

Stock Prediction System Orchestration Plan

Project Task List

The system consists of eight core tasks: Data acquisition is responsible for obtaining stock price data from data sources; Data validation checks data quality and completeness; Data preprocessing performs cleaning, standardization, and feature engineering; Model training updates LSTM model parameters.

Task Dependencies

After data acquisition is completed, data validation is triggered. Once validated, the process splits into two paths: one proceeds to data preprocessing and subsequently generates daily predictions; the other triggers the model training process.

Input/Output Specifications

The data acquisition task retrieves data via API interfaces and outputs raw data files; the data validation task reads raw data and outputs validation reports and cleaned data; the data preprocessing task generates processed data and standardized files; the model training task reads processed data and model configurations, outputting model weights and training logs.

Logging and Checkpoint Strategy

Checkpoint settings cover the entire process: the data pipeline saves intermediate data after each processing step, and model training saves the best weights based on validation loss. Checkpoints are stored in four designated directories: raw data checkpoints, processed data checkpoints, model checkpoints, and prediction archives.

Automation vs. Manual Task Division

The scope of immediate automation includes: repetitive data acquisition and validation. Tasks retained for manual control involve: model architecture changes requiring validation, hyperparameter tuning requiring careful evaluation, and major data source changes requiring risk assessment.

Rationale for Manual Task Retention

Model changes require rigorous backtesting to avoid significant financial losses in live trading environments. Automated model updates may introduce uncontrollable risks, and human oversight ensures compliance with risk management systems and regulatory requirements.

Error Handling Mechanism

If data acquisition fails, an exponential backoff strategy is used to retry three times. If daily data is unavailable, the system automatically falls back to the previous day's data. If model performance degrades by more than 20%, it automatically rolls back to the previous version.

Resource Requirements Configuration

The system requires 50 GB of storage space for data and model storage, 16 GB of memory for model training, 4 vCPUs for daily tasks, and GPU acceleration for weekly training. Data acquisition

depends on a stable network connection.