LLM Safety & Launch Package FastAPI + Redis Example Comprehensive Guide, Code, Deployment, Monitoring, and Legal Checklist

Author: Generated for Ray (user)

Date: 2025-09-28

Executive Summary

This document provides a comprehensive, production-ready safety package for launching a locally-trained LLM. It includes: - A defense-in-depth architecture. - Runnable example code (FastAPI) implementing input intent detection, model steering, post-generation filters, logging, rate limits, human review queue, and admin endpoints. - Deployment assets (Dockerfile, docker-compose, Kubernetes manifest). - Monitoring, alerting, and incident response guidance. - Legal, ethical, and red-team testing checklist. - Instructions for integrating HuggingFace/ONNX / custom classifiers and for extending to real moderation APIs.

Architecture (Defense-in-Depth)

Layers: 1) Authentication & Access Control (API keys, OAuth, KYC for high-risk flows). 2) Input Sanitizer & Intent Classifier (fast heuristics + ML classifier). 3) Capability Gate (disable risky tools; require additional checks for tool access). 4) Generation with System Steering (system prompts, refusal examples, constrained decoding). 5) Output Filters (regex deny-lists, ML content moderation). 6) Human-in-the-loop Review Queue for borderline/high-risk outputs. 7) Logging, Alerting, Telemetry, and Incident Response. 8) Rate-limiting, Quotas, and Reputation scoring for users.

Implementation Overview (what's included)

Files included (in this PDF): - safe_api.py : main FastAPI app showing the pipeline (auth, intent check, generation stub, post-filter). - classifiers.py : example placeholders for intent_classifier and content_classifier (simple heuristics + hooks). - human_review_queue.py : minimal human-review queue using Redis. - admin_app.py : admin endpoints to list/review flagged items. - docker/Dockerfile : container image for the API. - docker/docker-compose.yml : development compose with Redis. - k8s/deployment.yml : example Kubernetes deployment + service + Redis StatefulSet suggestion. - monitoring.md : Prometheus + Grafana + alert rules suggestions. - legal_and_compliance.md : checklist and notices. - README.md : quick start.

safe_api.py - key excerpt

```
# Intent detection + post-filter + generation (snippet)
intent = intent_classifier(prompt)
if intent["flag"]:
    log_incident("blocked_input", {...})
    raise HTTPException(status_code=400, detail=REFUSAL_MSG)

generated = generate_from_model(prompt, user_id)
post = content_classifier(generated)
if post.get("unsafe"):
    log_incident("blocked_output", {...})
    raise HTTPException(status_code=500, detail=REFUSAL_MSG)
```

classifiers.py - excerpt

```
# classifiers.py - pattern-based and ML hook
def intent_classifier(text):
```

```
    for p in BLOCK_PATTERNS:
        if re.search(p, text, flags=re.I):
            return {"flag": True, "reason": "deny_pattern_match", "score": 0.99}
    # call ML model here...
    return {"flag": False, "reason":"", "score": 0.0}
```

human_review_queue.py - excerpt

```
# human_review_queue.py - enqueue / dequeue
def enqueue_for_review(item):
    item["id"] = str(uuid.uuid4()); item["ts"] = time.time()
    r.rpush("review_queue", json.dumps(item))
```

Dockerfile - excerpt

```
# Dockerfile (see full file in repo)
FROM python:3.11-slim
WORKDIR /app
COPY requirements.txt .
RUN pip install --no-cache-dir -r requirements.txt
CMD ["uvicorn", "safe_api:app", "--host", "0.0.0.0", "--port", "8080"]
```

Monitoring & Incident Response

See monitoring.md for Prometheus/Grafana guidance, alert rules, and incident playbook.

Legal & Compliance

See legal_and_compliance.md for a checklist including AUP, ToS, retention policy, and counsel recommendations.

Appendix A: Full safe_api.py

```
[Full code file omitted in this PDF view for brevity in generation environment.
In the downloadable package, the repository includes full files:
- safe_api.py
- classifiers.py
- human_review_queue.py
- admin_app.py
- docker/Dockerfile
- docker/docker-compose.yml
- k8s/deployment.yml
- monitoring.md
- legal_and_compliance.md
- README.md
]
```