Case Study: GenAI Usage & Evaluation Monitor for SMBs on Azure

# 1. Problem

As SMBs adopt GenAI on Azure, especially in the Legal Tech domain, they often combine Azure OpenAI Service (AOAI) with open-source LLMs like LLaMA and Mistral. Despite Azure’s robust platform-level cost controls and content filtering, these teams face significant gaps:  
  
- No unified app-level visibility across AOAI + OSS model usage  
- Inability to trace GenAI cost drivers to specific prompts or agents  
- Lack of post-hoc output evaluations for hallucinations, PII leaks, or safety  
- Small teams with limited MLOps capacity to build such monitoring stacks

# 2. Solution

We built an open-source tool that provides:  
- Token + model usage tracking for AOAI and OSS models  
- Post-hoc AI evaluations for safety, hallucinations, and quality  
- Cost estimations using user-provided pricing configs  
- Streamlit dashboard and alerting system  
- Audit reports for compliance-readiness

# 3. Target Users & Use Cases

Target: Legal Tech SMBs on Azure using GenAI for document drafting, contract analysis, or legal research.  
  
Use Cases:  
1. Analyze cost drivers to optimize prompts and agents  
2. Run automated evals to flag hallucinations and PII risks  
3. Export audit-ready reports to show responsible AI practices  
4. Receive alerts when token usage or risk exceeds thresholds

# 4. Architecture

The tool is composed of:  
- Python middleware intercepting LLM API calls  
- Eval engine running Presidio, Detoxify, heuristics, Ragas  
- Cost estimator module  
- SQLite/Postgres storage backend  
- Streamlit dashboard or FastAPI + Grafana UI  
- Slack/email alerting system

# 5. OSS Stack

Tech Stack Includes:  
- v0.dev for UI prototyping  
- Decagon for eval orchestration (optional)  
- tiktoken and tokenizers for usage counting  
- Presidio, Detoxify, Ragas for AI evals  
- Docker for deployment

# 6. Impact & Differentiation

For Legal Tech SMBs, this tool delivers:  
- Up to 30% token cost savings via optimization  
- Detection of unsafe or hallucinated outputs before they cause harm  
- Compliance support via exportable eval logs  
- Minimal deployment overhead, no heavy infra required  
  
Compared to Azure’s native tools and other OSS libraries, this product uniquely unifies:  
- AOAI + OSS tracking  
- Token cost + output safety evals  
- App-level visibility + compliance export readiness

# 7. Next Steps

This tool can be deployed, extended, and presented to Legal Tech platforms like Intapp to gather feedback. It also serves as a strong portfolio asset demonstrating end-to-end AI product thinking, architecture design, and awareness of open-source AI ecosystem.