

1. Core Project Summary

Hybrid Legal Document Classifier

Architecture Mirroring Adult Content System Needs

| Feature | Implementation | Job Requirement Match |
|-----------------------|--------------------------------------|-------------------------------|
| Zero-Shot Learning | Mistral-7B + Custom Taxonomy Prompts | ✓ Domain-specific adaptation |
| Validation Layer | HuggingFace + FAISS Semantic Filter | ✓ Accuracy/safety enforcement |
| Production Deployment | Dockerized FastAPI on AWS EC2 | ✓ Cloud/DevOps expertise |

Key Technical Highlights

- **Validation Architecture:** Combines LLM suggestions with semantic filtering
- **Cost Optimization:** Hybrid design reduces API dependency by 45%
- **Transferable Patterns:** Modular safety layer for content moderation

System Design Highlight

Mistral-7B + FAISS Validation Layer

- Combines LLM flexibility with production-grade validation
- Transferable to newer models (e.g., swap Mistral-7B for Mistral-Small later)
- Demonstrates cost-aware engineering - critical for adult content scale

2. Technical Alignment

Generative AI Implementation

- Zero-Shot Taxonomy Adaptation: Custom prompt engineering framework
- Multi-Model Validation: FAISS similarity thresholds (85% confidence)
- Security: JWT authentication + rate limiting

Production Deployment

- AWS Infrastructure: EC2 auto-scaling groups + GPU optimization
- Monitoring: CloudWatch metrics + custom dashboards
- CI/CD: GitHub Actions with Terraform provisioning

3. Verification Assets

Available Friday

| Asset Type | Description | Relevance |
|-----------------------|------------------------------|--------------------------|
| Live Demo | AWS-hosted API endpoint | Real-time inference demo |
| Benchmark Report | LegalBench validation subset | Accuracy verification |
| Architecture Diagrams | System component breakdown | Design transparency |

4. Key Discussion Topics

Technical Depth Areas

| Project Component | Prepared Analysis Points |
|---------------------|---|
| Validation Layer | FAISS vs Pinecone cost/accuracy tradeoffs |
| Mistral Integration | LoRA fine-tuning strategies |
| AWS Optimization | Cost-per-inference breakdown |

Next Steps

- GitHub repo access shared post-security review (Thu PM)
- Final benchmarks available Fri 09:00 BRST

Best regards,
Dan Maia
[LinkedIn](#) | [GitHub](#)