Ragflow

Release 1

Jansen Jan

Nov 10, 2025

CONTENTS:

CHAPTER

ONE

ABOUT RAGFLOW: NAMED AMONG GITHUB'S FASTEST-GROWING OPEN SOURCE PROJECTS

October 28, 2025 · 3 min read

The release of **GitHub's 2025 Octoverse report** marks a pivotal moment for the open source ecosystem—and for projects like **RAGFlow**, which has emerged as **one of the fastest-growing open source projects by contributors this year**.

With a **remarkable 2,596% year-over-year growth in contributor engagement**, RAGFlow isn't just gaining traction—**it's defining the next wave of AI-powered development**.

—

1.1 The Rise of Retrieval-Augmented Generation in Production

As the Octoverse report highlights, **AI is no longer experimental—it's foundational**.

- **4.3 million+ AI-related repositories** on GitHub
- **1.1 million+ public repos** import LLM SDKs — a **178% YoY increase**

In this context, **RAGFlow's rapid adoption signals a clear shift**: developers are moving **beyond prototyping** and into **production-grade AI workflows**.

RAGFlow—an **end-to-end retrieval-augmented generation engine with built-in agent capabilities**—is perfectly positioned to meet this demand. It enables developers to build **scalable, context-aware AI applications** that are both **powerful and practical**.

> As the report notes: > *"AI infrastructure is emerging as a major magnet" for open source contributions.* > — **RAGFlow sits squarely at the intersection of AI infrastructure and real-world usability.**

—

1.2 Why RAGFlow Resonates in the AI Era

Several trends highlighted in the Octoverse report **align closely** with RAGFlow's design and mission:

1. **From Notebooks to Production** - Jupyter Notebooks: **+75% YoY** - Python codebases: **surging** - **RAGFlow supports this transition** with a **structured, reproducible framework** for deploying RAG systems in production.
2. **Agentic Workflows Are Going Mainstream** - GitHub Copilot coding agent launch - Rise of AI-assisted development - **RAGFlow's built-in agent capabilities** automate **retrieval, reasoning, and response generation**—key components of modern AI apps.

3. **Security and Scalability Are Top of Mind - 172% YoY increase** in Broken Access Control vulnerabilities - **RAGFlow's enterprise-ready deployment** helps teams address these challenges **secure-by-design**

—

1.3 A Project in Active Development

RAGFlow's evolution mirrors a **deliberate journey**—from solving foundational RAG challenges to **shaping the next generation of enterprise AI infrastructure**.

Phase 1: Solving Core RAG Limitations RAGFlow first made its mark by **systematically addressing core RAG limitations** through integrated technological innovation:

- **Deep document understanding** for parsing complex formats (PDFs, tables, forms)
- **Hybrid retrieval** blending multiple search strategies (vector, keyword, graph)
- **Built-in advanced tools**: **GraphRAG**, **RAPTOR**, and more
- Result: **dramatically enhanced retrieval accuracy and reasoning performance**

Phase 2: The Superior Context Engine for Enterprise Agents Now, building on this robust technical foundation, **RAGFlow is steering toward a bolder vision**:

> **To become the superior context engine for enterprise-grade Agents.**

- Evolving from a **specialized RAG engine** into a **unified, resilient context layer**
- Positioning itself as the **essential data foundation for LLMs in the enterprise**
- Enabling **Agents of any kind** to access **rich, precise, and secure context**
- Ensuring **reliable and effective operation across all tasks**

—

1.4 Conclusion

RAGFlow's **explosive growth** in the 2025 Octoverse is not a coincidence.

It reflects a **global developer movement** toward **production-ready, agentic, secure AI systems**—and RAGFlow is **leading the charge**.

From **deep document parsing** to **scalable agent workflows**, RAGFlow delivers the **infrastructure** and **usability** that modern AI demands.

The future of enterprise AI is context-aware, agent-driven, and open source—and RAGFlow is building it.

CHAPTER

TWO

WHY INFINIFLOW RAGFLOW USES A RERANKER, AN EMBEDDING MODEL, AND A CHAT MODEL

Infiniflow RAGFlow is a Retrieval-Augmented Generation (RAG) framework designed to build high-quality, traceable question-answering systems over complex data sources. To achieve accurate and contextually relevant responses, RAGFlow employs three distinct models that work in concert:

1. **Embedding Model** - **Purpose**: Converts both the user query and the chunks of retrieved documents into dense vector representations in the same semantic space. - **Role in Pipeline**: Enables semantic similarity search during the retrieval phase. By computing cosine similarity (or other distance metrics) between the query embedding and document chunk embeddings, RAGFlow retrieves the most semantically relevant passages from a large corpus—far beyond keyword matching.

2. **Reranker** - **Purpose**: Refines the initial retrieval results by re-scoring the top-*k* candidate chunks using a cross-encoder architecture. - **Role in Pipeline**: While the embedding model provides efficient approximate retrieval, the reranker applies a more computationally intensive but accurate relevance scoring. This step significantly improves precision by pushing the most contextually appropriate chunks to the top, reducing noise before generation.

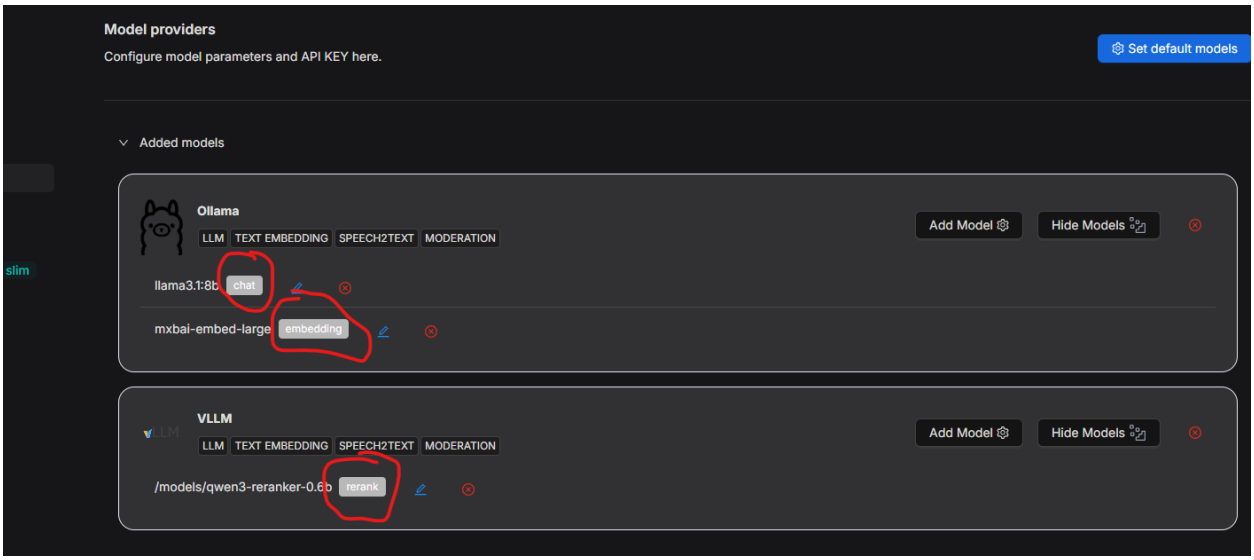

Fig. 1: **Figure 1**: The reranker evaluates query-chunk pairs to produce fine-grained relevance scores.

3. **Chat Model (LLM)** - **Purpose**: Generates the final natural language response grounded in the refined retrieved context. - **Role in Pipeline**: Takes the top reranked chunks as context and synthesizes a coherent, accurate, and fluent answer. The chat model (typically a large language model fine-tuned for instruction following) ensures the output is not only factually aligned with the source material but also conversational and user-friendly.

2.1 Synergy of the Three Models

- **Embedding Model** → Broad, fast, semantic retrieval
- **Reranker** → Precise, fine-grained reordering
- **Chat Model** → Coherent, grounded generation

This modular design allows RAGFlow to balance **speed**, **accuracy**, and **interpretability**, making it suitable for enterprise-grade RAG applications where both performance and trustworthiness are critical.

CHAPTER

THREE

WHY VLLM IS USED TO SERVE THE RERANKER MODEL

vLLM is a high-throughput, memory-efficient inference engine specifically designed for serving large language models (LLMs). In **Infiniflow RAGFlow**, the **reranker model**—responsible for fine-grained relevance scoring of retrieved document chunks—is served using **vLLM** to ensure low-latency, scalable, and production-ready performance.

3.1 Key Reasons for Using vLLM to Serve the Reranker

1. **PagedAttention for Memory Efficiency** - vLLM uses **PagedAttention**, a novel attention mechanism that manages KV cache in non-contiguous memory pages. - This dramatically reduces memory fragmentation and enables **higher batch sizes** and **longer sequence lengths** (up to 8192 tokens in this case), critical for processing query-chunk pairs during reranking.
2. **High Throughput & Low Latency** - Supports **continuous batching**, allowing dynamic batch formation as requests arrive. - Eliminates head-of-line blocking and maximizes GPU utilization—ideal for real-time reranking in interactive RAG pipelines.
3. **OpenAI-Compatible API** - Exposes a clean, standardized REST API compatible with OpenAI's format. - Enables seamless integration with RAGFlow's orchestration layer without custom inference code.
4. **Support for Cross-Encoder Rerankers** - Models like **Qwen3-Reranker-0.6B** are cross-encoders that take `[query, passage]` pairs as input. - vLLM efficiently handles the bidirectional attention required, delivering relevance scores via `logits[0]` (typically for binary classification: relevant/irrelevant).
5. **Ollama Does Not Support Reranker Models (Yet)** - **Ollama** is excellent for local LLM inference and chat models, but **currently lacks native support for reranker (cross-encoder) models**. - Rerankers require structured input formatting and logit extraction that Ollama's current API and model loading system do not accommodate. - vLLM, in contrast, supports any Hugging Face transformer model—including rerankers—with full access to outputs and fine-grained control.
6. **Scalability Advantage Over Ollama** - When scaling to **multiple concurrent users** or **high-throughput workloads**, vLLM is significantly more robust than Ollama. - vLLM supports **distributed serving**, **tensor parallelism**, **GPU clustering**, and **dynamic batching at scale**. - Ollama is primarily designed for **single-user, local development**, and does not scale efficiently in production environments.

3.2 Serving the Reranker Locally with vLLM

You can run the reranker model locally using vLLM with the following command:

```
vllm serve /models/qwen3-reranker-0.6b \
    --port 8123 \
    --max-model-len 8192 \
    --dtype auto \
    --trust-remote-code
```

Once running, the model is accessible via the OpenAI-compatible endpoint:

GET `http://localhost:8123/v1/models`

Example Response:


```
{
  "object": "list",
  "data": [
    {
      "id": "/models/qwen3-reranker-0.6b",
      "object": "model",
      "created": 1762258164,
      "owned_by": "vllm",
      "root": "/models/qwen3-reranker-0.6b",
      "parent": null,
      "max_model_len": 8192,
      "permission": [
        {
          "id": "modelperm-1a0d5938e30b4eeebb53d9e5c7d9599e",
          "object": "model_permission",
          "created": 1762258164,
          "allow_create_engine": false,
          "allow_sampling": true,
          "allow_logprobs": true,
          "allow_search_indices": false,
          "allow_view": true,
          "allow_fine_tuning": false,
          "organization": "*",
          "group": null,
          "is_blocking": false
        }
      ]
    }
  ]
}
```

3.3 RAGFlow Integration

RAGFlow configures the reranker endpoint in its settings:

```
reranker:
  provider: vllm
  api_base: http://localhost:8123/v1
  model: /models/qwen3-reranker-0.6b
```

During inference, RAGFlow sends batched `[query, passage]` pairs to the vLLM server, receives relevance scores, and reorders chunks before passing them to the chat model.

Result: Fast, accurate, and scalable reranking powered by optimized LLM inference—**where Ollama cannot currently follow, and where vLLM excels in both development and production.**

CHAPTER

FOUR

SERVING VLLM RERANKER USING DOCKER (CPU-ONLY)

To ensure **reproducibility**, **portability**, and **isolation**, **vLLM** can be deployed using **Docker**. This is especially useful in environments with restricted internet access (e.g., corporate networks behind proxies or firewalls), where **Hugging Face Hub** may be blocked or rate-limited.

In this setup, **vLLM runs on CPU only** because:

- **Laptop has no GPU**
- **Home server has an old NVIDIA GPU** (not supported by vLLM's CUDA requirements)

Thus, we use the **official CPU-optimized vLLM image** built from: https://github.com/vllm-project/vllm/blob/main/docker/Dockerfile.cpu

—

4.1 Docker Compose Configuration (CPU Mode)

```
version: '3.8'
services:
  qwen-reranker:
    image: vllm-cpu:latest
    ports: ["8123:8000"]
    volumes:
      - /home/naj/qwen3-reranker-0.6b:/models/qwen3-reranker-0.6b:ro
    environment:
      VLLM_HF_OVERRIDES: |
        {
          "architectures": ["Qwen3ForSequenceClassification"],
          "classifier_from_token": ["no", "yes"],
          "is_original_qwen3_reranker": true
        }
    command: >
      /models/qwen3-reranker-0.6b
      --task score
      --dtype float32
      --port 8000
      --trust-remote-code
      --max-model-len 8192
    deploy:
      resources:
        limits:
```

(continues on next page)

(continued from previous page)

```
        cpus: '10'
        memory: 16G
    shm_size: 4g
    restart: unless-stopped
```

—

4.2 Key Components Explained

- ``**image: vllm-cpu:latest**`` - Official vLLM CPU image (no CUDA dependencies). - Built from: vllm-project/vllm/docker/Dockerfile.cpu - Uses PyTorch CPU backend with optimized inference kernels.
- ``**ports: ["8123:8000"]**`` - Host port **8123** → container port **8000** (vLLM default).
- ``**volumes**`` - Mounts **locally pre-downloaded model** in **read-only** mode.
- ``**VLLM_HF_OVERRIDES**`` - Required for **Qwen3-Reranker** due to custom classification head and token handling.
- ``**command**`` - `--task score`: Enables reranker scoring (outputs relevance logits). - `--dtype float32`: Mandatory on CPU (no half-precision support). - `--max-model-len 8192`: Supports long query+passage pairs.
- **Resource Limits** - `cpus: '10'` and `memory: 16G` prevent system overload. - `shm_size: 4g` ensures sufficient shared memory for batched inference.

—

4.3 Why the Model Must Be Pre-Downloaded Locally

The container **cannot download the model at runtime** due to:

1. **Corporate Proxy / Firewall** - Outbound traffic to `huggingface.co` is blocked or requires authentication.
2. **Hugging Face Hub Blocked** - Git LFS and model downloads fail in restricted networks.
3. **vLLM Auto-Download Fails Offline** - vLLM uses `transformers.AutoModel` → attempts online download if model not found.

Solution: Download via mirror

```
HF_ENDPOINT=https://hf-mirror.com huggingface-cli download Qwen/Qwen3-Reranker-0.6B --
↪local-dir ./qwen3-reranker-0.6b
```

- Remark : for some models you need a token HF_TOKEN=xxxxxxxx (you have to specify the model in the token definition!)
- Remark2 : use "sudo" if non-root!!!
- Uses **accessible mirror** (`hf-mirror.com`).
- Saves model locally for volume mounting.

—

4.4 Why CPU-Only (No GPU)?

- **Laptop**: Integrated graphics only (no discrete GPU).
- **Home Server**: NVIDIA GPU too old (e.g., pre-Ampere) → **not supported** by vLLM's CUDA 11.8+ / FlashAttention requirements.
- **vLLM CPU image** enables full functionality without GPU.

> **Performance Note**: CPU inference is slower (~1–3 sec per batch), but sufficient for **development**, **prototyping**, or **low-throughput** use cases.

—

4.5 Start the Service

```
docker-compose up -d
```

4.6 Verify Availability

```
curl http://localhost:8123/v1/models
```

Expected output confirms the model is loaded and ready.

—

4.7 Integration with RAGFlow

Update RAGFlow config:

```
reranker:
  provider: vllm
  api_base: http://localhost:8123/v1
  model: /models/qwen3-reranker-0.6b
```

—

4.8 Benefits of This CPU + Docker Setup

- **Works on any machine** (laptop, old server, air-gapped systems)
- **No GPU required**
- **Offline-first** with pre-downloaded model
- **Consistent environment** via Docker
- **Secure**: read-only model, isolated container
- **Scalable later**: switch to GPU image when hardware upgrades

Ideal for local RAGFlow development and constrained production environments.

CHAPTER

FIVE

INTEGRATING VLLM WITH RAGFLOW VIA DOCKER NETWORK

To enable **Infiniflow RAGFlow** (running in Docker) to communicate with a **vLLM reranker container**, both services must be on the **same Docker network**. By default, containers are isolated and cannot resolve each other by service name unless explicitly networked.

In this setup, we ensure seamless internal communication between:

- **RAGFlow** (web + backend containers)
- **vLLM reranker** (serving *Qwen3-Reranker-0.6B*)

—

5.1 Why Network Configuration is Required

- RAGFlow runs inside Docker (typically via *docker-compose*).
- vLLM reranker runs in a **separate container** (e.g., CPU-only).
- RAGFlow needs to call: *http://<vllm-service-name>:8000/v1* internally.
- Without shared network → *Connection refused* or DNS lookup failure.

Solution: Attach both services to a **custom Docker bridge network** (e.g., *docker-ragflow*).

—

5.2 Step-by-Step: Configure Docker Network

1. Create a Custom Network

```
docker network create docker-ragflow
```

—

2. Update *docker-compose.yaml* for vLLM Reranker

Ensure the vLLM service uses the network:

Listing 1: docker-compose.yaml (vLLM)

```
version: '3.8'
services:
  qwen-reranker:
```

(continues on next page)

(continued from previous page)

```
    image: vllm-cpu:latest
    container_name: ragflow-vllm-reranker
    ports: ["8123:8000"]
    volumes:
      - /home/naj/qwen3-reranker-0.6b:/models/qwen3-reranker-0.6b:ro
    environment:
      VLLM_HF_OVERRIDES: |
        {
          "architectures": ["Qwen3ForSequenceClassification"],
          "classifier_from_token": ["no", "yes"],
          "is_original_qwen3_reranker": true
        }
    command: >
      /models/qwen3-reranker-0.6b
      --task score
      --dtype float32
      --port 8000
      --trust-remote-code
      --max-model-len 8192
    deploy:
      resources:
        limits:
          cpus: '10'
          memory: 16G
    shm_size: 4g
    restart: unless-stopped
    networks:
      - docker-ragflow

networks:
  docker-ragflow:
    external: true
```

—

3. Connect RAGFlow Containers to the Same Network

If RAGFlow is already running via its own *docker-compose*, **attach** it:

```
docker network connect docker-ragflow ragflow-web
docker network connect docker-ragflow ragflow-server
```

> Replace *ragflow-web*, *ragflow-server* with actual container names (check with *docker ps*).

—

4. Configure RAGFlow to Use Internal vLLM Endpoint

In RAGFlow settings (UI or config file), set:

```
reranker:
  provider: vllm
  api_base: http://ragflow-vllm-reranker:8000/v1
  model: /models/qwen3-reranker-0.6b
```

Key: Use **container name** (*ragflow-vllm-reranker*) — Docker DNS resolves it automatically within the network.

—

5.3 Architecture Diagram

—

5.4 Verification

1. **From RAGFlow container**, test connectivity:

```
docker exec -it ragflow-server curl http://ragflow-vllm-reranker:8000/v1/models
```

2. **Expected output**:

```
{
  "object": "list",
  "data": [
    {
      "id": "/models/qwen3-reranker-0.6b",
      ...
    }
  ]
}
```

—

5.5 Benefits of This Setup

- **Zero external exposure** (optional): vLLM accessible **only** within *docker-ragflow* network.
- **Secure & fast** internal communication.
- **Scalable**: Add more rerankers, LLMs, or vector DBs on same network.
- **Portable**: Works across dev, staging, production with same config.

—

5.6 Troubleshooting Tips

Issue | Solution |

|------|————-| | *Connection refused* | Check network: *docker network inspect docker-ragflow* | | *Unknown host* | Use **container name**, not *localhost* | | Port conflict | Ensure no other service uses *8000* inside network | | Model not loading | Verify volume mount and *trust-remote-code* |

—

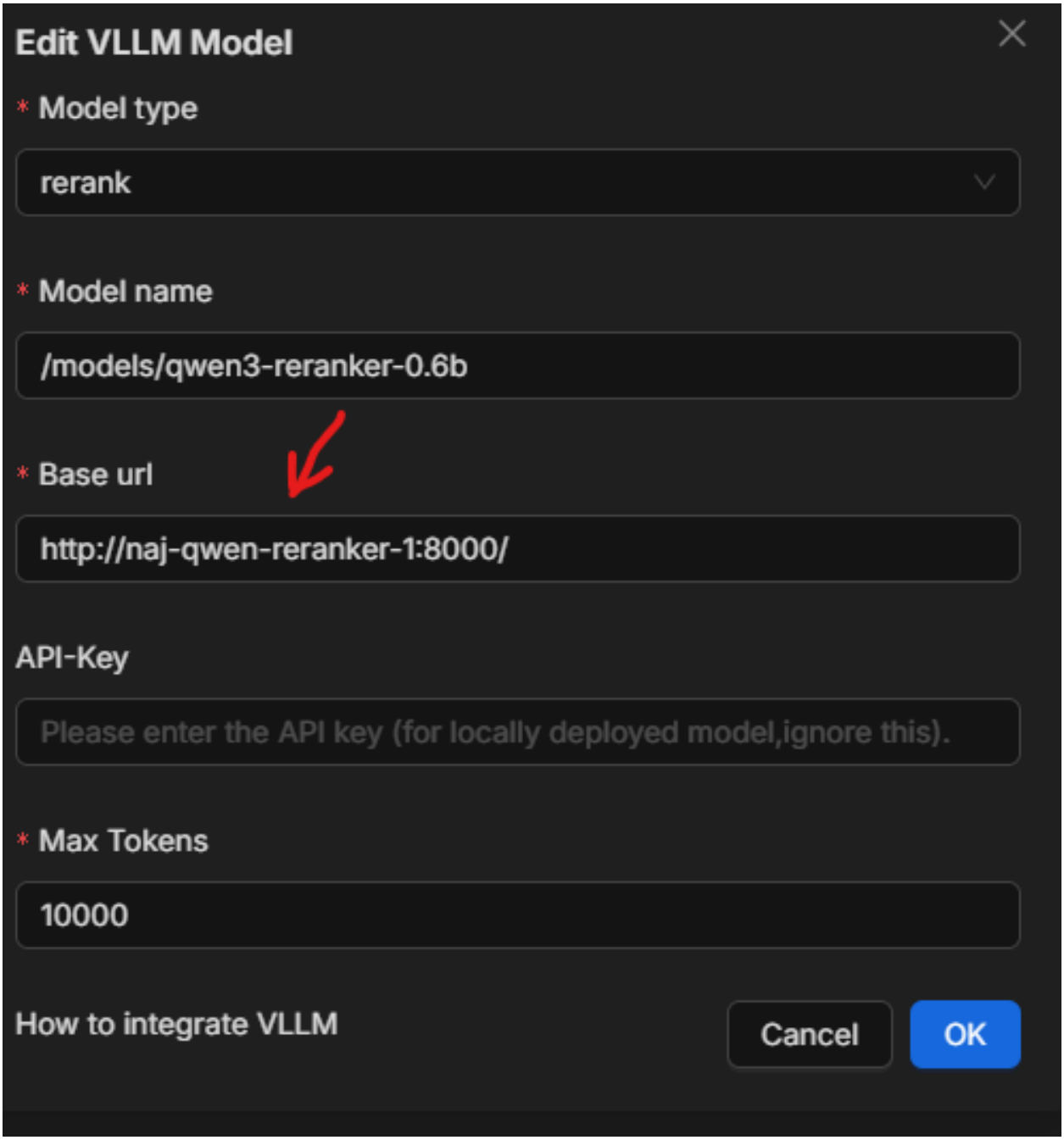

Fig. 1: **Figure 1**: RAGFlow containers communicate with vLLM reranker via internal Docker network *docker-ragflow*. External access (optional) via port *8123*.

5.7 Summary

To use **vLLM inside RAGFlow Docker environment**:

1. Create network: *docker network create docker-ragflow*
2. Connect both RAGFlow and vLLM containers
3. Use **container name** in *api_base*
4. Enjoy **fast, secure, internal reranking**

> **No need for public IPs, reverse proxies, or complex routing** — Docker handles it all.

CHAPTER

SIX

BATCH PROCESSING AND METADATA MANAGEMENT IN INFINIFLOW RAGFLOW

Infiniflow RAGFlow provides a **RESTful API** (*/api/v1*) that enables **programmatic control** over datasets and documents, making it ideal for **batch processing large volumes of documents**, **automated ingestion pipelines**, and **metadata enrichment**.

This is essential in enterprise settings where thousands of PDFs, reports, or web pages need to be:

- Ingested in bulk
- Tagged with structured metadata (author, date, source, category, etc.)
- Updated post-ingestion
- Queried or filtered later via the RAG system

—

6.1 API Base URL

```
http://<RAGFLOW_HOST>/api/v1
```

6.2 Authentication

All requests require a **Bearer token**:

```
Authorization: Bearer ragflow-<your-token>
```

> **Tip**: Obtain token via login or API key management in the RAGFlow UI.

—

6.3 Step 1: Retrieve Dataset and Document IDs

Before updating, you **must know** the target:

- **Dataset ID** (e.g., *f388c05e9df711f0a0fe0242ac170003*)
- **Document ID** (e.g., *4920227c9eb711f0bff40242ac170003*)

List all datasets:

```
curl -H "Authorization: Bearer ragflow-..." \
     http://192.168.0.213/api/v1/datasets
```

List documents in a dataset:

```
curl -H "Authorization: Bearer ragflow-..." \
     http://192.168.0.213/api/v1/datasets/<dataset_id>/documents
```

—

6.4 Step 2: Add Metadata to a Document (via PUT)

Use the **PUT** endpoint to **update metadata** of an existing document:

```
curl --request PUT \
     --url http://192.168.0.213/api/v1/datasets/f388c05e9df711f0a0fe0242ac170003/
↪documents/4920227c9eb711f0bff40242ac170003 \
     --header 'Content-Type: multipart/form-data' \
     --header 'Authorization: Bearer ragflow-QxNWIzMGNlOWRmMzExZjBhZjljMDI0Mm' \
     --data '{
       "meta_fields": {
         "author": "Example Author",
         "publish_date": "2025-01-01",
         "category": "AI Business Report",
         "url": "https://example.com/report.pdf"
       }
     }'
```

Request Breakdown:

- **Method**: *PUT*
- **Path**: */api/v1/datasets/<dataset_id>/documents/<document_id>*
- **Content-Type**: *multipart/form-data* (required even for JSON payload)
- **Body**: JSON string with *"meta_fields"* object

Response (on success):

```
{
  "code": 0,
  "message": "Success",
  "data": { "document_id": "4920227c9eb711f0bff40242ac170003" }
}
```

—

6.5 Use Case: Batch Metadata Enrichment

You can **automate metadata tagging** for **1000s of documents** using a script:

```
import requests
import json

BASE_URL = "http://192.168.0.213/api/v1"
TOKEN = "ragflow-QxNWIzMGNlOWRmMzExZjBhZjljMDI0Mm"
HEADERS = {
    "Authorization": f"Bearer {TOKEN}",
    "Content-Type": "multipart/form-data"
}

# Example: Load CSV with doc_id, author, date, url...
import pandas as pd
df = pd.read_csv("documents_metadata.csv")

for _, row in df.iterrows():
    dataset_id = row['dataset_id']
    doc_id = row['document_id']
    payload = {
        "meta_fields": {
            "author": row['author'],
            "publish_date": row['publish_date'],
            "category": row['category'],
            "url": row['source_url']
        }
    }

    files = {'': ('', json.dumps(payload), 'application/json')}
    resp = requests.put(
        f"{BASE_URL}/datasets/{dataset_id}/documents/{doc_id}",
        headers=HEADERS,
        files=files
    )
    print(doc_id, resp.json().get("message"))
```

Benefits:

- Enrich RAG context with **structured, queryable metadata**
- Enable **filtering** in UI or API (e.g., "Show reports from 2025 by Author X")
- Improve **traceability** and **auditability**

—

6.6 Other Batch-Capable Endpoints

Endpoint | Purpose |

|--------|———| | *POST /api/v1/datasets* | Create new dataset | | *POST /api/v1/datasets/{id}/documents* | Upload new documents (with metadata) | | *DELETE /api/v1/datasets/{id}/documents/{doc_id}* | Remove document | | *GET /api/v1/datasets/{id}/documents* | List + filter by metadata |

—

6.7 Best Practices

1. **Always use IDs** — never rely on filenames
2. **Batch in chunks** (e.g., 100 docs/sec) to avoid rate limits
3. **Validate metadata schema** in RAGFlow settings first
4. **Log responses** for retry logic
5. **Use dataset-level permissions** for access control

—

6.8 Summary

RAGFlow's **API-first design** enables:

- **Scalable batch ingestion**
- **Rich metadata attachment**
- **Full automation** of document lifecycle

> **Perfect for ETL pipelines, CMS integration, or enterprise knowledge base automation.**

With this API, you can manage **tens of thousands of documents** with full metadata — all programmatically.

CHAPTER

SEVEN

HOW THE KNOWLEDGE GRAPH IN INFINIFLOW/RAGFLOW WORKS

RAGFlow is an open-source Retrieval-Augmented Generation (RAG) engine developed by Infiniflow, designed to enhance LLM-based question-answering by integrating deep document understanding with structured data processing. The knowledge graph (KG) component plays a pivotal role in handling complex queries, particularly multi-hop question-answering, by extracting and organizing entities and relationships from documents into a graph structure. This enables more accurate and interconnected retrieval beyond simple vector-based searches.

7.1 Overview

The KG is constructed as an intermediate step in RAGFlow's data pipeline, bridging raw document extraction and final indexing. It transforms unstructured text into a relational graph, allowing for entity-based reasoning and traversal during retrieval. Key benefits include:

- **Multi-Hop Query Support**: Facilitates queries requiring inference across multiple documents or concepts (e.g., "What caused the event that affected company X?").
- **Dynamic Updates**: From version 0.16.0 onward, the KG is built across an entire knowledge base (multiple files) and automatically updates when new documents are uploaded and parsed.
- **Integration with RAG 2.0**: Part of preprocessing stages like document clustering and domain-specific embedding, ensuring retrieval results are contextually rich and grounded.

The KG is stored as chunks in RAGFlow's document engine (Elasticsearch by default or Infinity for advanced vector/graph capabilities), making it queryable alongside embeddings.

7.2 Construction Process

The KG construction occurs after initial document parsing but before indexing. Here's the step-by-step workflow:

1. **Document Ingestion and Extraction**: - Users upload files (e.g., PDF, Word, Excel, TXT) to a knowledge base. - RAGFlow's Deep Document Understanding (DDU) module parses the content, extracting structured elements like text blocks, tables, and layouts using OCR and layout models.
2. **Entity and Relation Extraction**: - Using NLP models (integrated via configurable LLMs or embedding services), RAGFlow identifies entities (e.g., people, organizations, events) and relations (e.g., "causes", "affiliated with") from extracted chunks. - This is model-driven: Preprocessing applies entity recognition and relation extraction to raw text, often leveraging domain-specific prompts for accuracy.
3. **Graph Building**: - Entities become nodes, and relations form directed/undirected edges. - The graph is unified across the entire dataset (not per-file since v0.16.0), enabling cross-document connections. - Acceleration features (introduced in later releases) optimize extraction speed, such as batch processing or efficient model inference.

4. **Storage**: - Graph chunks (nodes, edges, metadata) are serialized and stored in the document engine. - No separate graph database is required; it's embedded within the vector/full-text index for hybrid queries.

Note: Construction can be toggled per knowledge base and is optional, but recommended for complex domains like finance or healthcare.

7.3 Query and Retrieval Process

During inference, the KG enhances retrieval in the following manner:

1. **Query Parsing**: - Incoming user queries are analyzed to detect multi-hop intent (e.g., via LLM routing).
2. **Hybrid Retrieval**: - Combine vector similarity search (for semantic relevance) with graph traversal:
    - Start from query entities as seed nodes.
    - Traverse edges to fetch connected nodes (e.g., 1-2 hops).
    - Rank results by relevance scores, incorporating graph proximity.
    - Infinity engine (optional) supports efficient graph-range filtering alongside vectors.
3. **Augmentation and Generation**: - Retrieved graph-derived contexts (e.g., subgraphs or paths) are fused with text chunks. - Fed to the LLM for grounded generation, with citations traceable to source documents.

This process addresses limitations of pure vector RAG, such as hallucination in interconnected scenarios, by providing explicit relational paths.

7.4 Key Features and Limitations

- **Features**: - **Scalability**: Handles enterprise-scale knowledge bases with dynamic rebuilding. - **Customizability**: Configurable extraction models and hop limits. - **Agent Integration**: Supports agentic workflows for iterative graph exploration. - **Performance**: Accelerated extraction in v0.21+ releases.
- **Limitations**: - Relies on quality of upstream extraction; noisy documents may yield incomplete graphs. - Graph depth is configurable but can increase latency for deep traversals. - Arm64 Linux support is limited when using Infinity.

For implementation details, refer to the official guide at https://github.com/infiniflow/ragflow/blob/main/docs/guides/dataset/construct_knowledge_graph.md. To experiment, deploy RAGFlow via Docker and enable KG in your knowledge base settings.

CHAPTER

EIGHT

RUNNING LLAMA 3.1 WITH LLAMA.CPP

Table of Contents

8.1 1. Model Format: GGUF

`llama.cpp` uses the **GGUF** (GPT-Generated Unified Format) model format.

You can download a pre-quantized **Llama 3.1 8B** model in GGUF format directly from Hugging Face:

```
https://huggingface.co/QuantFactory/Meta-Llama-3-8B-GGUF
```

Example file: `Meta-Llama-3-8B.Q4_K_S.gguf` (~4.7 GB, 4-bit quantization, excellent quality/size tradeoff).

Note: The `Q4_K_S` variant uses ~4.7 GB RAM and runs efficiently on Intel i7 CPUs.

8.2 2. Compile llama.cpp for Intel i7 (CPU-only)

```
# Step 1: Clone the repository
git clone https://github.com/ggerganov/llama.cpp
cd llama.cpp

# Step 2: Build for CPU (Intel i7, AVX2 enabled by default)
make clean
make -j$(nproc) LLAMA_CPU=1

# Optional: Force AVX2 (most i7 CPUs support it)
make clean
make -j$(nproc) LLAMA_CPU=1 LLAMA_AVX2=1
```

The binaries will be in the root directory:

- `./llama-cli` → interactive CLI
- `./server` → web server (OpenAI-compatible API + full web UI)

> **Warning:** **Do not use vLLM** on older Xeon v2 CPUs — they **lack AVX-512**, which vLLM requires. **llama.cpp is a better choice** — it runs efficiently with just AVX2 or even SSE.

8.3 3. Run the Model with Web Interface

Place the downloaded GGUF file in a `models/` folder:

```
mkdir -p models
# Copy or symlink the model
ln -s /path/to/Meta-Llama-3-8B.Q4_K_S.gguf models/
```

Start the server on port **8087** using **12 threads**:

```
./server \
  --model models/Meta-Llama-3-8B.Q4_K_S.gguf \
  --port 8087 \
  --threads 12 \
  --host 0.0.0.0
```

8.3.1 Features

- **Full web UI** at: http://localhost:8087
- **OpenAI-compatible API** at: http://localhost:8087/v1
- List models:

```
curl http://localhost:8087/v1/models
```

Response:


```
{
  "object": "list",
  "data": [
    {
      "id": "models/Meta-Llama-3-8B.Q4_K_S.gguf",
      "object": "model",
      "created": 1762783003,
      "owned_by": "llamacpp",
      "meta": {
        "vocab_type": 2,
        "n_vocab": 128256,
        "n_ctx_train": 8192,
        "n_embd": 4096,
        "n_params": 8030261248,
        "size": 4684832768
      }
    }
  ]
}
```


8.4 4. Compile llama.cpp with NVIDIA GPU Support (CUDA)

If you have an **NVIDIA GPU** (e.g., RTX 3060, 4070, A100, etc.), enable **CUDA acceleration**:

8.4.1 Prerequisites

- NVIDIA driver (\geq{} 525)
- CUDA Toolkit (\geq{} 11.8, preferably 12.x)
- `nvcc` in `$PATH`

8.4.2 Build with CUDA

```
# Clean previous build
make clean

# Build with full CUDA support
make -j$(nproc) \
  LLAMA_CUDA=1 \
  LLAMA_CUDA_DMMV=1 \
  LLAMA_CUDA_F16=1

# Optional: Specify compute capability (e.g., for RTX 40xx)
# make LLAMA_CUDA=1 CUDA_ARCH="-gencode arch=compute_89,code=sm_89"
```

8.4.3 Run with GPU offloading

```
./server \
  --model models/Meta-Llama-3-8B.Q4_K_S.gguf \
  --port 8087 \
  --threads 8 \
  --n-gpu-layers 999 \   # offload ALL layers to GPU
  --host 0.0.0.0
```

Tip: Use `nvidia-smi` to monitor VRAM usage. For 8B Q4 (~4.7 GB), even a **6 GB GPU** can run it fully offloaded.

8.5 Summary

Feature	Command / Note
Model Format	**GGUF**
Download	https://huggingface.co/QuantFactory/...GGUF
CPU Build (i7)	`make LLAMA_CPU=1`
GPU Build (CUDA)	`make LLAMA_CUDA=1`
Run Server	`./server --model ... --port 8087`
Web UI	http://localhost:8087
API	http://localhost:8087/v1

llama.cpp = lightweight, CPU/GPU flexible, no AVX-512 needed → ideal replacement for vLLM on older hardware.

—

CHAPTER

NINE

RUNNING MULTIPLE MODELS ON LLAMA.CPP USING DOCKER

This guide demonstrates how to run multiple language models simultaneously using `llama.cpp` in Docker via `docker-compose`. The example below defines two services: a lightweight reranker model (Qwen3 0.6B) and a general-purpose chat model (Llama 3.1).

9.1 Example `docker-compose.yml`

```yaml
services:
  qwen-reranker:
    image: ghcr.io/ggerganov/llama.cpp:server-cpu
    ports:
      - "8123:8080"
    volumes:
      - /home/naj/qwen3-reranker-0.6b:/models/qwen3-reranker-0.6b:ro
    environment:
      - MODEL=/models/qwen3-reranker-0.6b
    command: >
      --model /models/qwen3-reranker-0.6b/model.gguf
      --port 8080
      --host 0.0.0.0
      --n-gpu-layers 0
      --ctx-size 8192
      --threads 6
      --temp 0.0
      --rpc
    deploy:
      resources:
        limits:
          cpus: '6'
          memory: 10G
    shm_size: 4g
    restart: unless-stopped

  llama3.1-chat:
    image: ghcr.io/ggerganov/llama.cpp:server-cpu
    ports:
      - "8124:8080"
    volumes:
      - /home/naj/llama3.1:/models/llama3.1:ro
```

(continues on next page)

(continued from previous page)

```
    environment:
      - MODEL=/models/llama3.1
    command: >
      --model /models/llama3.1/model.gguf
      --port 8080
      --host 0.0.0.0
      --n-gpu-layers 0
      --ctx-size 8192
      --threads 10
      --temp 0.7
      --rpc
    deploy:
      resources:
        limits:
          cpus: '10'
          memory: 20G
    shm_size: 8g
    restart: unless-stopped
```

9.2 Key Configuration Notes

- **Images**: Both services use the official CPU-optimized `llama.cpp` server image.
- **Ports**: - Reranker exposed on `8123` → internal `8080` - Chat model exposed on `8124` → internal `8080`
- **Volumes**: Model directories are mounted read-only (`:ro`) from the host.
- **Environment**: `MODEL` variable simplifies path references in commands.
- **Command Flags**: - `--n-gpu-layers 0`: Forces CPU-only inference. - `--ctx-size 8192`: Sets context length. - `--temp`: Controls randomness (0.0 for deterministic reranking, 0.7 for chat). - `--rpc`: Enables RPC interface for external control.
- **Resource Limits**: CPU and memory capped via `deploy.resources.limits`.
- **Shared Memory (shm_size)**: Increased to support larger contexts and batching.
- **Restart Policy**: `unless-stopped` ensures containers restart on failure or reboot.

9.3 Usage

Start the services:

```
docker compose up -d
```

Access the models:

- Reranker: `http://localhost:8123`
- Chat: `http://localhost:8124`

Send requests using the OpenAI-compatible API or `llama.cpp` client tools.

```rst .. _deployment-considerations:
```

CHAPTER

TEN

DEPLOYING LLMS IN HYBRID CLOUD: WHY LLAMA.CPP WINS FOR US

Table of Contents

10.1 1. Current Setup: Ollama in Testing

We have been using **Ollama** in a **test environment** with excellent results:

- **Easy to use** — *ollama run llama3.1* just works
- **Docker support** is first-class:

```
FROM ollama/ollama
COPY Modelfile /root/.ollama/
RUN ollama create my-llama3.1 -f Modelfile
```

- Models are pulled, versioned, and cached automatically
- Web UI and OpenAI-compatible API available out of the box

Verdict: Perfect for **prototyping**, **local dev**, and **small-scale testing**.

10.2 2. Production Requirements: Hybrid Cloud & Multi-User Access

When moving to **production in a hybrid cloud**, new constraints emerge:

We need a **lightweight, portable, hardware-agnostic** inference engine.

10.3 3. Evaluation: vLLM vs llama.cpp

Criteria	vLLM	llama.cpp
Hardware Requirements	Requires **AVX-512** (fails on Xeon v2, older i7)	Runs on **SSE2+**, AVX2 optional
GPU Support	Excellent (PagedAttention, high throughput)	CUDA, Metal, Vulkan — full offloading
CPU Performance	Poor without AVX-512	**Best-in-class** quantized inference
Binary Size	~200 MB + Python deps	**< 10 MB** (statically linked)
Deployment	Python server, complex deps	Single binary, *scp* and run
Multi-user / API	Built-in OpenAI API	*server* binary with full OpenAI compat + web UI
Quantization Support	FP16/BF16 only	Q4_K, Q5_K, Q8_0, etc. — **4–8 GB models fit in RAM**

Key Finding:

> **We cannot use vLLM** on our legacy Xeon v2 fleet due to missing **AVX-512**. > **llama.cpp runs efficiently** on the **same hardware** with **Q4_K_M** models.

10.4 4. Why llama.cpp Is Our Production Choice

Decision

llama.cpp is selected for **hybrid cloud LLM deployment** because:

- **Runs everywhere**: Old CPUs, new GPUs, laptops, edge
- **Single static binary**: No Python, no CUDA runtime hell
- **GGUF format**: Share models with Ollama, local files, S3
- **Built-in server**: OpenAI API + full web UI
- **Thread & context control**: *–threads*, *–ctx-size*, *–n-gpu-layers*
- **Kubernetes-ready**: Tiny image, fast startup

10.4.1 Example: Production-Ready Server

```
./llama.cpp/server \
  --model /models/llama3.1-8b-instruct.Q4_K_M.gguf \
  --port 8080 \
  --host 0.0.0.0 \
  --threads 16 \
  --ctx-size 8192 \
  --n-gpu-layers 0    # CPU-only on older nodes
  --log-disable
```

Deploy via Docker:

```
FROM alpine:latest
COPY llama.cpp/server /usr/bin/
COPY models/*.gguf /models/
EXPOSE 8080
CMD ["server", "--model", "/models/llama3.1-8b-instruct.Q4_K_M.gguf", "--port", "8080"]
```

10.5 5. Migration Path: From Ollama → llama.cpp

```
# 1. Reuse Ollama's GGUF
cp ~/.ollama/models/blobs/sha256-* /production/models/

# 2. Deploy llama.cpp server
kubectl apply -f llama-cpp-deployment.yaml

# 3. Point clients to new endpoint
export OPENAI_API_BASE=http://llama-cpp-prod:8080/v1
```

Zero model reconversion. Zero downtime.

10.6 Summary

Use Case	Recommended Tool
Local dev / prototyping	**Ollama**
Hybrid cloud, old hardware, scale	**llama.cpp**
High-throughput GPU cluster	vLLM (if AVX-512 available)

> llama.cpp = the Swiss Army knife of LLM inference.

—

CHAPTER

ELEVEN

INDICES AND TABLES

- genindex
- modindex
- search