate seamlessly with business workflows

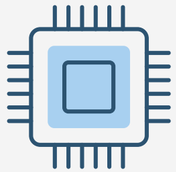
MODELS Pretrained models, fine-tuned models

- Need retraining & fine-tuning for changing data distributions
- Risk of model drift; requires periodic validation

DATA Data pipelines, storage, ETL, feature stores

- Data quality is the bottleneck; 80% of ML work is data prep
- Handling unstructured data (text, images, logs) at scale

AI APPLICATION INFRASTRUCTURE



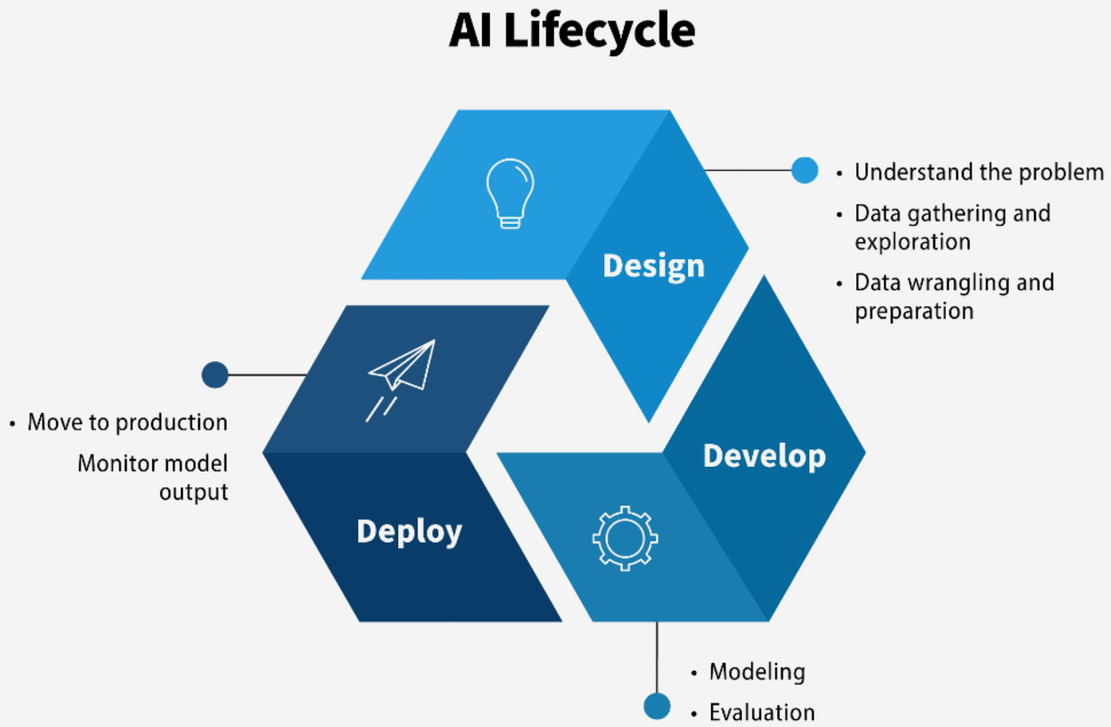
COMPUTE GPUs/TPUs, cloud infrastructure, orchestration

- Training large models is compute-bound, requiring scaling strategies
- Optimize cost with hybrid/on-demand cloud

MLOps Monitoring, CI/CD for ML, deployment pipelines

- Keeps models reliable post-deployment, ensures reproducibility
- Automate monitoring & retraining to avoid “silent failures”

THE AI PRODUCT LIFECYCLE



https://coe.gsa.gov/coe/ai-guide-for-government/understanding-managing-ai-lifecycle/

DESIGN

Shape the foundation for AI solutions

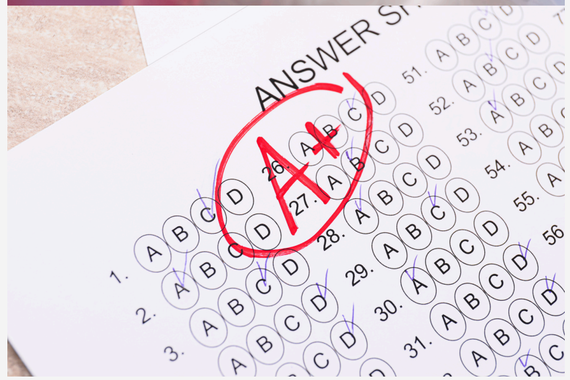
- Understand the Problem
  - Define business objectives & success criteria.
  - Confirm problem is suitable for AI.

- Data Gathering & Exploration
  - Identify available datasets, check quality.
  - Plan how to access, store, and use data.

- Data Wrangling & Preparation
  - Clean and transform raw data into usable formats.
  - Handle missing values, normalization, feature creation.

DEVELOP

Build and validate the AI system



- Modeling
  - Experiment with different algorithms.
  - Train, test, refine — expect multiple iterations.
  - Ensure infrastructure supports compute needs.

- Evaluation
  - Use relevant metrics (accuracy, fairness, bias).
  - Test with unseen data to ensure generalization.
  - Confirm the model meets business objectives.

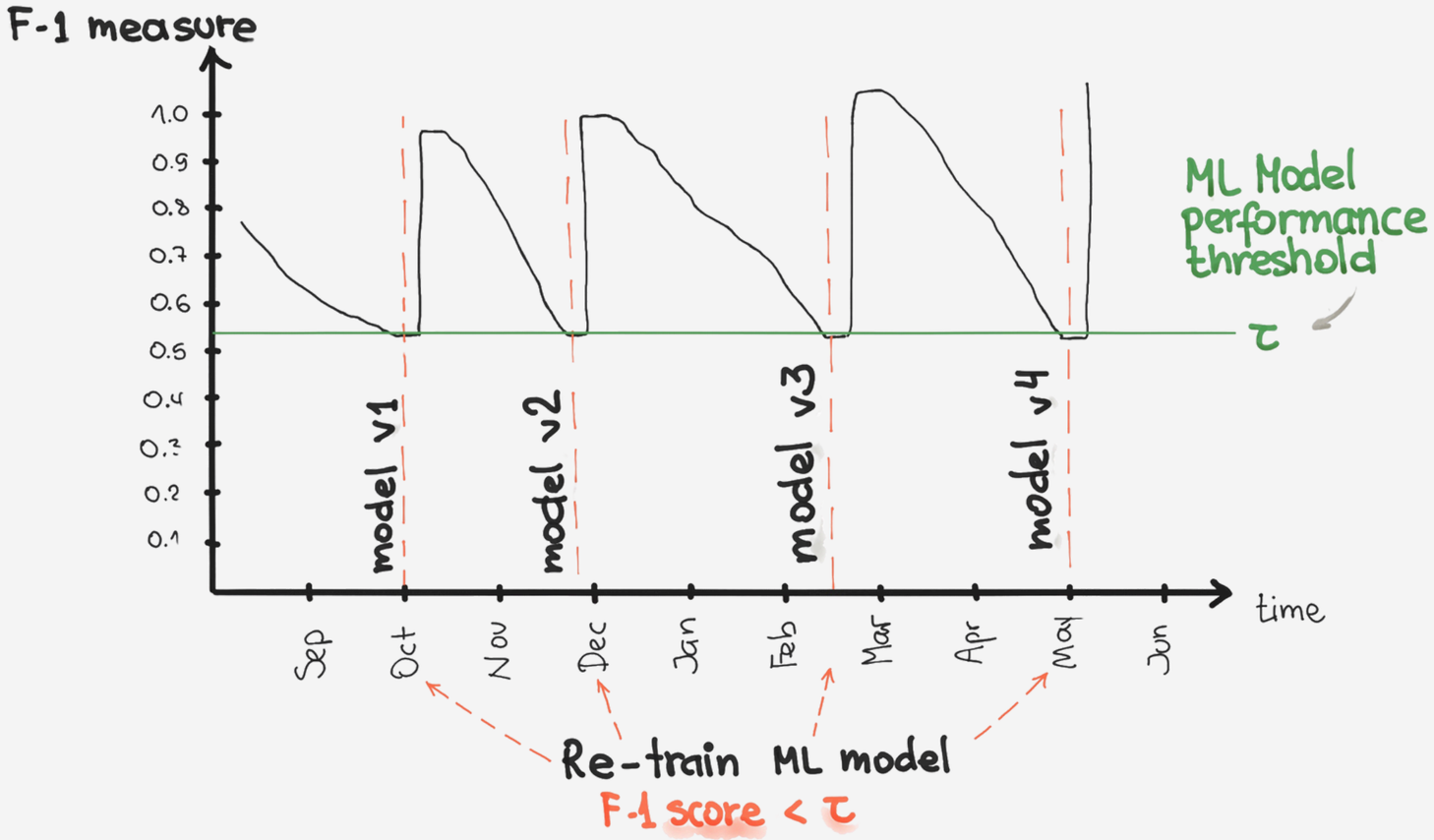
DEPLOY

Deliver real-world value and maintain performance

- Production Transition
  - Move best-performing model into live environment.
  - Handle new and unseen input data effectively.

- Monitoring & Maintenance
  - Track outputs for drift, bias, and errors.
  - Continuously update and retrain models as needed.
  - Apply agile, iterative improvements.

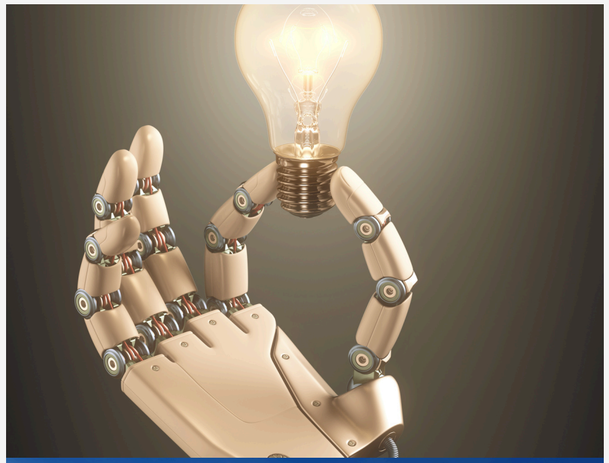
WHY RETRAIN?



- **Model performance decays over time**: an ML model's accuracy (F1 score) gradually drops as data changes.
- **Threshold for action:** When performance falls below a set limit (τ), it's like vitals crossing a danger zone — intervention is needed.
- **Retraining as treatment:** Retraining the model restores performance, similar to giving another dose of medication to bring levels back up.

https://ml-ops.org/content/mlops-principles

DEPLOYMENT TOOL: FastAPI



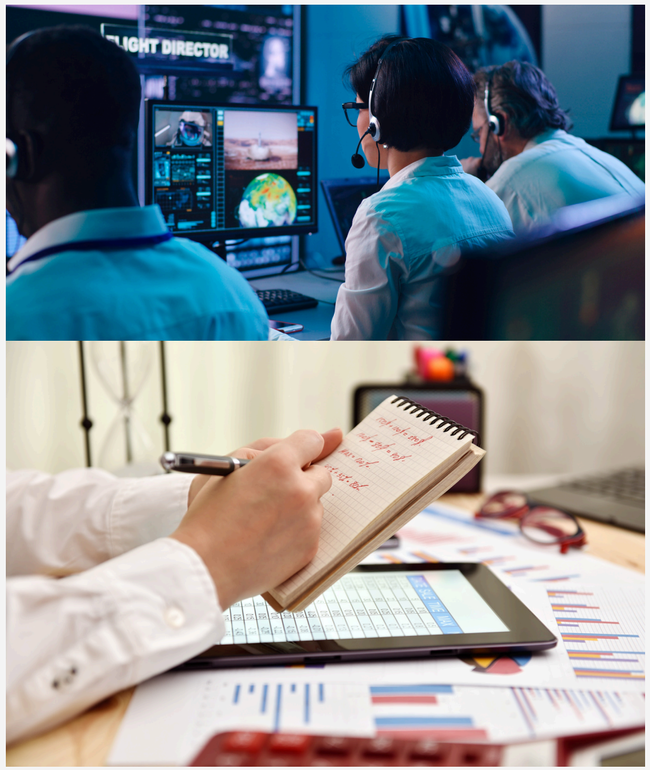
What is FastAPI?

- A tool to turn ideas into apps quickly.
- Lets developers create simple websites or services that computers can talk to.
- Comes with a “self-explanatory manual” (automatic docs).

Why it’s Useful

- Fast → Makes apps respond quickly.
- Easy → Fewer steps to go from idea → working app.
- Ready to Share → Runs smoothly on cloud platforms (AWS, Google, Azure).
- Scalable → Can grow from small demo to full business service.

MONITORING & RETRAINING TOOL: MLFlow



What is MLflow?

- A tool to keep track of machine learning projects.
- Think of it as a “control center” for experiments, models, and results.
- Helps teams organize, compare, and share their work.

Why it’s Useful

- Tracks Experiments → Keeps a record of what worked and what didn’t.
- Manages Models → Stores different versions safely.
- Easy Deployment → Push models into apps or services with less hassle.
- Collaboration → Teams can work together without losing progress.

Conclusion

- AI Engineering = **more than training** → it’s about deploying, monitoring, and evolving
- Lifecycle (**Design** → **Develop** → **Deploy**) keeps AI reliable and valuable
- Tools like **FastAPI** & **MLflow** make AI practical, scalable, and sustainable
- Goal: Deliver **accessible, reliable, safe** AI that adapts to real-world change