Building a Private GPT-like ChatBot - An Operational Guide

Baridhi Malakar, PhD[*]

June 19, 2025

Abstract

This technical guide introduces building a private GPT-like chat-bot and provides a structured implementation. It is intended to support practitioners and researchers working on artificial intelligence applications in finance. Relevant tools are discussed with practical code examples and potential future extensions.

1 Introduction

Natural Language Processing (NLP) has become a critical tool in financial research, enabling the extraction of insights from unstructured text such as earnings calls, filings, and news. Early applications focused on sentiment analysis, tone, and topic modeling to quantify qualitative disclosures and predict asset prices or firm fundamentals (e.g., Loughran and McDonald (2011); Tetlock (2007)). Recent scholarly work such as Chava et al. (2021), Gurun et al. (2021), and Jiang et al. (2023) illustrate the growing sophistication of NLP techniques in extracting economic signals and enabling real-time financial decision-making. Subsequent applications have led to the development of more dynamic tools like large language model (LLM)-based chatbots.

This guide details the implementation of a private chatbot for financial document analysis using Retrieval-Augmented Generation (RAG) and locally hosted LLMs. Particularly, the framework is meant to support non-computer science majors who work at the intersection

[*] Affiliation: Independent Researcher. Email: baridhi.malakar@scheller.gatech.edu

of finance, machine learning, and artificial intelligence from an applied perspective. The code repository for this guide is included on Github. This example focuses on PDF documents which form a large portion of unstructured data in financial documents. The system addresses two critical needs in investment research:

- **Data Privacy**: No API calls or cloud dependencies (Menick et al., 2022).
- **Domain-Specific Accuracy**: RAG enhances factual consistency for financial texts (Lewis et al., 2020).

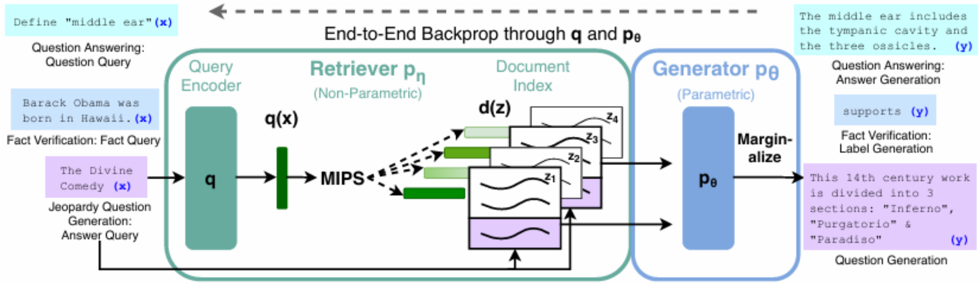

Define "middle ear" (x)
Question Answering: Question Query
Barack Obama was born in Hawaii. (x)
Fact Verification: Fact Query
The Divine Comedy (x)
Jeopardy Question Generation: Answer Query
End-to-End Backprop through q and pθ
Query Encoder
Retriever pη
(Non-Parametric)
Document Index
q
q(x)
MIPS
d(z)
z4
z3
z2
z1
Generator pθ
(Parametric)
pθ
Margin-alize
The middle ear includes the tympanic cavity and the three ossicles. (y)
Question Answering: Answer Generation
supports (y)
Fact Verification: Label Generation
This 14th century work is divided into 3 sections: "Inferno", "Purgatorio" & "Paradiso" (y)
Question Generation

Figure 1: Overview of system architecture, Lewis et al. (2020).

2 Technical Foundations

This project was implemented on MacBook Air with an M3 chip, eight cores, and 16GB memory. The following packages were used in this project:

- `Python`: version 3.12
- `tqdm`: version 4.67.1
- `ollama`: version 0.4.2
- `langchain_community`: version 0.3.8
- `langchain_huggingface`: version 0.1.2
- `langchain-chroma`: version 0.1.4
- `langchain-ollama`: version 0.2.0
- `chromadb`: version 0.5.20

- `sentence_transformers`: version 3.3.1
- `pymupdf`: version 1.24.14
- `streamlit`: version 1.40.2

2.1 Retrieval-Augmented Generation (RAG)

RAG combines dense retrieval with generative LLMs to reduce hallucinations in financial analysis (Lewis et al., 2020). Key components deployed in this architecture include:

Table 1: RAG Components in this implementation

Component	Role
ChromaDB	Vector store for document chunks
Mistral-7B (default)	Generative LLM for answer synthesis
PyMuPDF	Text extraction from PDFs
Ollama	Local suite to host open-source LLMs

2.2 Chain-of-Thought (CoT) Prompting

CoT improves reasoning by decomposing questions into intermediate steps (Wei et al., 2022). The `pvtgpt_cfa_ui.py` implements this via:

```
CUSTOM_PROMPT = """
As an investment research analyst:
1. Identify key terms in: {question}
2. Retrieve relevant covenants from: {context}
3. Summarize risks with page references.
"""
```

3 Phase 1: Document Ingestion

3.1 Text Extraction

The `ingest.py` script uses `PyMuPDFLoader` for PDF parsing, `TextLoader` to read in TXT files, etc. in order to choose for its:

- Preservation of financial table structures
- Metadata extraction (e.g., page numbers)

3.2 Chunking Strategy

Optimal chunk size balances:

- **Context Preservation**: 1,024 tokens capture full covenants
- **Retrieval Efficiency**: Smaller chunks reduce noise (Karpukhin et al., 2020)

```
# ingest.py (optimized for financial docs)
text_splitter = RecursiveCharacterTextSplitter(
    chunk_size=1024,
    chunk_overlap=128,  # Ensures covenant continuity
    length_function=len
)
```

3.3 Embedding Generation

`HuggingFaceEmbeddings` creates vectors using the following items:

- `all-MiniLM-L6-v2`: Balance of speed/accuracy for finance terms
- Metadata injection (source, page numbers) for auditability

This dense vector representation (embedding) aims to capture semantic meaning. These vectors enable efficient similarity search in Retrieval-Augmented Generation (RAG) pipelines.

4 Phase 2: Query Processing

4.1 Retrieval Optimization

The system employs:

- **Hybrid Search**: Combines semantic + keyword search
- **Re-Ranking**: Prioritizes chunks with financial terminology

```
# pvtgpt_cfa_ui.py retrieval configuration
retriever = db.as_retriever(
    search_type="mmr",  # Max Marginal Relevance
    search_kwargs={"k": 5, "filter": {"document_type": ".pdf"}}
)
```

4.2 Generation with Local LLMs

Key considerations when using Mistral:

- **Temperature**: 0.3 for factual responses
- **Stop Sequences**: "\n\n" prevents rambling

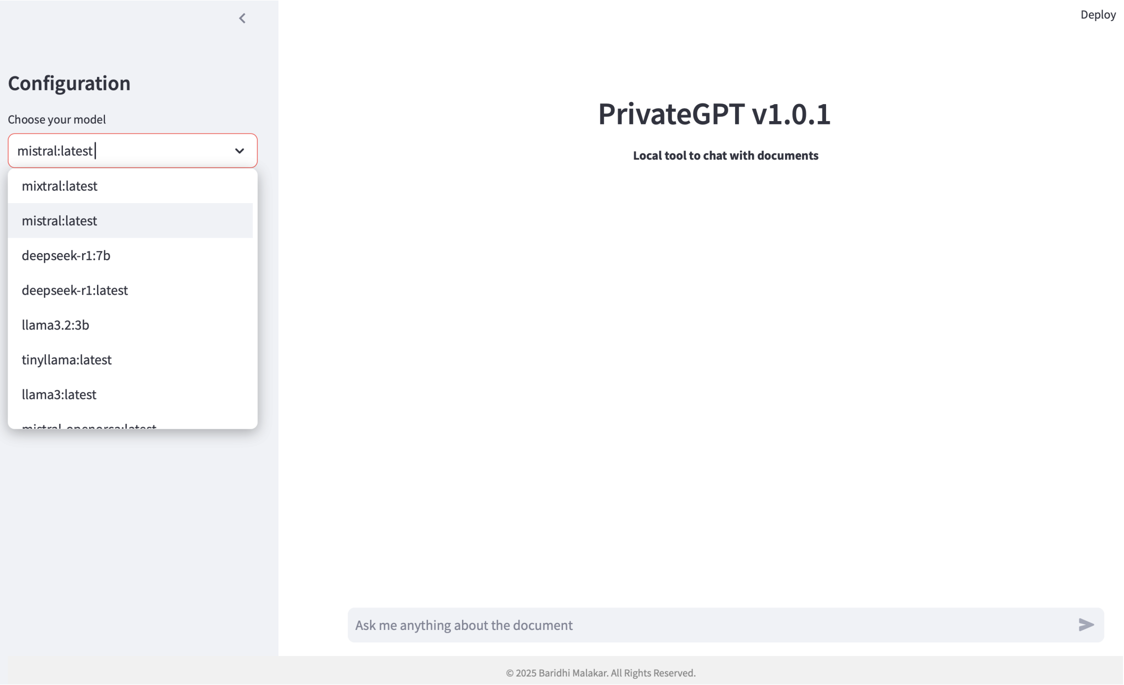
```
# Ollama configuration in app.py
llm = OllamaLLM(
    model="mistral",
    temperature=0.3,
    stop=["\n\n"],
    top_k=40  # Broader search for rare terms
)
```

5 Phase 3: User Interaction

5.1 Streamlit User-Interface Design

The interface incorporates:

- Model switching (Mistral ↔ DeepSeek)
- Source citation expander (Figure 2b)
- Conversation persistence



Configuration

Choose your model

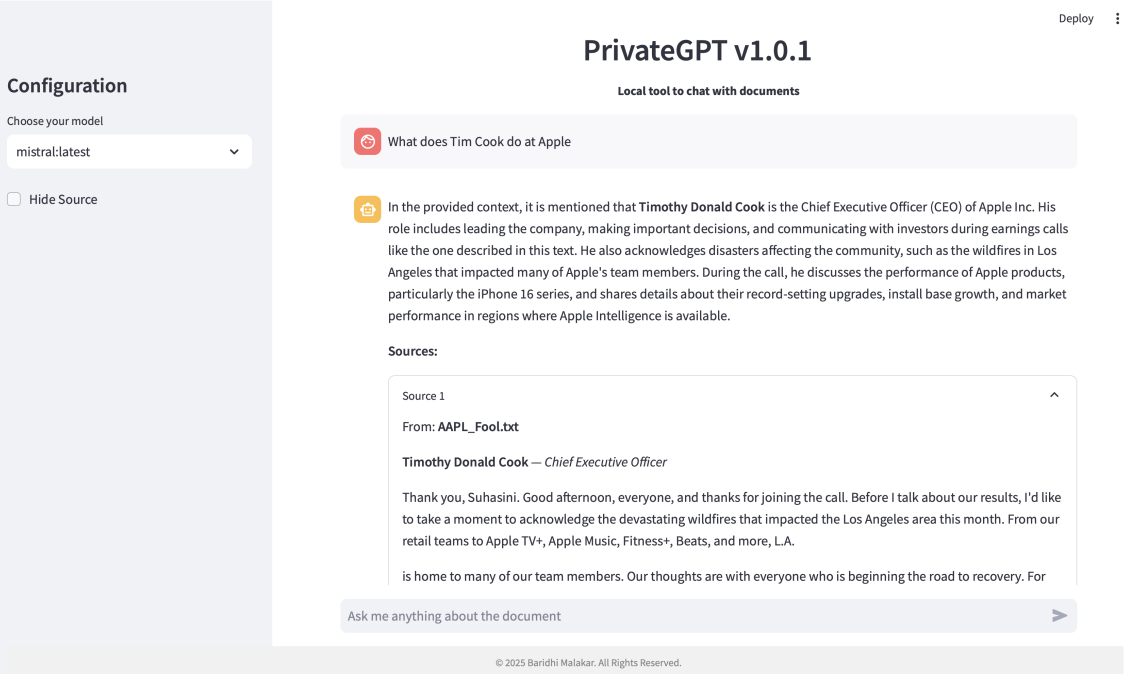
mistral:latest

PrivateGPT v1.0.1

Local tool to chat with documents

(a) Model selection dropdown

Figure 2: User-Interface components (continued on next page)


Deploy
□ Hide Source
Ask me anything about the document
© 2025 Baridhi Malakar. All Rights Reserved.

Configuration

Choose your model

mistral:latest

PrivateGPT v1.0.1

Local tool to chat with documents

What does Tim Cook do at Apple

In the provided context, it is mentioned that **Timothy Donald Cook** is the Chief Executive Officer (CEO) of Apple Inc. His role includes leading the company, making important decisions, and communicating with investors during earnings calls like the one described in this text. He also acknowledges disasters affecting the community, such as the wildfires in Los Angeles that impacted many of Apple's team members. During the call, he discusses the performance of Apple products, particularly the iPhone 16 series, and shares details about their record-setting upgrades, install base growth, and market performance in regions where Apple Intelligence is available.

Sources:

Source 1

From: **AAPL_Fool.txt**

Timothy Donald Cook — *Chief Executive Officer*

Thank you, Suhasini. Good afternoon, everyone, and thanks for joining the call. Before I talk about our results, I'd like to take a moment to acknowledge the devastating wildfires that impacted the Los Angeles area this month. From our retail teams to Apple TV+, Apple Music, Fitness+, Beats, and more, L.A.

is home to many of our team members. Our thoughts are with everyone who is beginning the road to recovery. For

(b) Source citation expanders

Figure 2: User-Interface components (continued from previous page)

5.2 Chain-of-Thought in Practice

Example query processing:

1. User asks: “What does Tim Cook do at Apple”
2. System retrieves relevant number of chunks
3. Mistral synthesizes response with page references
4. UI displays sources in expandable sections

See additional examples in the Appendix shownig Figure 3.

5.3 Quantization Tradeoffs

4-bit quantization reduces memory by 4x in bash with minimal accuracy loss (Dettmers et al., 2023):

```
$ ollama pull mistral:7b-instruct-q4_0
```

6 Advanced Use Cases

6.1 Multi-Document Analysis

Extend `ingest.py` to:

- Cluster documents by industry
- Compare management discussion across stocks

6.2 Fine-Tuning with LoRA

Low-Rank Adaptation (LoRA) is a parameter-efficient fine-tuning method for large language models. The technique involves freezing the original weights of the LLM and injecting trainable low-rank matrices to adapt the model to new tasks with minimal computational overhead. Domain adaptation steps may include:

1. Prepare industry(or stock)-specific Q&A pairs
2. Train adapter layers (Hu et al., 2021)
3. Merge with base Mistral

7 Conclusion

This guide demonstrates how open-source tools can create secure, reasonably accurate AI assistants for investment research. This private GPT does not rely on (paid) subscription or transferring data to an external website. This makes it cost-effective and secure. Future directions of possible enhancements include:

- Fine-tuning on proprietary research documents
- Integrating learning from internal memos and meeting minutes

References

Chava, S., W. Du, and B. Malakar (2021). Do managers walk the talk on environmental and social issues? *Georgia Tech Scheller College of Business Research Paper* (3900814).

Dettmers, T., A. Pagnoni, A. Holtzman, and L. Zettlemoyer (2023). Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems 36*, 10088–10115.

Gurun, U. G., N. Stoffman, and S. E. Yonker (2021). Unlocking clients: The importance of relationships in the financial advisory industry. *Journal of Financial Economics 141*(3), 1218–1243.

Hu, E. et al. (2021). Lora: Low-rank adaptation of large language models. *ICLR*.

Jiang, J., B. Kelly, and D. Xiu (2023). (re-) imag (in) ing price trends. *The Journal of Finance 78*(6), 3193–3249.

Karpukhin, V., B. Oguz, S. Min, P. S. Lewis, L. Wu, S. Edunov, D. Chen, and W.-t. Yih (2020). Dense passage retrieval for open-domain question answering. In *EMNLP (1)*, pp. 6769–6781.

Lewis, P., E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, et al. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems 33*, 9459–9474.

Loughran, T. and B. McDonald (2011). When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *The Journal of finance 66*(1), 35–65.

Menick, J., M. Trebacz, V. Mikulik, J. Aslanides, F. Song, M. Chadwick, M. Glaese, S. Young, L. Campbell-Gillingham, G. Irving, et al. (2022). Teaching language models to support answers with verified quotes. *arXiv preprint arXiv:2203.11147*.

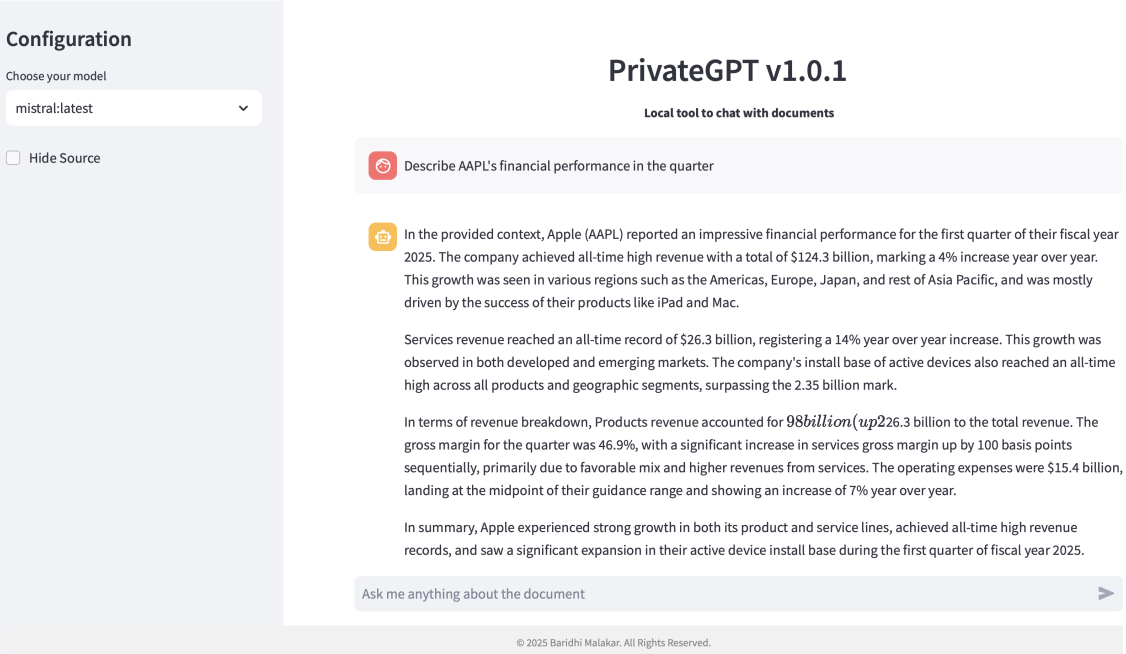
Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. *The Journal of finance 62*(3), 1139–1168.

Wei, J. et al. (2022). Chain-of-thought prompting elicits reasoning in large language models. *NeurIPS*.

Appendix

Additional Examples

The following instances provide two more examples of how the interactive privateGPT can handle questions posed by the user. The user may choose to toggle the choice of hiding source references from which the answers are composed by the application.



Configuration

Choose your model

mistral:latest

□ Hide Source

PrivateGPT v1.0.1

Local tool to chat with documents

Describe AAPL's financial performance in the quarter

In the provided context, Apple (AAPL) reported an impressive financial performance for the first quarter of their fiscal year 2025. The company achieved all-time high revenue with a total of $124.3 billion, marking a 4% increase year over year. This growth was seen in various regions such as the Americas, Europe, Japan, and rest of Asia Pacific, and was mostly driven by the success of their products like iPad and Mac.

Services revenue reached an all-time record of $26.3 billion, registering a 14% year over year increase. This growth was observed in both developed and emerging markets. The company's install base of active devices also reached an all-time high across all products and geographic segments, surpassing the 2.35 billion mark.

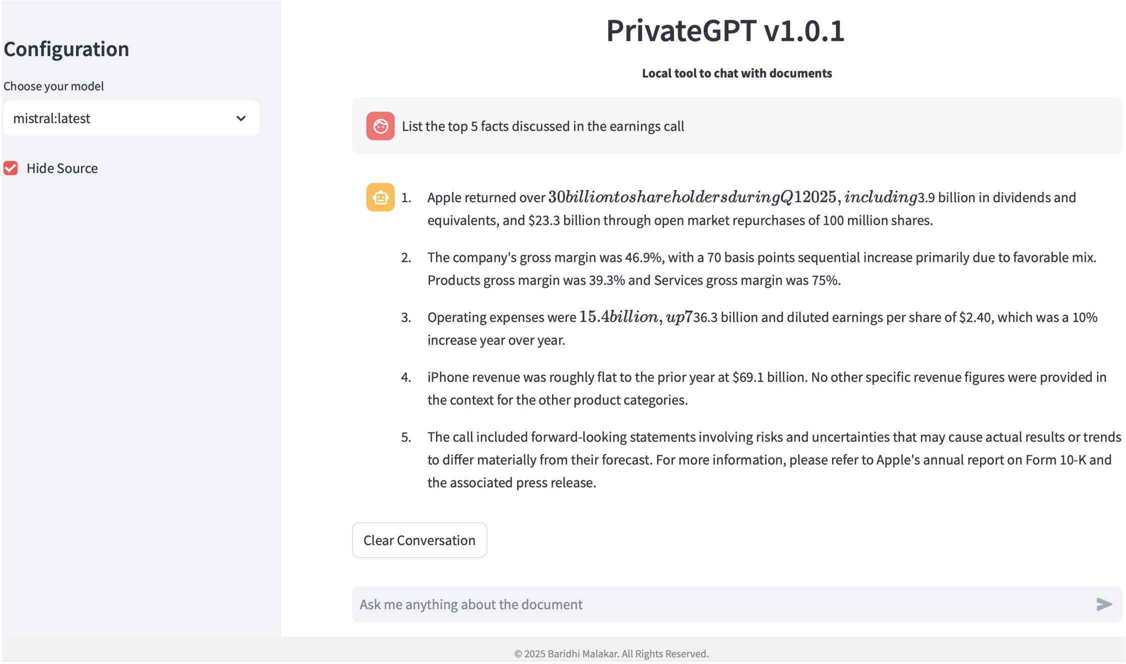
In terms of revenue breakdown, Products revenue accounted for $98billion(up2$26.3 billion to the total revenue. The gross margin for the quarter was 46.9%, with a significant increase in services gross margin up by 100 basis points sequentially, primarily due to favorable mix and higher revenues from services. The operating expenses were $15.4 billion, landing at the midpoint of their guidance range and showing an increase of 7% year over year.

In summary, Apple experienced strong growth in both its product and service lines, achieved all-time high revenue records, and saw a significant expansion in their active device install base during the first quarter of fiscal year 2025.

© 2025 Baridhi Malakar. All Rights Reserved.

(a) Default usage

Figure 3: Example Demonstration



Configuration

Choose your model

mistral:latest

☑ Hide Source

PrivateGPT v1.0.1

Local tool to chat with documents

List the top 5 facts discussed in the earnings call

1. Apple returned over $30billiontoshareholdersduringQ12025, including$3.9 billion in dividends and equivalents, and $23.3 billion through open market repurchases of 100 million shares.
2. The company's gross margin was 46.9%, with a 70 basis points sequential increase primarily due to favorable mix. Products gross margin was 39.3% and Services gross margin was 75%.
3. Operating expenses were $15.4billion, up7$36.3 billion and diluted earnings per share of $2.40, which was a 10% increase year over year.
4. iPhone revenue was roughly flat to the prior year at $69.1 billion. No other specific revenue figures were provided in the context for the other product categories.
5. The call included forward-looking statements involving risks and uncertainties that may cause actual results or trends to differ materially from their forecast. For more information, please refer to Apple's annual report on Form 10-K and the associated press release.

(b) Source toggled off

Figure 3: Example Demonstration