# Frequently Asked Questions

### *Q: What file formats are supported?*

LotoAI supports PDF, DOCX, DOC, TXT, MD, HTML, and HTM files. The maximum file size is 50MB per file.

### *Q: How does the search work?*

The system uses hybrid search combining vector embeddings and keyword matching. Documents are split into semantic chunks and indexed. When you search, the system finds relevant chunks and optionally reranks them for better accuracy.

### *Q: Is my data private?*

Yes! You can run LotoAI completely locally with no external API calls. Use the local embedding mode (EMBEDDING_PROVIDER=local) for full privacy. Your documents never leave your infrastructure.

### *Q: How many documents can I upload?*

There's no hard limit on document count. The system scales to thousands of documents. Performance depends on your hardware resources.

### *Q: What is reranking?*

Reranking is a two-stage retrieval process. First, we find candidate documents using fast vector search. Then, a more accurate cross-encoder model reorders results by relevance. This improves answer quality significantly.

### *Q: Can I use this offline?*

Yes! Configure EMBEDDING_PROVIDER=local and you can run the entire RAG system without internet access. You'll need to download the embedding models once (about 500MB).

### *Q: How do I improve search quality?*

Enable reranking (ENABLE_RERANKING=1), use smaller chunk sizes (CHUNK_SIZE_CHARS=600), and ensure documents are well-formatted. For multilingual content, use the BGE-M3 embedding model.