Frequently Asked Questions - Agentic RAG System

Frequently Asked Questions

GENERAL QUESTIONS:

Q1: What is an Agentic RAG system?
A: Agentic RAG (Retrieval-Augmented Generation) is an AI system that intelligently decides when to search documents and when to answer directly. It combines document retrieval with language generation for accurate answers.

Q2: What types of documents can I use?
A: Currently, the system supports PDF documents. We recommend using text-based PDFs rather than scanned images for best results.

Q3: How many documents can I process?
A: There's no hard limit, but performance is optimal with up to 1,000 documents. For larger collections, consider organizing into separate indexes.

Q4: Is my data secure?
A: Yes! The system runs entirely locally on your machine. No data is sent to external servers. All processing happens on your computer.

INSTALLATION QUESTIONS:

Q5: What are the system requirements?
A: You need Python 3.8+, 8GB RAM (16GB recommended), and 10GB free disk space. The first run will download AI models (approximately 2GB).

Q6: How long does installation take?
A: Initial setup takes 10-15 minutes, including downloading dependencies and AI models. Subsequent runs are much faster.

Q7: I'm getting installation errors. What should I do?
A: Ensure you're using a virtual environment and have the latest pip version. Try: pip install --upgrade pip, then reinstall requirements.

USAGE QUESTIONS:

Q8: How do I add new documents?
A: Place PDF files in the data folder, then run the build command to reindex.

The system will process all PDFs in the folder.

Q9: How accurate are the answers?
A: Accuracy depends on document quality and question clarity. The system provides confidence scores to help you assess answer reliability.

Q10: Can I use this for multiple projects?
A: Yes! Use different persist-dir paths for different document collections. Each project can have its own index.

Q11: What's the difference between search and direct mode?
A: Search mode retrieves relevant document chunks before answering. Direct mode answers without searching. The agent automatically chooses based on your question.

PERFORMANCE QUESTIONS:

Q12: Why is the first query slow?
A: The AI model loads into memory on first use. Subsequent queries are faster. Consider keeping the system running for multiple queries.

Q13: How can I improve answer quality?
A: Use clear, specific questions. Ensure your documents are well-formatted. Increase the number of retrieved chunks (k parameter) for complex questions.

Q14: Can I use GPU acceleration?
A: Yes! If you have a CUDA-compatible GPU, the system will automatically use it for faster processing. Ensure you have the GPU version of PyTorch installed.

TROUBLESHOOTING:

Q15: The system says no documents found. Why?
A: Check that PDF files are in the correct folder and the path is correct. Ensure PDFs are not corrupted or password-protected.

Q16: I'm getting memory errors. What should I do?
A: Reduce the number of documents or use a smaller AI model. Close other applications to free up RAM. Consider processing documents in batches.

Q17: Answers seem irrelevant. How to fix?
A: Rebuild the index with different chunk sizes. Try rephrasing your question.

Check that the relevant information exists in your documents.

ADVANCED QUESTIONS:

Q18: Can I customize the AI model?
A: Yes! Edit the model name in rag_agent.py. You can use any Hugging Face text2text-generation model compatible with the transformers library.

Q19: How do I backup my index?
A: Simply copy the entire persist-dir folder (default: chroma_db). Restore by copying it back.

Q20: Can I integrate this into my application?
A: Yes! The code is modular. Import the functions you need and integrate them into your Python application.

SUPPORT:

For additional help:
- Email: support@agenticrag.example.com
- Documentation: https://docs.agenticrag.example.com
- GitHub Issues: https://github.com/agenticrag/issues

Last Updated: January 4, 2024
Version: 1.0