Real-Time AI Toolkit

For FinTech Founders of Tech Champions 5 Jeddah Cohort

The 3 Golden Levers

1. **Model Choice**: Smaller, task-specific models beat giant LLMs for production.
   → Use distilled models (DistilBERT) or tabular models (LightGBM/XGBoost).

2. **Infra Tricks**: Speed comes from caching, batching, and approximate nearest neighbor (ANN) search.
   → Cache repeat results (Redis), batch queries, use FAISS/Pinecone for lookups.

3. **Decision Design**: Decide fast when confident; defer when not.
   → Confidence thresholds, fallback rules, or human-in-loop (HITL).

Key Trade-Off

Real-time AI is always a balance between **Accuracy**, **Latency**, and **Cost**.
In FinTech, a timely **92% accurate decision** can beat a late **98%** one.

Quick FinTech Patterns

- **BNPL (Gersh / Sariat)** → Tabular model + rules; cache approvals for 10–30 min; manual review if uncertain.

- **Trading (Qunfin)** → Precompute features; stream deltas; use ANN search; aim for p95 < 200ms.

- **Phishing (TPhish)** → DistilBERT email classifier at edge; escalate to LLM only if uncertain.
- **Donations (Eyrad / Halala)** → Low-latency donor intent classifier; cache recs; explain decisions.
- **Compliance (Adalah Chain / Meta Works / Invora)** → Rules first; ML for gray zones; always log decisions.

Starter Tools

- HuggingFace Distil models (text tasks)
- LightGBM / XGBoost (tabular risk scoring)
- Redis (caching) + FastAPI (prototyping)
- FAISS / Pinecone (vector search)

Action Step

This week:

1. Pick one decision in your product.
2. Identify its bottleneck.
3. Apply one lever (Model, Infra, or Decision).

Created for Tech Champions 5 Jeddah Cohort by Adeen Atif
www.adeenatif.com/rt-ai