🚀 Quick Start Guide

Get your Mini RAG Chatbot running in 5 minutes!

Prerequisites Checklist

- ☐ Python 3.9 or higher installed
- ☐ Ollama installed (ollama.ai)
- ☐ 5-10 research papers (PDFs) ready

Step-by-Step Setup

1. Install Ollama & Llama 3.2

bash

```
# macOS/Linux: Download from https://ollama.ai
# Then pull the model
ollama pull llama3.2
```

Verify it works:

bash

```
ollama list
# Should show llama3.2 in the list
```

2. Clone/Download Project

bash

```
# If you have the project
cd mini-rag-chatbot

# Make setup script executable (Linux/macOS)
chmod +x setup.sh

# Run setup
./setup.sh
```

Or install manually:

bash

```
python3 -m venv venv
source venv/bin/activate  # On Windows: venv\Scripts\activate
pip install -r requirements.txt
```

3. Add Your Documents

```
bash

# Copy your research papers
cp /path/to/your/papers/*.pdf data/

# Verify they're there
ls -la data/
```

Sample datasets (if you don't have papers):

- ArXiv papers: arxiv.org
- PubMed articles: pubmed.ncbi.nlm.nih.gov
- Any research PDFs from your field

4. Ingest Documents

```
bash

python src/ingest.py
```

Expected output:

```
Loading documents from data...
Loaded 50 document pages
Chunking documents...
Created 142 chunks
Creating embeddings and vector store...
✓ Vector store saved with 142 chunks
```

This takes 2-5 minutes depending on document size.

5. Run the Chatbot!

Option A: Web Interface (Recommended)

```
bash
```

```
streamlit run src/app.py
```

Opens at `http://localhost:8501`

Option B: Command Line

```bash
python src/chatbot.py
```

Option C: Jupyter Notebook

```bash
jupyter notebook notebooks/demo.ipynb
```

First Questions to Try

Once running, try these questions:

1. **"What are the main contributions of these papers?"**
2. **"Summarize the methodology used"**
3. **"What datasets were used in the experiments?"**
4. **"What are the key findings?"**
5. **"What limitations are mentioned?"**

Common Issues & Fixes

Issue 1: "Vector store not found"

```bash
# Solution: Run ingestion first
python src/ingest.py
```

Issue 2: "Could not connect to Ollama"

```bash
# Solution: Make sure Ollama is running
ollama serve  # In a separate terminal
```

Issue 3: "Model llama3.2 not found"

```bash
# Solution: Pull the model
ollama pull llama3.2
```

Issue 4: "No PDF files found"

```bash
# Solution: Add PDFs to data/ directory
ls data/  # Should show .pdf files
```

Issue 5: Out of memory

```bash
# Solution: Reduce chunk size or use fewer documents
python src/ingest.py --chunk-size 500
```

Testing Your Setup

Run the test suite:

```bash
python test_pipeline.py
```

This will verify:

- ✓ Document ingestion works
- ✓ Retrieval returns results
- ✓ Chatbot generates answers
- ✓ Failure cases are handled

Configuration Options

Adjust Chunk Size

```bash
```

```
python src/ingest.py --chunk-size 1500 --chunk-overlap 300
```

Change Number of Retrieved Chunks

```bash
python src/chatbot.py --top-k 6
```

Use Different Model

```bash
# First, pull the model
ollama pull llama3.1

# Then use it
python src/chatbot.py --model llama3.1
```

Adjust Temperature

```bash
# Lower = more focused, higher = more creative
python src/chatbot.py --temperature 0.3
```

Project Structure

```
mini-rag-chatbot/
├── data/              ← Put your PDFs here
├── src/
│   ├── ingest.py      ← Run this first
│   ├── chatbot.py     ← Main application
│   └── app.py         ← Web interface
├── vectorstore/       ← Generated by ingest.py
└── notebooks/         ← Interactive demos
```

Next Steps

1. **Try the Jupyter notebook** for detailed explanations:

```bash
```

```
jupyter notebook notebooks/demo.ipynb
```

2. **Read the full README** for advanced features
3. **Customize the prompts** in `src/chatbot.py`
4. **Add more documents** and re-run ingestion
5. **Share your results** and get feedback!

Performance Expectations

With 100 research papers (~500 pages):

- **Ingestion**: 3-5 minutes
- **Query response**: 3-5 seconds
- **Retrieval only**: <1 second
- **Accuracy**: ~85% (based on manual evaluation)

Getting Help

- Check `README.md` for detailed documentation
- Review `notebooks/demo.ipynb` for examples
- Run `test_pipeline.py` to diagnose issues
- Check Ollama docs: ollama.ai/docs

Minimal Working Example

```
python

from src.chatbot import RAGChatbot

# Initialize
chatbot = RAGChatbot(vectorstore_dir="vectorstore")

# Ask question
result = chatbot.answer("What is this paper about?")

# Print answer
print(result['answer'])
```

Tips for Best Results

1. **Use high-quality PDFs** (not scanned images)
2. **Start with 10-20 papers** to test
3. **Keep questions specific** to your documents
4. **Review sources** to verify answers
5. **Adjust top_k** if answers are too narrow/broad

Recording Your Demo

For the deliverable, record:

1. **Terminal output** of ingestion

```bash
script -c "python src/ingest.py" ingestion.log
```

2. **Screen recording** of the app:
   - Use QuickTime (Mac), OBS (all platforms), or Screen Recorder
   - Show: asking questions, viewing sources, demonstrating features
3. **Screenshots** of interesting results

Success Checklist

□ Ollama installed and llama3.2 pulled
□ Documents ingested successfully
□ Can ask questions and get answers
□ Sources are displayed correctly
□ Test suite passes
□ Ready to demo!

🎉 **You're all set! Start asking questions about your research papers.**

For more details, see the main `README.md`