HackRx 5.0

Ideate • Co-create • Impact

BAJAJ FINSERV

GPTEAM

next slide →

HackRx 5.0

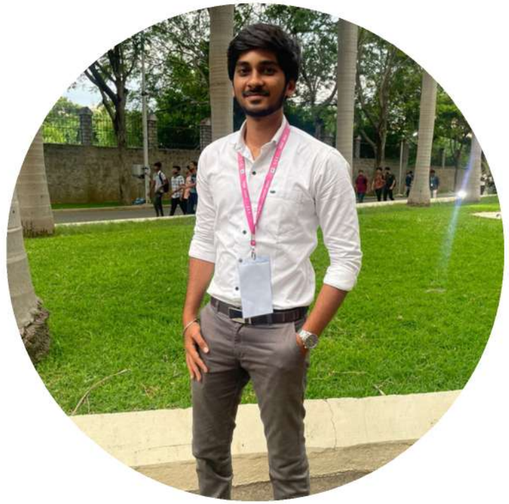
Ideate • Co-create • Impact

BAJAJ FINSERV

The team: Who are we?

Syed Nabiel Hasaan M

- Former Intern @ SHARP Software Development India.
- Former Intern @ Daira Edtech private limited.
- Top 10 @ DeFy 24 Blockchain National Hackathon
- Qualifier @ Microsoft Imagine Cup



Arvindhan K

- Current Intern @ Samsung Prism
- Winner - Python coding challenge
- Author of research paper based on image classification using deep learning
- Former intern @ Daira EdTech private limited

Yashwanth B

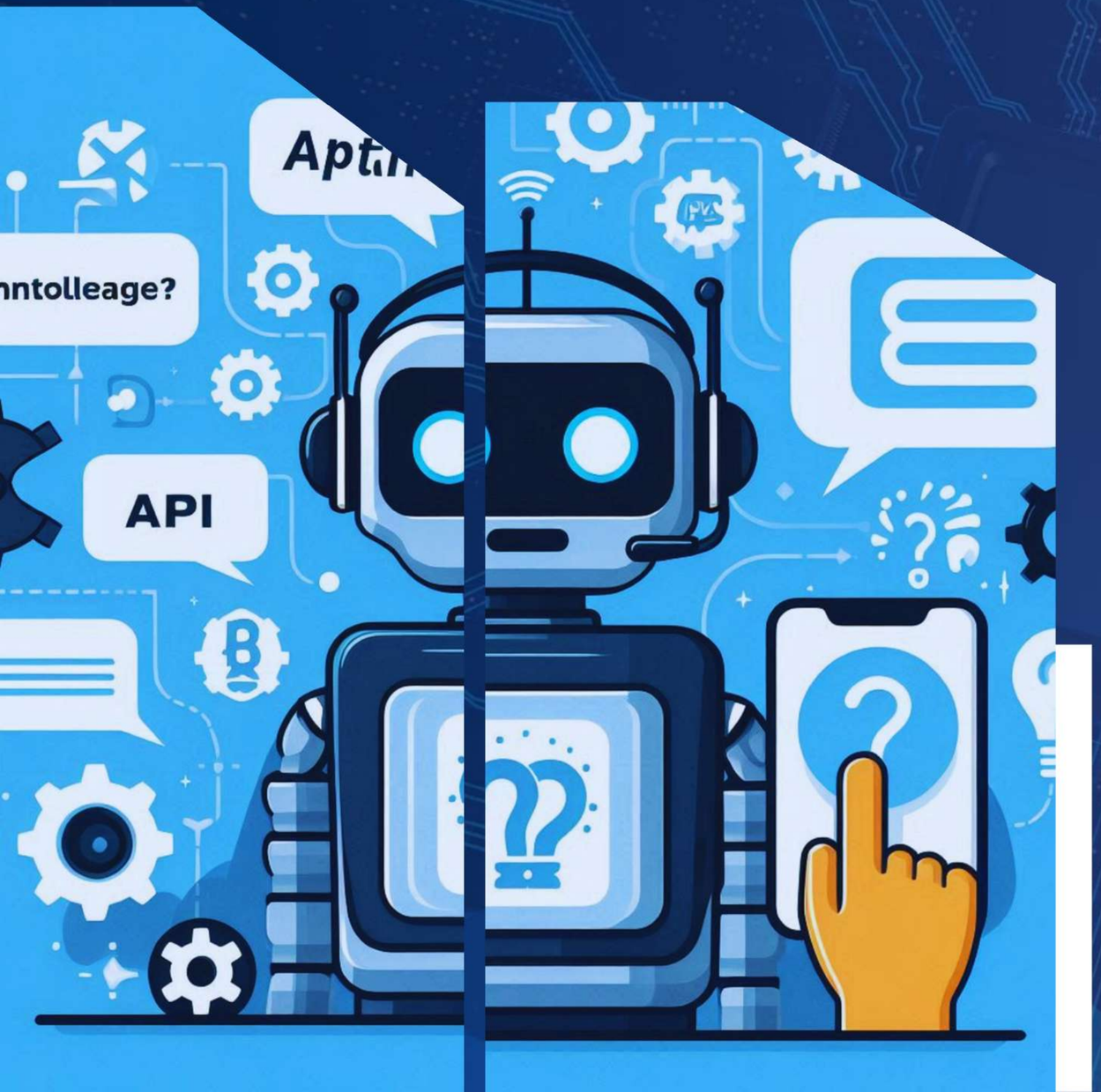
- Current Intern @ Daira Edtech private limited.
- HR @Microsoft Azure Skillup Tamilnadu
- OCI GEN AI Professional
- Freeelancer -UI/UX design

next slide →

HackRx 5.0

Ideate • Co-create • Impact

BAJAJ FINSERV



PROBLEM STATEMENT

To develop a context-aware chatbot that delivers precise information from a vast knowledge base while seamlessly executing tasks via API integration for a smooth user experience.

next slide →

HackRx 5.0
Ideate • Co-create • Impact

BAJAJ FINSERV

WHAT DOES IT DO?

1. Automates information retrieval from diverse documents, providing accurate, up-to-date responses without manual searches.

2. Integrates APIs for efficient task execution, enabling seamless order management and payment processing within conversations.

next slide →

HackRx 5.0
Ideate • Co-create • Impact

BAJAJ FINSERV

OUR PRODUCT

RX-ASSISTANT

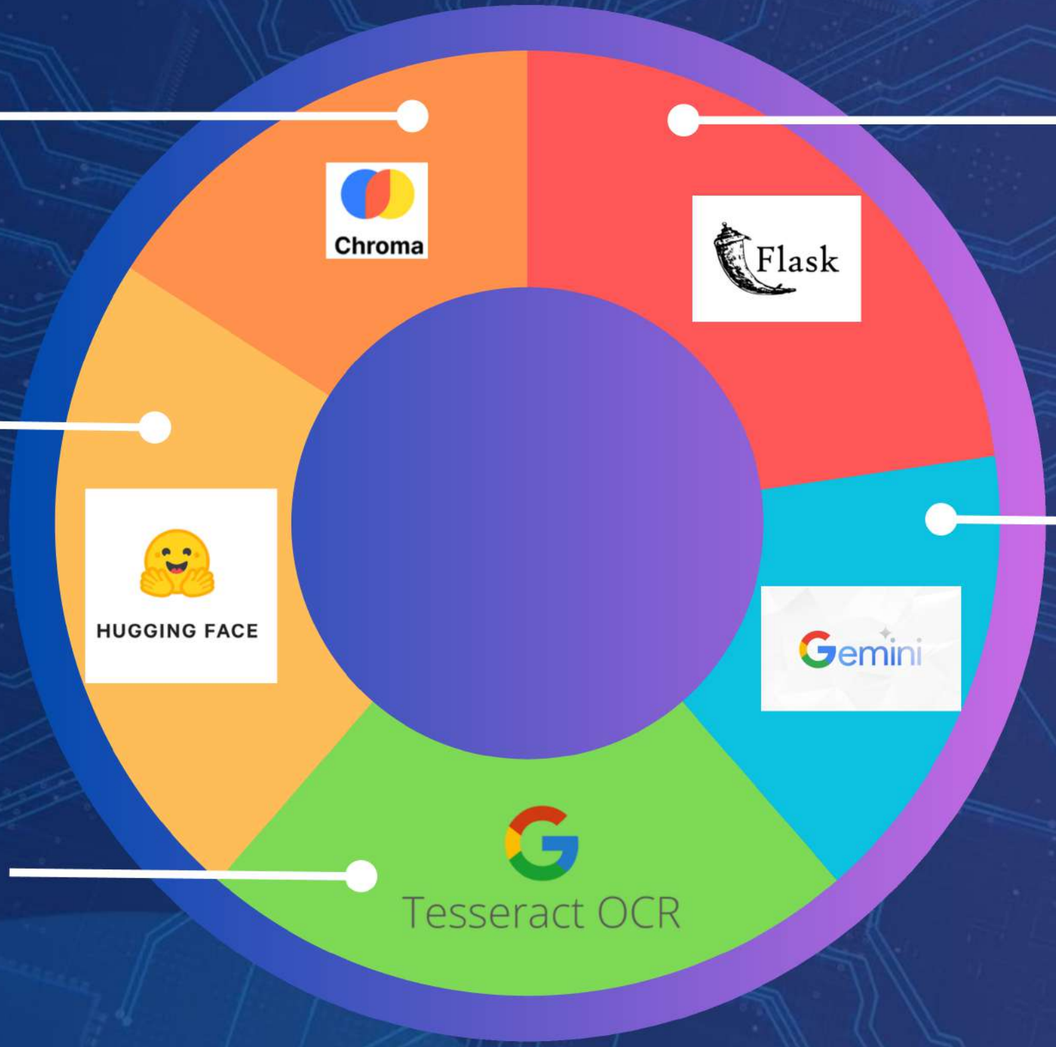
RX-Assistant is a context-aware chatbot that processes user queries related to documents in a designated database/folder, utilizing ChromaDB for storage, Google Generative AI for responses, and Hugging Face Transformers for embeddings.

next slide →

HackRx 5.0
Ideate • Co-create • Impact

BAJAJ FINSERV

TECH STACK



next slide →

HackRx 5.0
Ideate • Co-create • Impact

BAJAJ FINSERV

DETAILED DESCRIPTION OF SOLN

1. **Document Ingestion and Embedding**:
   - Documents (PDFs, PPTs, images) are processed using `load_documents.py`, with text extraction via libraries like `pdfplumber`, `python-pptx`, and `pytesseract`.
   - Extracted text is embedded using the SentenceTransformer model.

2. **Vector Database (ChromaDB)**:
   - Embeddings are stored in ChromaDB for semantic search, with metadata tracking document origin.

3. **User Query Handling**:
   - The Flask server (`app.py`) processes user queries via an API, retrieving relevant embeddings from ChromaDB.

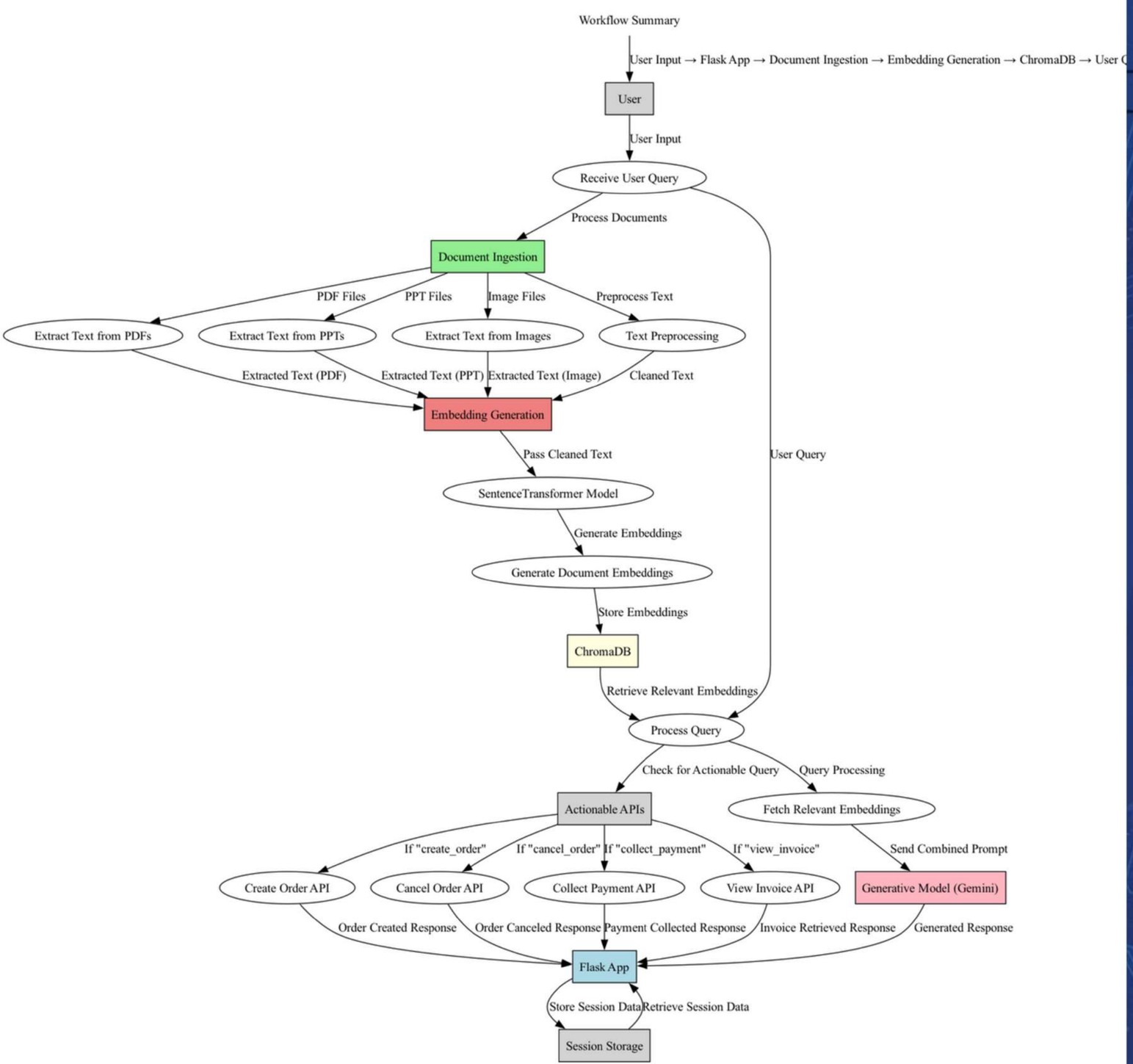
4. **Context Management and Prompt Engineering:**
   - Context is built from embeddings to match user queries, and prompt engineering ensures actionable queries (e.g., "create_order") are recognized.

5. **Accessing the LLM**:
   - The combined prompt (query + context + history) is sent to Gemini, which generates a response.

6. **Session Storage:**
   - User sessions are tracked, mapping query and response history for personalized interactions.

Workflow Summary:
User sends a query → Flask app retrieves embeddings from ChromaDB → Context is built → Prompt is sent to Gemini → Response generated, actionable queries processed, and session data updated.

next slide →



BAJAJ FINSERV

BLOCK DIAGRAM

next slide →

HackRx 5.0
Ideate • Co-create • Impact

BAJAJ FINSERV

WHY EMBEDDINGS?

Aspect	Passing Text to LLM	Using Embeddings
Scalability	Limited by token limits (max ~4K-32K tokens)	Scales well with large datasets and knowledge bases
Token Efficiency	Consumes more tokens, leading to higher costs	Reduces token usage by retrieving relevant chunks
Context Length	Can handle short or moderately sized contexts	Can manage long contexts by retrieving only necessary info
Performance	Slower for large inputs due to token limits	Faster and more efficient for large knowledge bases
Cost	Higher cost due to token consumption	Lower cost by optimizing API calls
Context Preservation	Limited by token capacity	Embeddings can preserve and retrieve context effectively
Complexity	Simpler to implement	Requires setting up vector storage and retrieval system
Use Case	Suitable for short or real-time interactions	Ideal for handling large-scale or multi-format data

next slide →

Vector Database

Aspect	Pinecone (Cloud-based Vector DB)	ChromaDB (Self-hosted, Open-source Vector DB)
Scalability	**Highly scalable**, auto-scaling resources	**Scalable** based on local/cloud infrastructure you provision
Setup & Maintenance	**Managed service**, no setup required	**Self-hosted**, requires setup, management, and scaling
Cost	**Pay-as-you-go** (pricing below)	**Free** (open-source), but incurs **infrastructure** costs
Cost Example (Pinecone)	Starter Plan: $0.07/hour per pod; standard pod: $0.096/hour	**Infrastructure costs** only (e.g., cloud VMs, storage, bandwidth)
Total Monthly Cost	~$50/month (small scale), scales with more pods/ usage	Varies (~$20-$100+/month) based on your **cloud setup** (e.g., AWS)
Latency	**Low latency** due to global infrastructure	Depends on infrastructure, generally **higher latency**
Performance	Optimized with **global replication**	Dependent on hardware and cluster size
Ease of Use	**Simple API**, fully managed, minimal DevOps	Requires managing DevOps, manual scaling
Storage	**Cloud storage** with various region options	Custom storage solutions (local or cloud)
Security	**Enterprise-level** security, SOC 2, GDPR, encryption	Security is **your responsibility**, can be customized
Integrations	Integrated with major platforms (AWS, GCP, Azure, Hugging Face)	Requires custom integrations, manual work
Community Support	Strong support, enterprise-ready, large community	Small but growing open-source community

HackRx 5.0
Ideate • Co-create • Impact

BAJAJ FINSERV

LLM

Model/Provider	Cost (per 1M tokens/embeddings)	Typical Accuracy Range (e.g., retrieval tasks)	Best Use Case in Business
Gemini (Google)	$0.15 - $0.20	85-95%	**Enterprise Applications**: Suitable for large-scale operations where high accuracy is critical and cost is less of a constraint.
OpenAI (GPT-4)	$0.03 - $0.06	90-98%	**Mid-to-Large Businesses**: Ideal for businesses that require high accuracy but need to balance cost efficiency, such as customer service bots.
Hugging Face Models	Free to $0.05 (depending on the model and usage)	75-95%	**Startups and Cost-Sensitive Businesses**: Good for smaller operations or experimental projects where budget constraints are tight.

next slide →

HackRx 5.0
Ideate • Co-create • Impact

BAJAJ FINSERV

FUTURE POSSIBLE ENHANCEMANTS

Pinecone Integration: Planning to use Pinecone for hosting the vector database, allowing remote access for fetching embeddings via API to enhance scalability

Expanding Multi Channel Support: to allow user interactions across web, mobile apps, social media, and messaging services, enhancing accessibility and providing a seamless experience

Multi language : to allow uer interactions based on their language preference for providing more native experience.

Hybrid Architecture : Implement dynamic model selection to optimize chatbot performance based on real-time user interactions and load.

Security for the files: Ensuring file security features such as SHA-256, Encryption.

HackRx 5.0
Ideate • Co-create • Impact

BAJAJ FINSERV

RISKS/CHALLENGES/DEPENDENCIES

1. Ensure security compliance to protect sensitive user data
2. API failures can disrupt the user experience.
3. Poor embeddings reduce response accuracy and relevance
4. Large models and vector databases can incur high production costs

next slide →

HackRx 5.0
Ideate • Co-create • Impact

BAJAJ FINSERV

ACCEPTANCE CRITERIA COVERAGE

1 Knowledge Base Integration

- Multiple Document Formats:
  - Integrated with PDFs, PPTs, images, and text documents.
  - Utilizes specialized libraries for text extraction (e.g., pdfplumber, python-pptx, pytesseract).
- Accurate Retrieval:
  - Documents converted to embeddings using SentenceTransformer.
  - Stored in ChromaDB for efficient retrieval based on user queries.
- Knowledge Base Updates:
  - Seamless ingestion and updates of documents via documents.py.

2 Action Execution

- Actionable Queries:
  - Identifies specific user intents (e.g., "create_order", "cancel_order") through prompt engineering and context-aware analysis.
- API Integration:
  - Architecture is ready for API calls.
  - Currently uses a dummy function to simulate action execution; will be replaced by actual API calls in the future.

3 Context Preservation

- **Session Management:**
  - Tracks user sessions to maintain context and history throughout interactions.

next slide →

HackRx 5.0
Ideate • Co-create • Impact

BAJAJ FINSERV

ANYTHING ELSE?

- **Comprehensive Documentation:** Detailed guides and instructions included.
- **User Interface:** Built using Chainlit for seamless interaction.
- **Model Support:** Compatible with Gemini, OpenAI, and Lamini (offline) models.
- **Access:** Available via the GitHub repository: **github/msnabiel/opengpt**.

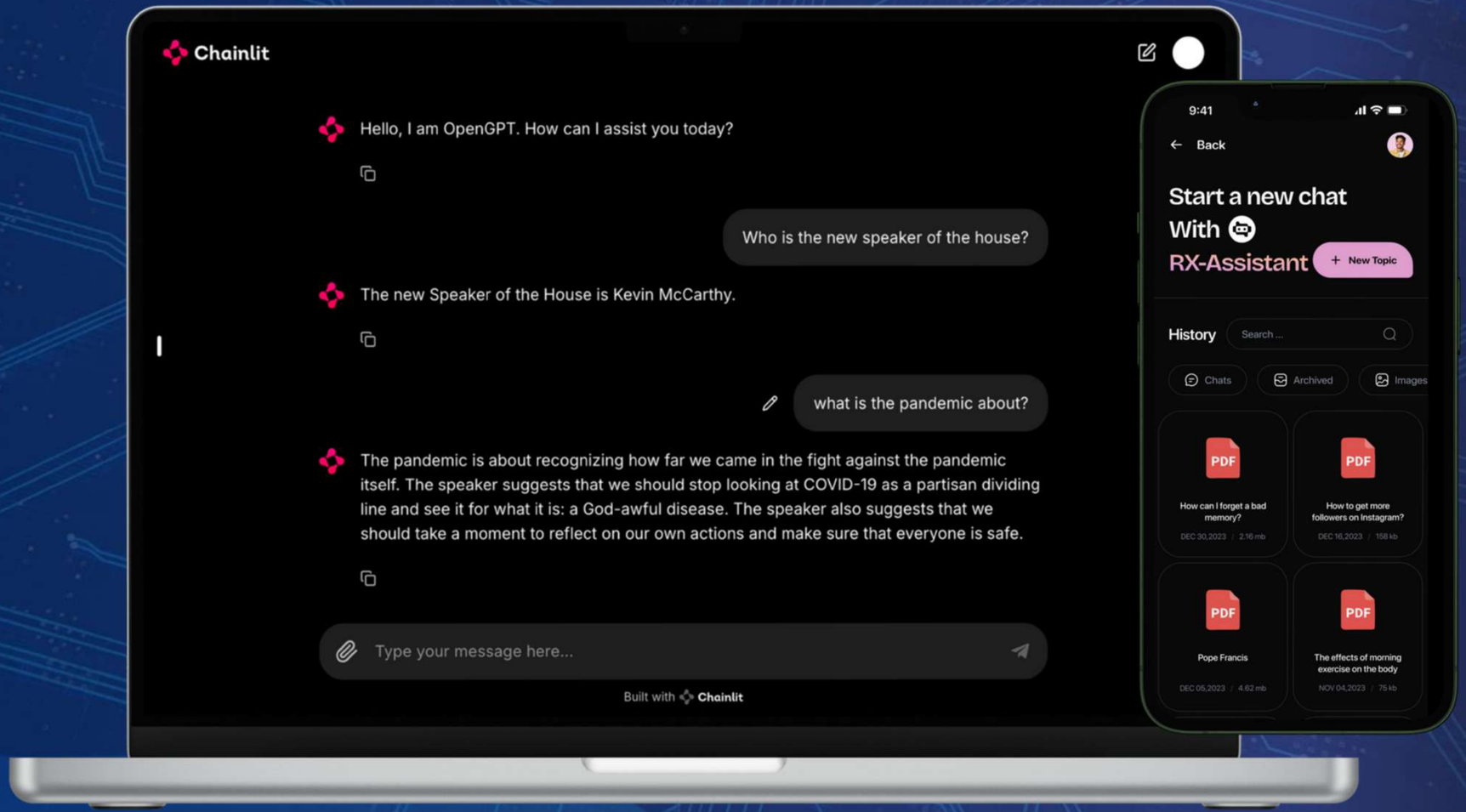
- **Model Base**: Built on the Gemini model.
- **OpenAI Compatibility**: Easy implementation of OpenAI-based alternatives.
- **Flexible Switching**: Supports seamless transitions between local and cloud-based models.
- **Versatile Integration**: Allows use of various embeddings and vector databases.
- **Operational Flexibility**: Adaptable to specific requirements and operational needs.

next slide →

HackRx 5.0
Ideate • Co-create • Impact

BAJAJ FINSERV

USER INTERFACE



next slide →

HackRx 5.0

Ideate • Co-create • Impact

BAJAJ FINSERV

REFERENCE LINK

https://github.com/msnabiel/RX-Asisstant