Task 4. Deployment Strategy

Infrastructure

The system is divided into modular services for maintainability and scalability:

- **FastAPI App**: Main API handling sentiment classification and similar review retrieval via ChromaDB. It can optionally connect to an external model.
- **Chroma Vector DB**: Self-hosted and persisted for similarity search over embedded reviews.
- **Embedding Service**: Runs periodically (for example, every night) to detect new reviews, generate embeddings and update the vector DB.
- **Review Database**: Stores raw reviews in plain text, used as input for embedding generation.
- **Optional Model Endpoint**: The sentiment model can be hosted externally to offload compute from the API.

Scalability Approach

- Use containerized services with Docker and optionally Kubernetes for orchestration.
- Asynchronous tasks like embedding updates are decoupled from the main API.
- Externalizing the model allows independent scaling of inference.

Model & Data Storage

- **Model**: The system uses DistilBERT for sentiment classification due to its excellent balance between performance and efficiency. It delivers strong accuracy on sentiment tasks while being significantly smaller and faster to train/infer than larger models like BERT or RoBERTa. If higher accuracy is needed, roberta-base could be considered as an alternative. For future multilingual support, variants such as distilbert-base-multilingual-cased or xlm-roberta-base may be evaluated.
- **Embeddings**: Stored in ChromaDB.

- **Reviews**: Stored in a relational DB (example: SQLite or PostgreSQL) and synchronized with the vector store.

- **Embeddings pipeline**: Automates ingestion and updates via scheduled background tasks.

Resource & Cost Considerations

- Deploy on low-cost VMs or serverless containers (example: GCP Cloud Run or AWS).

- Optimize inference cost with Lightweight transformer model.

- Schedule heavy tasks (like embedding) during off-peak hours.

- Externalizing the model allows pay-per-inference options if usage is moderate.