DRC SYSTEMS

Revolutionizing Customer Experience with PrivateGPT

Client Overview

The client is an educational institution aiming to enhance information accessibility and user experience for students, faculty, and website visitors. The institution faced challenges in efficiently delivering topic-specific academic content, resulting in information overload on the website and difficulty in finding relevant resources. To address these issues, the institution sought an AI-powered solution capable of providing instant, accurate, and context-aware responses within a secure environment eliminating the need for manual searches, streamlining academic support, and improving overall student engagement.

What Client Needed

To create a PrivateGPT for closed environments that can:

- Enable rapid, precise information retrieval for students and website visitors without exposing sensitive data to public AI services.
- Streamline user experience by providing instant, context-aware responses tailored to proprietary content (syllabus, documentation, product info).
- Ensure compliance with data privacy and sovereignty regulations by deploying AI within a closed, on-premise environment.

Tech Stack

LLM Framework

Llama 3 (Meta), PyTorch, Hugging Face Transformers for language tasks and fine-tuning on domain-specific data.

Model Optimization

LoRA (Low-Rank Adaptation) for efficient fine-tuning on proprietary datasets with reduced compute requirements.

Containerization & Orchestration

Docker, Kubernetes, Red Hat OpenShift

Document Processing

Apache Tika for parsing and extracting text from PDFs, word docs, and HTML

RAG Orchestration

LangChain, Haystack for document retrieval and synthesis for context-aware answers

Conversational AI

Rasa for dialogue management, intent recognition, context tracking

NLP Libraries

spaCy for multilingual named entity recognition (NER)

LMS Integration

LTI 1.3, Moodle, Canvas for seamless integration with learning management systems

Experiment Tracking

MLflow for auditing and tracking AI model training and responses

Business Challenges

Inefficient Information Search

Students and clients spent excessive time manually searching through books, chapters, or website pages to find specific information, leading to frustration and lost productivity.

Delayed Client Engagement

Website visitors often waited for email or phone responses, risking missed business opportunities due to slow information delivery.

Data Privacy Concerns

Existing public AI solutions posed risks of data leakage and non-compliance with regulations (e.g., GDPR, ISO 27001), especially when handling sensitive or proprietary content.

Lack of Customization

Off-the-shelf chatbots and search tools could not be tailored to the organization's unique data, terminology, or compliance requirements.

What We Built

To address these challenges, we developed PrivateGPT, a tailored conversational AI solution leveraging advanced AI and machine learning technologies.

Deployment of PrivateGPT

- **On-Premise LLM Framework:** Deployed **Llama 3 (Meta)** as the base model, fine-tuned using **PyTorch** and **Hugging Face Transformers** for domain-specific tasks.
- **Containerization:** Hosted in a **Kubernetes cluster** (on-premise) using **Docker** for scalable, isolated microservices.
- **Fine-Tuning Pipeline:** Leveraged **LoRA (Low-Rank Adaptation)** to optimize model training on proprietary datasets (syllabi, product docs) with minimal compute overhead.

Retrieval Augmented Generation (RAG)

- **Text Extraction:** Apache Tika for parsing PDFs, Word docs, and web-scraped HTML.
- **Vectorization:** Used **sentence-transformers (all-MiniLM-L6-v2)** to convert text into embeddings.
- **Vector Database:** Stored embeddings in **Pinecone** for low-latency semantic search.
- **Contextual Answering:** Combined **LangChain orchestration** with **Haystack** pipelines to dynamically retrieve and synthesize answers from documents.

Secure Data Handling

- **Network Isolation:** Deployed within a Zero Trust Architecture using Tailscale for encrypted peer-to-peer connections.
- **Data Encryption:** Applied **AES-256** for data-at-rest and **TLS 1.3** for data-in-transit.
- **Authentication:** Integrated **Keycloak** for SSO and **OAuth 2.0** with Active Directory/LDAP.
- **Authorization:** Enforced **RBAC** using **Open Policy Agent (OPA)** for granular permissions.

Multi-language Support

- **NLP Libraries:** Leveraged **spaCy** for multilingual named entity recognition (NER).
- **Offline Translation:** Deployed **MarianMT** models for on-device translation in restricted environments.
- **Locale-Specific Tuning:** Fine-tuned the LLM on **parallel corpora** (e.g., EUROPARL) for idiomatic accuracy.

Scalable Integration

- **REST & GraphQL:** Exposed endpoints via **FastAPI (Python)** and **Apollo Server (Node.js)**.
- **Asynchronous Workflows:** Managed high-volume requests with **Celery** and **RabbitMQ** task queues.
- **LMS Integration:** Prebuilt connectors for **Moodle and Canvas** using LTI 1.3 standards.
- **CRM Compatibility:** Webhooks to sync with **Salesforce** and **HubSpot** for client engagement tracking.

Business Impact

The PrivateGPT solution empowered the client to deliver a transformative user experience for both students and clients:

Efficiency Gains

Students and website visitors could instantly access precise information, reducing search and response times by over 70%.

Enhanced Security

All data remained within the organization's private infrastructure, ensuring full compliance with data privacy laws and eliminating the risk of external data leaks.

Customization and Control

The AI assistant was fine-tuned on proprietary data, providing contextually relevant answers and supporting industry-specific terminology and workflows.

Scalability

The solution was easily integrated across educational, corporate, and customer-facing platforms, supporting a wide range of use cases from student revision to customer support automation.

Core Achievements

Improved client engagement, reduced operational bottlenecks, and built organizational AI expertise, resulting in measurable ROI and a competitive edge.

PrivateGPT demonstrates how secure, on-premise generative AI can revolutionize information access and user engagement while maintaining strict data governance and compliance.