FIELD JOURNAL: DOCUMENT INSIGHT DEPLOYMENT
Day 1: We connected the service to a dusty archive of scanned contracts.
The OCR spun up, redis warmed, and embeddings began to trickle into FAISS.
Latencies hovered around 2.8s uncached â good, but we wanted better.
We pinned the QA model to the ’best’ preset; confidence climbed on legal clauses.

Day 2: We toggled to DistilBERT for a few hours to profile speed.
Cache hits jumped after we re-asked common billing questions.
One user searched for "termination notice" and got the right page instantly.
We celebrated with coffee and a cleaned-up Swagger page.

Day 3: A surprise request: highlight entities across all scanned NDAs.
spaCy models (en, hr) tagged names, dates, and orgs with solid precision.
We exported JSON snippets with page numbers, then bulk-shared via the UI.
Performance note: after warming, median ask latency hit 0.9s cached.