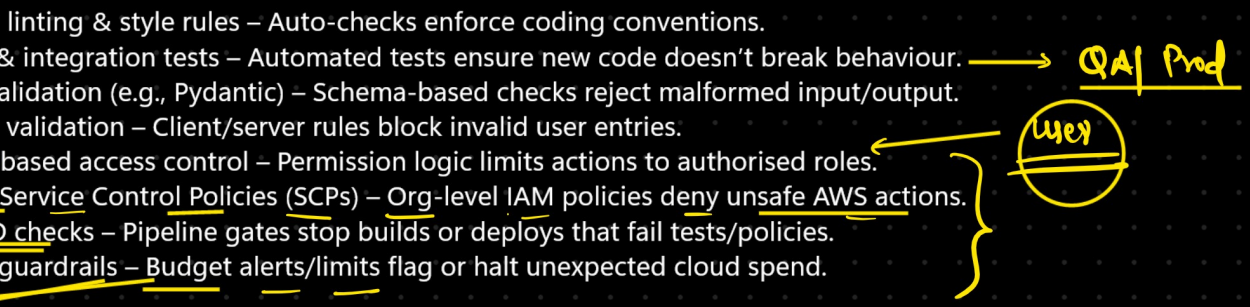
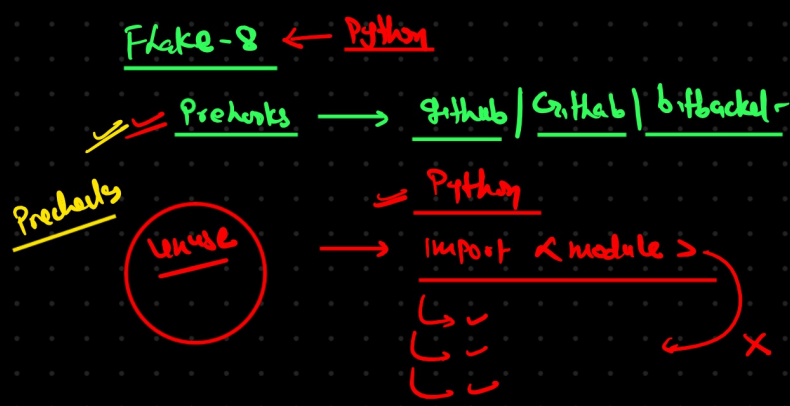
Guardrails

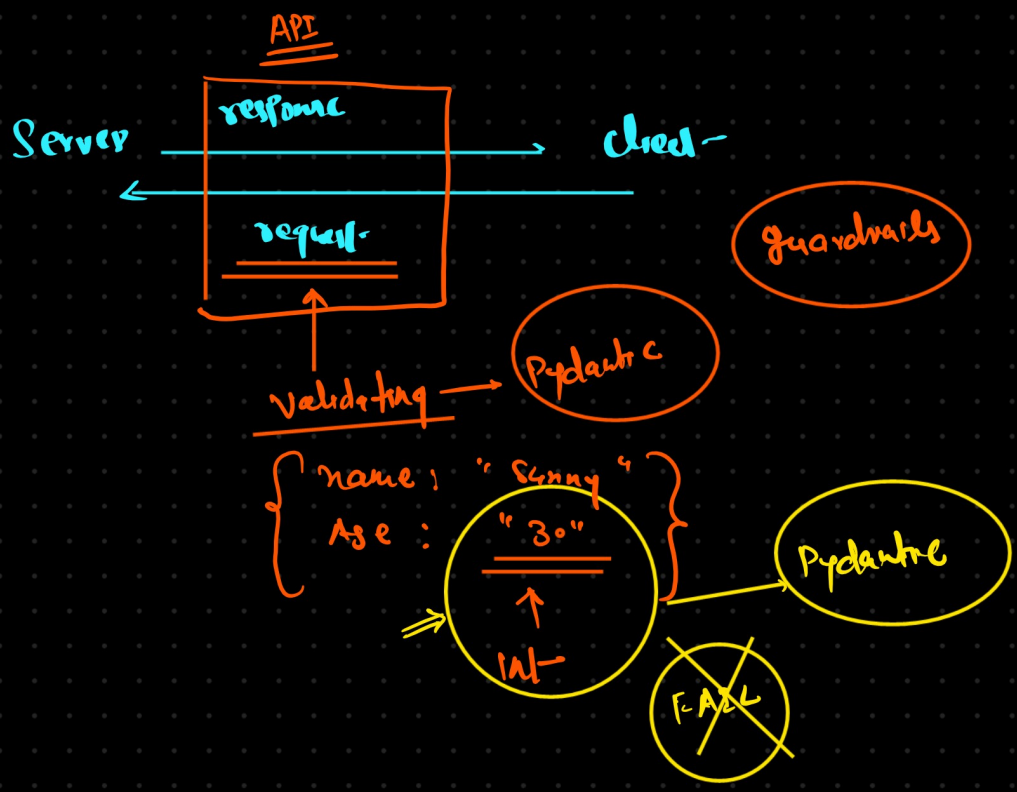
Guardrails mean Safety boundaries, protective limits, or control mechanisms.

In Software Engineering / Programming / Web Development / DevOps / Cloud & Infrastructure

- Code linting & style rules – Auto-checks enforce coding conventions.
- Unit & integration tests – Automated tests ensure new code doesn't break behaviour.
- API validation (e.g., Pydantic) – Schema-based checks reject malformed input/output.
- Form validation – Client/server rules block invalid user entries.
- Role-based access control – Permission logic limits actions to authorised roles.
- AWS Service Control Policies (SCPs) – Org-level IAM policies deny unsafe AWS actions.
- CI/CD checks – Pipeline gates stop builds or deploys that fail tests/policies.
- Cost guardrails – Budget alerts/limits flag or halt unexpected cloud spend.







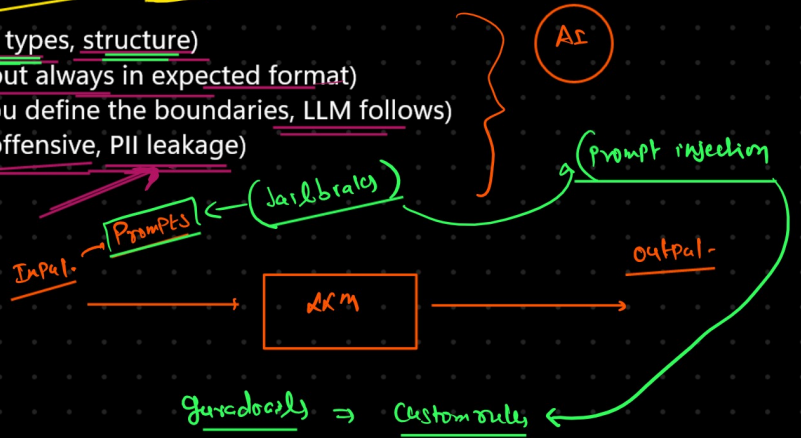
Guardrails in AI

Guardrails in AI (LLM Applications) are the safety boundaries—protective limits and control mechanisms—that keep a model's behavior on the right, safe, and controlled path.

Definition: Guardrails are rules + safety nets that ensure LLMs stay safe, correct, consistent, and under your control — just like a seatbelt keeps you safe in a fast car.

They ensure:

- Correctness (data types, structure)
- Consistency (output always in expected format)
- Controllability (you define the boundaries, LLM follows)
- Safety (no toxic, offensive, PII leakage)



Chatbot ⇒ Give me a Product review in Json.

→ LLM Response: Product is great and the rating is 4.5.

Error

Guardrails ⇒ { review: Product is great, rating: 4.5 }

Medical bot :- User question : I have chest pain. Suggest me medicine.

LLM Response : take Aspirin and lie down on bed. ✗

Emergency

Patient condition — Not safe response ✗

After implement the Guardrails — take this medicine ------

LLM response : I am not a medical professional. Please consult with a doctor

Safe

Guardrails-AI: https://www.guardrailsai.com/docs
OpenAI Guardrails: https://openai.github.io/openai-guardrails-python/
NeMo Guardrails: https://docs.nvidia.com/nemo/guardrails/latest/index.html
LMQL (Language Model Query Language): https://lmql.ai/docs/