



# Gateway ML Studio: A Self-Optimizing ML Stack – Whitepaper

## Introduction and Problem Background

Gateway ML Studio is a **self-optimizing machine learning stack** designed for robust, real-time analytics across diverse tasks. Modern data-driven applications often require a combination of supervised learning for prediction, anomaly detection for rare events, and time-series forecasting for trends. Managing separate models for each task can be complex, and static ensembles may struggle to adapt to changing data patterns. Gateway ML Studio addresses these challenges by integrating multiple model types (classification, anomaly detection, forecasting) into a unified system with an **RL-governed ensemble routing** mechanism. This system can dynamically route inputs to the best model or combination of models, leveraging reinforcement learning (RL) to continuously improve decision policies. The platform is intended for use by internal developers, external ML contributors, and stakeholders who need interpretable, reliable model outputs for critical applications (e.g. financial depeg detection, operational monitoring). In this introduction, we outline the vision: Gateway ML Studio provides a *single, orchestrated ML pipeline* that is **adaptive, fault-tolerant**, and **self-monitoring**, aiming to reduce integration effort and improve predictive performance across scenarios.

*Background:* Traditional ML systems often operate in silos – one model per task – making it hard to coordinate when multiple insights are needed. For example, a financial monitoring solution might need a classifier to detect fraud, anomaly detectors to catch market deviations (depegs), and forecasters to project future metrics. Without a unifying framework, such solutions can suffer from inconsistent outputs, unhandled edge cases, and difficulty in monitoring. Gateway ML Studio’s “self-optimizing” approach means it can learn from feedback (e.g. errors, drifts, rewards) to automatically tune which models are used when, aiming for **continuous improvement**. This whitepaper provides a detailed overview of the system’s architecture and components, from core models to orchestration and deployment, serving as documentation for developers and stakeholders.

## Overview of System Architecture

**Gateway ML Studio's architecture** is composed of multiple layers of models and controllers orchestrated in a cohesive pipeline. At a high level, data flows from ingestion through preprocessing into a suite of models (supervised learners, anomaly detectors, forecasters). These models feed into an **ensemble routing system** governed by a reinforcement learning agent (the *Router/Governor*), which decides how to combine or select model outputs. An *Overseer LLM* (a large language model) acts as a meta-controller, providing high-level analysis or policy adjustments based on system-wide outputs (for example, explaining anomalies or adjusting routes based on contextual rules). The entire workflow is coordinated by an orchestration layer (built with **n8n** workflows) and exposed via a **FastAPI gateway**. This gateway provides unified API endpoints for predictions and alerts, abstracting away the complexity behind a simple interface.

(Figure 1: Gateway ML Studio Architecture – data flows from sources through a preprocessing pipeline into multiple models (classification, anomaly detection, forecasting). Model outputs are aggregated by a Router/Governor (RL agent) which decides final outputs or routes. An Overseer LLM monitors outputs and suggests adjustments. All components are orchestrated by n8n workflows, and a FastAPI service provides an external API. Monitoring (Prometheus) taps into each component to track metrics.)

### Components:

- **Data Pipeline:** Ingests and preprocesses incoming data (batch or streaming). This stage handles transformations, feature engineering, and ensures data is formatted for each model type.
- **Model Suite:** A collection of specialized models (see next section) including supervised classifiers, anomaly detectors (depeg sentinels), and time-series forecasters. Each model outputs its prediction or anomaly score.
- **Ensemble Router & Governor:** A decision system (powered by RL) that dynamically routes inputs to the optimal model or weights model outputs to form an ensemble prediction. The *Router* is the mechanism of choosing/combining models, while the *Governor* applies policies or thresholds (e.g., suppressing an overly uncertain prediction or enforcing business rules).
- **Overseer LLM (Meta-Controller):** A large language model component that monitors the system's outputs and state. It can interpret complex scenarios (e.g. multiple anomaly alerts across models) and apply higher-level policies or explain results in human-readable form. It may also serve as a fallback reasoning engine to suggest actions when the standard models are uncertain.
- **Orchestration Layer:** Built with n8n (a low-code workflow automation tool), this layer sequences tasks such as data fetching, model invocation, conditional logic, notifications, and retraining triggers. The orchestration ensures that when an event occurs (e.g., new data arrives or an anomaly is detected), the appropriate sequence of actions (run models, apply router logic, send alerts, etc.) is executed in order ① ② .
- **FastAPI Gateway:** A lightweight API service that serves as the entry point for external requests. Clients (internal or external) hit the FastAPI endpoints (e.g. `/predict`, `/forecast`, `/analyze`) and the gateway triggers the orchestrated workflow and returns results. FastAPI was chosen for its high performance and easy integration, offering intuitive structure and automatic documentation for testing endpoints ③ .
- **Monitoring & Alerting:** A Prometheus-based monitoring subsystem collects metrics from models (e.g. error rates, drift metrics) and the infrastructure (latency, throughput). Grafana dashboards display these metrics and alert rules trigger notifications if Service Level Objectives (SLOs) are violated or if data drift or anomalies are detected beyond thresholds (details in a later section).

This architecture ensures **modularity** (each model/component can be updated independently), **adaptability** (the RL router learns optimal combinations over time), and **observability** (extensive monitoring of each part). In the next sections, we drill down into each major part of this architecture.

## Full Model Lineup

Gateway ML Studio comes with a comprehensive lineup of models, each geared toward specific types of insights. These models are grouped into categories: core supervised learners for classification/regression tasks, anomaly detection models (the “Depeg Sentinel” suite) for outlier detection, forecasting models for time-series predictions, and meta-models for ensemble scoring and severity assessment. Below we describe each category and the included algorithms.

## Core Supervised Models

The core supervised models provide predictions for well-defined outcomes (e.g., classification of an event or regression for a KPI). We include a mix from simple linear to advanced ensemble learners:

- **Logistic Regression:** A linear model for binary or multi-class classification that estimates odds via a sigmoid function. It's fast, interpretable, and often serves as a baseline. Logistic regression is effective for linearly separable data and provides probability outputs; for example, it can flag the probability of a transaction being fraudulent <sup>4</sup>.
- **Random Forest:** An ensemble of decision trees that improves generalization by averaging many tree outputs. Each tree is trained on a random subset of data/features, reducing overfitting and exploiting different aspects of the data. Random Forests excel at handling non-linear patterns and are robust to noise.
- **XGBoost (Extreme Gradient Boosting):** A high-performance gradient boosting framework that builds trees in sequence, each correcting errors of the previous. XGBoost includes regularization and efficient handling of missing data, often achieving state-of-the-art results in structured data competitions. It's included both as a core model and later as a meta-ensemble model due to its versatility. Notably, XGBoost can also be used for forecasting tasks by treating time features appropriately <sup>5</sup> (for example, it's used in some demand forecasting systems alongside ARIMA/ETS).

These core models are typically trained on labeled historical data (supervised learning). Logistic regression offers simplicity and interpretability, Random Forest offers a balance of accuracy and robustness, and XGBoost offers cutting-edge accuracy for complex patterns. During operation, the Router may choose one of these models for a given input if it has learned that one model outperforms the others in that scenario (for instance, XGBoost might be favored for complex nonlinear data, whereas logistic regression might be sufficient for simpler patterns).

## Anomaly/Depeg Sentinel Models

For anomaly detection – especially critical in scenarios like monitoring a stablecoin **peg** (where a deviation or “depeg” event must be caught) – the studio provides several unsupervised models that detect outliers or shifts without labeled anomalies. These “Depeg Sentinels” include:

- **Isolation Forest (IF):** A tree-based anomaly detector that explicitly isolates anomalies by random partitioning. The idea is that anomalies, being few and different, are isolated in fewer splits than normal points <sup>6</sup>. The Isolation Forest assigns an *anomaly score* based on the average path length to isolate a point – shorter paths (point isolated quickly) indicate more anomalous points. This method works well in high-dimensional data and makes no distribution assumptions <sup>7</sup>.
- **Local Outlier Factor (LOF):** A density-based detector that compares the local density of a point to that of its neighbors. If a point's density is significantly lower than its neighbors, it's considered an outlier <sup>8</sup>. LOF is effective for detecting context-dependent anomalies (points that are isolated with respect to their local region) and can handle complex data distributions, though it may struggle in very high dimensions.
- **One-Class SVM:** A support vector machine variant for novelty detection. It learns a decision boundary (often a hypersphere) around normal data such that most training points lie within it, and anomalies (novelties) lie outside <sup>9</sup>. One-Class SVM is useful when only “normal” data is available for training – it attempts to carve out the region of normality and flags anything elsewhere as an anomaly.

- **CUSUM (Cumulative Sum Control Chart):** A statistical method for change detection in time series. CUSUM charts accumulate deviations from a target mean and signal an anomaly when the cumulative sum exceeds a threshold. In practice, CUSUM is great for detecting small persistent shifts in a metric's mean or variance <sup>10</sup>. For example, it can detect a subtle depeg of a currency by picking up a series of small deviations that accumulate into a significant shift.
- **LSTM Autoencoder:** A neural network-based approach where an LSTM (Long Short-Term Memory) encoder-decoder network is trained to reconstruct time-series sequences. The model is trained on normal behavior; when a new sequence that contains anomalies is fed in, the autoencoder reconstructs it poorly, yielding a high reconstruction error. This error serves as the anomaly score <sup>11</sup>. LSTM autoencoders can capture complex temporal patterns and are especially useful for detecting anomalies in sequences with seasonality or long-range dependencies that simpler models might miss.

Each of these sentinel models outputs either a binary flag (anomaly or not) or, more usefully, a **score** indicating degree of anomaly or novelty. In Gateway ML Studio, these scores are standardized into a *Novelty Score*. The Novelty Score is a unified metric of how far a data point deviates from expected patterns, as assessed by one or more detectors. For instance, Isolation Forest provides an anomaly score based on isolation depth, LOF provides a local density ratio, etc., and these can be combined or normalized. In general, *novelty detection* refers to identifying previously unseen, abnormal data points as different from the "normal" training data <sup>12</sup>. The Novelty Score in our system encapsulates this, giving downstream components (like the Router or meta-classifier) a single metric to gauge how "strange" an input is.

## Forecasting Models

Time-series forecasting is a key capability for predictive analytics (e.g., forecasting prices, demand, or system load). Gateway ML Studio includes a mix of classic statistical models and modern approaches:

- **ARIMA:** AutoRegressive Integrated Moving Average, a classic time-series model that uses its own past values and past errors to predict future values <sup>13</sup>. ARIMA assumes some level of stationarity (which can be achieved via the Integrated (d) term that differences the series) and models the autocorrelation (AR term) and moving average of errors (MA term). ARIMA is powerful for many seasonal or trend-free series and often serves as a baseline for comparison. It requires careful parameter tuning (p, d, q) or the use of auto-ARIMA algorithms that test different combinations. (*In Gateway ML Studio, ARIMA might be applied for short-term forecasts or when a quick classical approach is needed. It's noted that ARIMA works best on stationary data* <sup>14</sup>.)
- **ETS (Exponential Smoothing, a.k.a. Error-Trend-Seasonality models):** A family of forecasting models (including Holt-Winters) that explicitly model level (average), trend, and seasonality with smoothing factors. ETS models excel when data shows clear trend or seasonal patterns and when recent observations should be given more weight <sup>5</sup>. For example, if a metric has linear growth and annual seasonality, an ETS model can capture that structure. These models are relatively fast and can provide prediction intervals. In Gateway ML Studio, ETS methods are valuable for business or operational metrics that have repeatable patterns over time.
- **Prophet:** A forecasting tool developed by Facebook (Meta) that fits an additive model with components for trend, seasonality, and holidays <sup>15</sup>. Prophet is designed to be intuitive (for analysts) and to handle common business time-series (which often have multiple seasonalities, holiday effects, etc.) with minimal tuning. It automatically detects changepoints (changes in trend) and can be robust to outliers and missing data. Prophet works well on complex real-world series where an analyst might otherwise manually add trend/seasonal terms <sup>5</sup>. In our stack, Prophet might be

used for, say, forecasting a cryptocurrency's price trajectory where weekly and yearly cycles and event-driven spikes all matter.

- **LSTM (Long Short-Term Memory) network for forecasting:** A deep learning approach using LSTM, which is a type of recurrent neural network adept at sequence modeling. LSTMs can learn arbitrary long-range dependencies and are not constrained to a specific time-series structure (unlike ARIMA or ETS)<sup>16</sup>. They can incorporate multiple features (exogenous variables) easily. The downside is they require more data and computational power, and can be harder to interpret. In Gateway ML Studio, an LSTM-based forecaster might be employed when the time series is complex, nonlinear, and rich in patterns that simpler models cannot capture (for example, if the data has nonlinear reactions to external variables or very long seasonal periods). We often pair LSTM forecasts with confidence estimation or anomaly detection on forecast errors to monitor performance.

Each forecasting model can produce point forecasts and possibly prediction intervals. The ensemble Router (discussed later) might choose to combine forecasts (e.g., by averaging ARIMA and Prophet for a robust forecast, or using an LSTM for short-term and ARIMA for long-term). The availability of multiple methods also allows a "best model selection" approach, where the system could pick the model with the historically lowest error for the current conditions, similar to a best-fit algorithm<sup>17</sup>. In practice, our platform could even run several forecasters in parallel and then use the Router's policy (or a simple average) as the final forecast to hedge against any single model's failure.

## Meta and Severity Scoring

On top of the base models, Gateway ML Studio uses **meta-models** to consolidate and interpret the outputs:

- **XGBoost Meta-Classifier:** In addition to being a core model, XGBoost is also used as a meta-learning algorithm that takes inputs from other models' outputs. For example, consider a scenario where we have predictions or scores from several models (say, logistic regression, an Isolation Forest anomaly score, a Prophet forecast error, etc.) – we can feed these features into an XGBoost classifier that was trained to predict an outcome of interest (such as "will this situation require an alert?"). This approach is essentially **stacking**, an ensemble technique where a "level-1" meta-learner combines the predictions of base learners<sup>18</sup>. The XGBoost meta-classifier can learn which model's output to trust more under which conditions. By leveraging the strengths of different algorithms and feature outputs, a stacking ensemble can improve performance over individual models<sup>19</sup>. In Gateway ML Studio, the meta-classifier might, for instance, output a classification of overall system state (Normal / Anomaly / Critical) using inputs like: classification model prediction, anomaly scores, forecast deviations, etc.
- **Novelty Score & Severity Index:** As mentioned, each anomaly detector yields a novelty/anomaly score. Gateway ML Studio can aggregate these into a single **Severity Index** that indicates how severe or unusual an event is. For example, if multiple anomaly detectors all flag an observation as extreme (each with a high score), the combined severity would be very high. Alternatively, if one mild anomaly is detected by one model but others see nothing, the system might treat it as low severity. The Severity Index can also incorporate contextual factors (for example, an anomaly during a peak traffic period might be rated differently than the same anomaly in a quiet period). This index is used to decide when to trigger alerts or when to engage the *Governor* logic to perhaps override model outputs (e.g., "if severity is extremely high, escalate to fail-safe mode"). **Novelty detection** techniques ensure that the system is aware of when it's encountering data that is fundamentally different from its training distribution<sup>20</sup>. By quantifying that as a score, we give the ensemble a sense of its own uncertainty or the need for human review.

In summary, the meta layer of models in Gateway ML Studio serves to **combine**, **interpret**, and **score** the raw outputs of the core models. This layer is crucial for translating raw model outputs into decisions or alerts that stakeholders care about. A single extreme anomaly score might not be meaningful by itself, but a meta-classifier can consider *which* anomaly detector fired and in what context, and then decide if this indeed signals a critical event (for instance, a depeg warning) or a false alarm. The use of XGBoost as a meta-learner provides a powerful, nonlinear aggregator of signals, which is appropriate given its proven ability to integrate diverse features in ensemble learning.

## Ensemble Routing and Governance

Perhaps the most distinguishing feature of Gateway ML Studio is its **Ensemble & Routing system**, which is governed by *reinforcement learning* and augmented with policy logic. The ensemble routing system determines *which model(s)* should be applied to a given input or how their outputs should be combined. This goes beyond static ensemble methods by using an RL agent (the Router/Governor) that learns optimal routing decisions over time.

Key aspects of this subsystem:

- **Dynamic Ensemble Selection (Router):** Instead of a one-size-fits-all ensemble (like a fixed weighted average of models), the Router can select different models or combinations depending on the current input and context. This approach treats model selection as a sequential decision problem that can be learned. We use **reinforcement learning** to train the router: it observes state (features of the input, recent performance of models, etc.), takes an action (choose model A, or blend A and B, etc.), and receives a reward based on the outcome (e.g., accuracy of prediction, or detection of a true anomaly). Over time, the router improves its policy to maximize long-term reward. Research in dynamic ensemble selection supports this approach: an RL agent can adaptively choose which predictor to trust for each instance, yielding better performance in changing environments <sup>21</sup>. In our case, for example, if the data regime shifts, the router might learn to rely more on a model that handles concept drift (like an LSTM) and less on a static model (like ARIMA) – adjusting on the fly.
- **Governor – Policy and Constraints:** While the RL router learns from data, we also incorporate a rule-based *Governor* that can enforce safety and policy constraints. These are like guardrails. For instance, if an ensemble prediction deviates wildly from known physical limits, the Governor can catch that (this could be implemented as a simple **Math/Policy SLM** – a Scoring Logic Module – that says “if output > X, cap it”). The mention of **SQL/Math/Policy SLMs** refers to simple *rule-based modules* integrated into the ensemble. A **SQL SLM** might run a database query or threshold check (e.g., checking if an account ID is in a blacklist stored in SQL and flagging the prediction if so). A **Math SLM** could be a formula-based check (like rate-of-change limit, or a formula that combines certain input fields to detect inconsistency). A **Policy SLM** could encode business rules (e.g., “if system load is above 90% and anomaly score is high, route to safe mode model”). These SLMs operate alongside ML models to provide deterministic signals. Hybrid anomaly detection systems often combine rule-based thresholds with ML models to catch obvious issues while ML catches subtler patterns <sup>22</sup>. In Gateway ML Studio, such modules ensure that obvious conditions trigger immediate known responses (for example, a rule might directly trigger an alert if a stablecoin price deviates >10% from peg, regardless of what ML models say, because that’s a critical depeg scenario).
- **Overseer LLM (Large Language Model):** The Overseer LLM is a unique component acting as a meta-controller and interpreter. It can be thought of as an AI agent that has an overview of the system’s state and can provide reasoning or adjustments in a flexible manner. Concretely, the LLM

might take as input a summary of recent model outputs, alerts, and context (maybe in a prompt form) and output a suggested action or analysis. For instance, if multiple models disagree, the LLM could analyze *why* (perhaps by correlating with external data or its embedded knowledge) and suggest which model is likely correct, or generate a natural language explanation for developers ("Model A flagged anomaly due to X, but Model B predicts normal. Given historical patterns, this might be a false positive."). The Overseer LLM thus acts as both a **judge** and a **translator**: it can judge ensemble outputs (a concept akin to *ensemble verification* using diverse "agents" for reliability<sup>23</sup>) and can translate the numeric outputs into insights or human-friendly reports. It might also inject external knowledge or perform on-the-fly checks that static models cannot (for example, querying a knowledge base via its prompt). The LLM's suggestions aren't taken as final blindly (since LLMs can hallucinate), but they inform the Router/Governor. One could imagine the LLM generating a policy update in natural language that the system parses – for now, however, it primarily provides analyses and can be reviewed by human operators.

- **Meta-Controller:** We often refer to the combination of the Router, Governor, and Overseer LLM together as the *meta-controller* layer. It's the brain that decides how models work together. The Meta-Controller's goal is to maximize overall system performance (accuracy, timeliness) while adhering to reliability constraints. It learns from both data (RL rewards) and possibly from human feedback (if stakeholders flag an alert as false positive, that could be fed back as a negative reward or via the LLM analysis). Over time, this should make the system **self-optimizing** – improving its routing policy to handle new conditions. For example, if a new type of anomaly occurs (unseen in training), initially the meta-controller might be uncertain; but once it's confirmed (perhaps by an operator or a triggered retraining), the meta-policy can adjust (the next time, it might route that scenario to a different model or trigger an ensemble vote).

To illustrate, imagine an input comes in: the supervised model says "no issue" with moderate confidence, but an Isolation Forest says "anomaly score very high" and Prophet indicates an unexpected spike in its forecast error. The Router/Governor could decide this combination is concerning and either output "anomaly" or at least escalate for review. The RL agent within it has learned from past events that whenever the anomaly detectors all agree on a high score, it should trust them even if the classifier didn't flag anything. Conversely, if only one detector is noisy but others and the classifier are fine, it might decide to ignore the single outlier (reducing false alarms). This kind of nuanced decision-making is exactly what the RL-governed ensemble is for – *contextual ensemble decisions* rather than fixed rules. Studies have shown that such dynamic weighting or selection via RL can significantly improve ensemble predictions in non-stationary environments<sup>24</sup>, as the agent continuously optimizes model weights or selections based on feedback.

In summary, the Ensemble & Routing system in Gateway ML Studio is a **multi-tiered decision engine**. It brings together straightforward rule-based logic and advanced learning to ensure that the right model (or models) are applied at the right time. This self-optimizing loop (with RL and oversight) is what allows the platform to maintain performance as conditions change, making it more resilient than static pipelines. Stakeholders benefit by getting a system that is both **accurate** and **reliable**, with fewer false negatives (missed events) and false positives (spurious alerts) thanks to this intelligent routing.

## Data Pipeline, Training & Evaluation, and Deployment Strategy

An effective ML system requires not just models, but also a robust **data pipeline** and systematic processes for training, evaluating, and deploying those models. This section covers how data flows into Gateway ML

Studio, how models are trained and evaluated (both initially and iteratively), and the strategy for deploying models into the production environment.

#### **Data Ingestion and Preprocessing Pipeline:**

Gateway ML Studio's data pipeline can handle both batch and streaming data. In a typical setup, raw data (for example, transactions, sensor readings, or market data) is first collected from source systems (databases, message queues, APIs). An ingestion layer (which could be a scheduled job or real-time stream) feeds data into the **preprocessing workflow**. Using n8n workflows, the pipeline may include nodes for data validation (dropping or fixing malformed entries), feature extraction (e.g., computing moving averages, encoding categorical variables, scaling features), and possibly enrichment (joining with reference data or adding engineered features like "time since last event"). The preprocessing ensures consistency: all the models in the lineup expect data in certain formats with specific features. This pipeline is configured through a combination of environment settings and n8n workflow definitions, so it can be adjusted without changing code – e.g., adding a new feature computation is as simple as adding a node in the flow.

Important considerations in the pipeline include handling of missing data (for time-series, we might forward-fill or interpolate; for tabular, we might use model-specific defaulting) and handling concept drift in features (the pipeline can compute summary statistics and feed them to the monitoring system to watch for drift in input distributions).

#### **Training and Evaluation:**

Each model in Gateway ML Studio has its own training process: - Supervised models (Logistic Regression, Random Forest, XGBoost meta-classifier, etc.) are trained on labeled datasets. We use standard practices: split data into training/validation sets, use cross-validation as needed, and tune hyperparameters (Grid search or Bayesian optimization can be integrated into the n8n workflow or done offline). Models like XGBoost and Random Forest can output feature importance, which we log for interpretability. - Anomaly detectors are typically unsupervised. They are trained on a sample of "normal" data. For instance, the Isolation Forest or One-Class SVM might be trained on a dataset assumed to contain mostly normal instances (any anomalies in training will dilute performance, so sometimes manual curation is needed). We evaluate anomaly detectors by their detection rates on known events if available, or via synthetic anomalies injected for testing. Metrics include precision, recall on anomalies, or simply visualizing their score distributions. The LSTM Autoencoder is trained to minimize reconstruction error on normal sequences <sup>25</sup>. Evaluation can involve checking that known anomalies produce high error. - Forecasting models are trained on historical time-series. We set aside recent data as a test set to evaluate forecast accuracy (metrics like MAE, RMSE, MAPE). For ARIMA/ETS, we often use automated parameter selection (e.g., auto-ARIMA) and rolling forecasts for evaluation. Prophet can fit automatically, but we still validate its performance on hold-out periods. LSTM models are trickier – we often train them with a sliding window approach and may need to retrain periodically as new data comes in. We monitor forecast bias and error over time.

A critical part of training is that it's **continuous or periodic**. Gateway ML Studio supports scheduled re-training or fine-tuning. Using n8n orchestration, we can schedule nightly or weekly jobs to retrain models on the latest data (especially the RL router which might need to retrain as it collects new experience). Additionally, if drift is detected by the monitoring system, it can trigger an on-demand retraining workflow. For example, if data drift for a feature crosses a threshold (via Prometheus alert), an n8n workflow could initiate retraining of the affected model (say, the Random Forest) and deploy the new model version, all automated (with notifications to the team).

Evaluation metrics are logged into the system. We store the performance of each model (perhaps in a Model Registry database) including training date, dataset used, and metrics. This ties into the deployment strategy: only models that meet certain performance criteria are promoted to production.

### **Deployment Strategy:**

Models are containerized or saved as artifacts to be loaded by the FastAPI service. There are a few deployment patterns: - **In-process deployment:** For fast models with small memory footprint (like logistic regression or small trees), we might load them directly into the FastAPI app process. The FastAPI endpoint then simply runs the model prediction in memory (this is feasible for scikit-learn models or even XGBoost with moderate size). - **Microservice deployment:** For heavier models (like a deep LSTM or the LLM or if using GPUs), we encapsulate them in separate microservices. For instance, an LSTM model might run in a TensorFlow Serving or a PyTorch backend container. The FastAPI gateway would call this service (possibly via HTTP or gRPC) when that model is needed. Similarly, the LLM overseer might be an API call to a model server or an external API (if using a large cloud-hosted model). - **On-demand ensemble invocation:** The orchestrator (n8n) can also handle invoking multiple models and assembling results. For example, an incoming request triggers the orchestrator: it calls the necessary model APIs (in parallel, potentially), collects all results, feeds them to the Router which could be a function or service, and then returns a consolidated result. This approach treats each model as a function that can scale separately.

The **model registry** holds versioned models. We use a consistent approach: when a new model is trained and evaluated, if it outperforms the current one, it gets a new version tag. The deployment can then either automatically pick up the new version or require a manual approval (depending on the criticality – e.g., for an anomaly detector, we might auto-deploy; for a financial risk model, perhaps a human must approve).

Continuous Integration/Continuous Deployment (CI/CD) pipelines are set up (for example, using GitHub Actions or Jenkins) to automate testing and deployment of the infrastructure and models. Infrastructure-as-code scripts (Dockerfiles, Kubernetes manifests if in K8s, etc.) ensure that deploying Gateway ML Studio in a new environment (staging/production) is reproducible.

For retraining the RL Router (Governor), the system logs the outcomes of decisions (reward signals). Periodically, an offline training can run (this could use algorithms like DQN or policy gradient methods on the accumulated experience). Once a new policy is derived, it can update the Router's parameters. Alternatively, we might use an online learning approach (continually updating the policy), but careful—stability and exploration need to be managed (some form of epsilon-greedy or softmax exploration in decisions). In practice, a safer approach is batched updates: let the router run with the current policy for a day, collect data, then retrain and redeploy policy if it's better.

### **Validation and Testing:**

Before deployment, we perform rigorous testing: - Unit tests for each model (ensuring that given known inputs, expected outputs or behaviors occur). - Integration tests where the entire pipeline is run on a sample input (perhaps simulating a full scenario: a known anomaly event flows through and we verify the system raises an alert). - Performance tests to ensure latency is within SLO (especially important when multiple models are called; we test concurrency, e.g., sending N requests at once to see if any bottleneck). - Fallback testing: we simulate a model failure (like disable one model) to ensure the Router can handle missing inputs (the Governor might have a policy like "if primary model fails, use secondary").

The deployment is usually containerized. We provide a **Docker Compose** setup for local deployment (with containers for the FastAPI gateway, the n8n orchestrator, a PostgreSQL or other DB if needed for storage, and a Prometheus/Grafana container for monitoring). In production, these could be deployed on Kubernetes (with each component as a deployment and perhaps using Helm charts to configure). The FastAPI and n8n containers scale horizontally if needed (they are stateless, aside from any cache or loaded models – for stateful things like the RL agent's memory, we either persist it or use external storage).

Finally, once deployed, the system enters the monitoring phase, which we describe next. Summing up this section: Gateway ML Studio's data pipeline and training/deployment processes are designed for **automation and reliability** – new data is continuously folded in, models are retrained and evaluated regularly, and deployment is made smooth through containerization and orchestration. This ensures the system remains up-to-date and performance doesn't degrade as data evolves.

## Alerts, Drift Detection, and SLO Monitoring with Prometheus

Operating a complex ML system in production demands continuous monitoring to ensure models perform as expected and to catch issues like data drift or component failures. Gateway ML Studio incorporates a comprehensive monitoring framework using **Prometheus** for metric collection and **Grafana** for visualization/alerting. Additionally, Service Level Objectives (SLOs) are defined to keep the system's performance (both ML metrics and operational metrics) within acceptable bounds. This section details how we handle alerts, drift detection, and SLO monitoring.

### Prometheus Metrics Collection:

Each component of the system (data pipeline, models, router, etc.) exposes metrics to Prometheus. We follow a metrics exposition approach wherein the FastAPI gateway and other services have endpoints (e.g., `/metrics`) that Prometheus scrapes periodically. Some of the key metrics include:

- **Model metrics:** For supervised models, we log the prediction confidence distribution, number of predictions made, and if possible, online accuracy (when true outcomes become known later). For anomaly detectors, we log the rate of anomalies flagged per time window. For forecasters, we log forecast error on recent data (where actuals are available).
- **Data distribution metrics:** We capture statistics of input features (mean, std, min, max, perhaps histograms) in production. Tools like Boxkite can automate this by creating histograms of feature values for both training and production data <sup>26</sup>. These histograms are exposed to Prometheus, allowing us to detect data drift by comparing distributions over time. For example, if a feature's distribution in production diverges significantly from training (quantified via a metric like KL divergence or population stability index), an alert can be triggered.
- **System metrics:** These include standard service metrics like response latency (to ensure real-time requirements are met), request throughput, error rates (e.g., if a model service fails to respond or an exception occurs, we count it), and resource usage (CPU, memory of each service). Prometheus can scrape system exporters or Kubernetes metrics for this.

By leveraging Prometheus' powerful querying, we can combine metrics for sophisticated monitoring. For instance, we might have a PromQL query that computes the difference between training and live distribution for a given feature and triggers if it exceeds a threshold (Boxkite actually suggests using KL divergence or K-S test metrics in PromQL with thresholds for drift alerts <sup>27</sup>).

### Data Drift Detection:

Data drift (input drift) and concept drift (output drift or model performance drift) are key concerns. Our approach:

- **Feature Drift:** As mentioned, we track feature distributions. Using the histograms from training and production, we compute divergence metrics. For categorical features, **KL divergence** or chi-square can measure drift <sup>27</sup>. For continuous features, a **Kolmogorov-Smirnov (K-S) test** or Wasserstein distance can be used. These are computed either offline and pushed as metrics, or directly via Prometheus queries if histograms are available (e.g., one could write a PromQL to calculate the probability mass difference). When drift exceeds a set threshold, Prometheus fires an alert. In Grafana, we have dashboards showing side-by-side histograms of training vs. current data (Boxxite's Grafana dashboard plots production vs training distribution for features over time <sup>28</sup>, allowing a visual check for drift).
- **Model Performance Drift:** If we have ground truth coming in (for instance, in a fraud detection scenario, later you find out which transactions were fraudulent), we can compute model accuracy or precision in production over a sliding window. If accuracy falls below a certain SLO (say, below 90% for a day), that's concept drift or model degradation. We set an alert on such a metric. In cases where ground truth is not immediately available, we use proxy metrics: e.g., anomaly detectors might be unsupervised, so instead we monitor the rate of anomalies. If an anomaly detector suddenly flags 5x more points than usual, either the world changed (something bad is happening) or the model's behavior changed (maybe it's seeing out-of-training-distribution data). Either way, that triggers investigation.
- **Feedback Loop:** The system can also incorporate human feedback. If users of the system (analysts, operators) mark certain alerts as false positives or flag missed events, those can be logged as metrics too (like "false\_positive\_count"). Monitoring those can help identify when a model starts giving too many false alerts, prompting retraining or threshold adjustments.

### Alerts and Notifications:

Prometheus Alertmanager is configured with various alerting rules, tied to email, Slack, or other channels for notifications. Some key alerts:

- **Anomaly Alert:** If the Severity Index (from the ensemble) goes above a critical value, that might trigger an immediate alert to stakeholders (e.g., an email/SMS for on-call engineers, or a message to a Slack channel). This is a core functionality: the whole point of the anomaly detection pipeline is to alert when something is off (like a potential depeg event). The alert includes details like which models fired, the values, etc., often formatted via templates in Alertmanager.
- **Data Pipeline Alert:** If data hasn't arrived (no new data in X minutes when expected) or if the pipeline fails (n8n can send a metric or Alertmanager can watch for workflow failures), an alert is sent to the engineering team to investigate upstream issues.
- **Latency SLO Violation:** For example, if the prediction API latency exceeds, say, 500 ms for more than 5% of requests in the last 5 minutes (this SLO depends on requirements), fire an alert. This could indicate a model is running slow (maybe the LLM or LSTM is dragging) or infrastructure issues.
- **Error Rate:** If any service (FastAPI, model microservice) error rate > 1% (HTTP 5xx responses) in the last 5 minutes, alert. We want to catch model serving errors quickly.
- **Drift Alert:** As described, if feature drift metric > threshold or if model accuracy (or other performance indicator) drops, send an alert. In some setups, this might be an *FYI* alert to data scientists to retrain models, or it could automatically trigger the retraining pipeline (with a notification that it did so).

### Service Level Objectives (SLOs):

We define SLOs for both **service performance** and **model quality**. Examples include:

- **Availability SLO:** The prediction API should be up 99.9% of the time (leading to certain allowed downtime per month). This is

tracked by uptime probes and error rates.

- *Latency SLO*: As mentioned, say 95% of predictions should be served under 300ms. We track histogram of response times and compute the 95th percentile. Grafana can show an SLO dashboard and Alertmanager can notify if the budget is being eaten into (e.g., if over the last hour, 95th percentile is 500ms, we're violating).

- *Detection SLO*: For the ML side, perhaps: "System should detect and alert on any confirmed depeg event within 1 minute, with at most 1 false alarm per week." Tracking this is more complex; it may involve tagging real incidents and measuring detection time and count of false alarms. While hard to fully automate, the monitoring can track proxy metrics: number of alerts, and known incidents missed. If false alerts exceed threshold, that triggers a review of model thresholds; if something known was missed (maybe via a manual log entry), that's a serious event to investigate.

- *Drift SLO*: We may set an objective like "Model data drift (by KL divergence) should remain below X; if it exceeds, model retraining should occur within Y hours." The monitoring can ensure the first part (drift below X). The second part, retraining, could be partially automated as discussed.

Using **Grafana**, we create a **Monitoring Dashboard** for Gateway ML Studio that includes panels for: - Real-time predictions per minute, anomaly count per minute (with anomalies highlighted). - Feature drift indicators (like a gauge for each major feature's drift level). - Model metrics like latest accuracy or average anomaly score. - System metrics like CPU, memory, response times. - A summary of alerts and their status.

Grafana's alerting (in newer versions) or Prometheus Alertmanager handles sending notifications. Alerts are categorized by severity (warning vs critical). For example, a mild drift might just create a warning (so data scientists can schedule retraining), whereas a severe anomaly triggers a high-severity incident alert.

One of the benefits of using Prometheus/Grafana is that we can unify ML monitoring with traditional monitoring in one place <sup>29</sup>. Our monitoring stack treats model metrics just like application metrics. If the team already uses Grafana for microservices, the ML metrics can be integrated so that an ops engineer can see a timeline and notice "Ah, around 2:00 AM the data distribution shifted drastically (as shown in one panel) and at the same time latency spiked (another panel) and an anomaly alert was sent." This holistic view is crucial for troubleshooting complex issues that may span data and infrastructure.

#### **Logging and Tracing:**

Although the question focuses on Prometheus (metrics), it's worth noting we also implement logging (each prediction/alert is logged, possibly to an ELK stack or cloud logging service) and tracing (with tools like OpenTelemetry) for the pipeline. This helps in post-mortems (like investigating false positives/negatives or system slowdowns, by tracing the path of a request through n8n tasks and model calls). For each alert triggered, we might store a snapshot of the input data and model outputs that led to it, aiding debugging.

In summary, Gateway ML Studio's monitoring strategy uses **Prometheus for continuous metric surveillance** and **Grafana for visualization and alerting**. It covers *data drift*, *model performance*, and *system health*, ensuring that any divergence from expected behavior is quickly detected. By defining SLOs, we set clear targets for the system's reliability and accuracy, and the monitoring tools automatically enforce those by alerting when we're close to breaching. This proactive monitoring allows the team to **maintain trust** in the system's outputs – a critical requirement especially if stakeholders are making decisions based on Gateway ML Studio's alerts and predictions.

## n8n Orchestration Logic, Workflows, and FastAPI Gateway

Orchestration is the glue that holds together Gateway ML Studio's components, ensuring they operate in the correct sequence and under the right conditions. We leverage **n8n**, a popular low-code workflow automation tool, to build and manage these orchestration workflows. In tandem, the **FastAPI gateway** provides an interface for external users/systems to interact with the ML Studio. This section describes how n8n is used to create robust flows and how the FastAPI gateway is structured as a stub that ties into those flows.

### n8n Orchestration Workflows:

n8n acts as the **central conductor** of our platform, coordinating data flow and decisions across various services <sup>1</sup>. Think of n8n workflows as directed graphs of nodes, where each node can perform an action (HTTP request, data transformation, decision logic, etc.). We have designed multiple workflows for different purposes:

- **Prediction Workflow:** This is triggered whenever a prediction request comes in via the FastAPI endpoint (likely through a webhook or direct function call). The workflow steps might be:
  - **Receive Input:** Start node gets the input data (either passed from FastAPI or pulled from a queue).
  - **Preprocess:** Nodes perform any needed preprocessing (e.g., normalization) – though major preprocessing is often done prior, this ensures format is correct.
  - **Parallel Model Calls:** The workflow can branch into parallel HTTP Request nodes or Function nodes that call each required model. For example, one branch calls the classification API, another computes anomaly scores (perhaps by calling an internal function or service), another gets a forecast. n8n allows parallel branches, so this can happen concurrently to minimize latency.
  - **Combine/Route (IF/Switch logic):** After model outputs are collected, an IF node or Switch node can implement simple routing logic. For instance, an IF node can check “IF any anomaly score > threshold OR classifier output = fraud, then follow the anomaly branch, else follow normal branch.” <sup>2</sup>. This is a simple form of routing. For more complex decisions, we might instead send all results to a **Router function** (perhaps implemented as a custom node or an HTTP call to a Router microservice) which applies the RL policy and returns a decision.
- **Post-processing & Response:** Based on the route, the workflow may add contextual info. For example, if an alert is triggered, the workflow might call another service (like an email or Slack node) to send notification, and then format the API response to indicate an alert was raised. If no issues, it simply returns the prediction result. The end of the workflow sends the result back to the FastAPI handler (if using a webhook mechanism, n8n would send an HTTP response).
- **Scheduled Workflows:** n8n is also used for maintenance tasks. For instance, a workflow might run nightly at 2 AM to:
  - Pull the latest data from a database,
  - Retrain one or more models (by calling a training script or service),
  - Evaluate the model, and if performance is acceptable, push the new model artifact to the model registry and maybe trigger a deployment pipeline. This could include a human approval step if needed (n8n can pause for manual confirmation).
  - Send a summary report via email about the training (metrics, etc.).

- **Alert Handling Workflow:** When the system generates an alert (e.g., a high severity anomaly), aside from the Prometheus/Grafana alerts, we can have n8n handle certain automated remediation. For example, an n8n workflow could be triggered by a webhook from Alertmanager (Prometheus alerts can fire webhooks). The workflow might:
  - Enrich the alert: gather recent data around the event, maybe query a database for related records.
  - Create a ticket in an incident management system or notify specific people.
  - If the alert is of a certain type, execute predefined mitigations (for instance, if it's an anomaly in a certain system, call a REST API to scale up resources or switch off something).
- **Workflow for Data Drift:** Similarly, an automated workflow could trigger on a drift alert to, say, launch a retraining job or simply log the drift and notify the data science team.

One of the advantages of n8n is the **visual clarity** of these processes. Non-developers can open the n8n editor and see the “map” of how data flows and decisions are made, which fosters transparency <sup>1</sup>. For internal developers, it’s easy to modify the logic (like adjusting thresholds or adding new branches when new models are added) without diving deep into code – they can update the workflow visually and test it out.

We ensure that complex Python logic (like the RL routing policy) is either encapsulated in a single n8n Function node (where you can write custom JavaScript code) or, more cleanly, exposed as an HTTP endpoint that n8n calls. For example, the **Router + Governor** might be a Python function deployed in the FastAPI backend. The n8n workflow at decision stage simply calls “POST /routeDecision” with all model outputs, and gets back the decision.

Another workflow use-case is **FastAPI error handling**: If a request times out or fails mid-way, n8n can catch that and respond gracefully. For instance, in the workflow, if a model HTTP request node fails, we can have an error branch that perhaps retries a fallback model or returns a message “Service currently unavailable, please try again” (and triggers an alert behind the scenes). This prevents the entire API from crashing on single point failures.

#### **FastAPI Gateway Stub:**

The FastAPI service in Gateway ML Studio is relatively thin. Its responsibilities are to: - Expose well-defined API endpoints (RESTful HTTP endpoints) for clients. E.g., `POST /predict` for getting a prediction/anomaly assessment, `GET /health` for health checks, `POST /forecast` for a forecast query, etc. - Validate and parse incoming requests (using Pydantic models for data validation, ensuring required fields are present and correctly typed). - Pass the request data to the orchestration layer and wait for the result. - Format and return the response (in JSON), including appropriate HTTP status codes.

In some implementations, we directly embed the n8n trigger. For example, FastAPI could call an internal function that triggers an n8n workflow execution. However, n8n provides webhooks: one can create a workflow that starts with a **Webhook node** listening at a certain URL path. We can register a webhook like `/webhook/predict` which corresponds to an n8n workflow. In that case, the FastAPI might not even need to handle the logic – clients could call the n8n webhook URL directly. However, for security (auth, etc.) and unified documentation, we typically route through FastAPI.

So, the FastAPI “stub” might do something like: when `/predict` is hit, it internally performs an HTTP request to the n8n webhook endpoint (which could be local or separate) with the data, or calls an n8n SDK if available. Alternatively, if we converted n8n workflows to a Python function using tools like n8n2py, we

could even have the workflow steps executed in-process <sup>30</sup> (but that's advanced; likely we keep n8n running separately).

Why FastAPI? It's a **modern, high-performance web framework** with automatic docs (via Swagger/OpenAPI UI) which is very handy for both internal testing and external integration <sup>31</sup>. Developers can go to `/docs` and see all endpoints, their schemas, and even test calls. FastAPI also handles async concurrency well, which is useful since some model calls or orchestrations might be I/O bound. The gateway ensures that if we scale out to multiple instances (behind a load balancer), each can service requests statelessly.

We call it a "gateway stub" to indicate that it mostly just forwards to the real logic in workflows. However, it can contain lightweight logic if needed, like simple caching. For example, if a client requests a forecast for a common date range that was recently computed, FastAPI could cache that response for a short time to reduce load on the forecasting models.

#### **Authentication & Security:**

If Gateway ML Studio is used by external clients, the FastAPI gateway can enforce auth (API keys or OAuth tokens). This can be integrated into the dependency system of FastAPI (requiring a token and validating it). n8n workflows can also include auth logic, but it's cleaner to keep it at the gateway. We also validate payload sizes and types to avoid injection attacks or overload (since an unexpectedly large input might cause heavy processing in models).

#### **Example Flow (End-to-End):**

To tie it together, let's walk through a typical event – say a new data point comes for anomaly analysis: 1. An external system (or user) sends an HTTP request to Gateway's `/predict` endpoint with the data (for instance: features of a transaction to evaluate). 2. FastAPI receives it, authenticates the request, parses the JSON into a Pydantic model (which might define the fields like amount, account\_id, etc.). 3. FastAPI then triggers the n8n workflow. This could be a synchronous call to a webhook (FastAPI waits for the response from n8n). The n8n workflow executes as described: runs models, does IF logic, etc. behind the scenes. 4. n8n returns a result to FastAPI. Let's say it returns `{"anomaly": true, "severity": 8.7, "explanation": "Deviation in pattern X detected"}`. 5. FastAPI takes that and returns it as the HTTP response to the client. Additionally, n8n might have, in parallel, sent out an alert email due to severity or logged something. 6. The client gets the response in near real-time (assuming the workflow took maybe a few hundred milliseconds with parallel model execution). 7. Meanwhile, Prometheus metrics for this request are updated (FastAPI exposes latency, and maybe within n8n or model calls metrics were incremented).

If the client calls `/forecast` endpoint, perhaps that triggers a different workflow focusing on time series models. If they call an admin endpoint like `/retrain_model?model=xgboost`, we might secure that to internal use and have it kick off a retraining workflow.

One cool aspect is **extensibility**: internal developers can add a new model to the lineup, and it mostly involves updating an n8n workflow to call that model and possibly retraining the meta-classifier. No need to rewrite orchestration code from scratch – the low-code nature means faster iteration. This encourages external contributors to plug in their own models; as long as they expose an endpoint or write an n8n function node for it, it can be integrated.

### **Reliability of Orchestration:**

We run n8n in a highly-available manner (maybe multiple instances with a shared database for state). If n8n is down, the system's functionality is affected, so we monitor n8n itself. FastAPI has basic fallback – for example, if the orchestrator doesn't respond in time (timeout), FastAPI can return a safe error message or even attempt a simpler direct model prediction as a fallback if appropriate.

In conclusion, **n8n** provides a clear and flexible way to define the logic of how Gateway ML Studio processes data, while **FastAPI** offers a robust interface for clients to interact with that logic. This separation of concerns (workflow vs. API) allows us to use the best tools for each job: n8n's visual workflows make complex orchestration easier to manage and modify, and FastAPI's speed and developer-friendly features ensure that integration and scaling in a microservice environment are handled gracefully. Together, they orchestrate the *brains* (models and decisions) and the *communication* (requests and responses) of Gateway ML Studio.

## **Configuration, Environment Variables, and Secrets Management**

Effective configuration management is essential for deploying Gateway ML Studio in different environments (development, testing, production) without code changes. Equally important is secure handling of sensitive information like API keys or database credentials. In this section, we outline how configuration is organized, how environment variables are used following best practices, and how we manage secrets.

### **Configuration via Environment Variables:**

Gateway ML Studio follows the principle of **externalized configuration**. All configuration values that may change between deployments or contain sensitive data are kept out of the code and instead provided through environment variables. This is aligned with the Twelve-Factor App methodology which advocates storing config in the environment <sup>32</sup>. By doing so, we ensure that: - We can change settings (like model thresholds, logging levels, URLs of services) without modifying code – just by changing an env var and restarting the service. - There's no risk of accidentally committing secrets to version control, since they're not hardcoded. - The same codebase can be deployed to staging vs production with different settings simply by using different env var values.

**What configuration do we have?** Many aspects: - **Model Config:** e.g., threshold values for anomaly scores to consider something critical, or the frequency of retraining schedules. Instead of hardcoding "if `anomaly_score > 0.8 then alert`", we might have an env var `ANOMALY_ALERT_THRESHOLD=0.8` so it can be tuned. - **Service URLs and Ports:** The FastAPI service might need to know the URL of the n8n service, or the Prometheus pushgateway, etc. These are set via env (like `N8N_URL`, `PROMETHEUS_ENDPOINT`). For local dev, `N8N_URL` might be `http://localhost:5678`, while in production it could be an internal DNS name. - **Database Credentials:** If we use a database for logging or model registry, things like `DB_HOST`, `DB_USER`, `DB_PASS` are set as env vars (often injected via Docker secrets or k8s secrets). - **API Keys/Tokens:** Any external service integration (like if the overseer LLM calls an external API such as OpenAI, or if n8n sends Slack messages, etc.) – those keys are stored in env or in n8n's credentials manager (which itself likely uses env or a vault reference). - **Feature Flags:** We might have env flags to toggle components (e.g., `USE_OVERSEER_LLM=true/false` if we want to turn it off easily). Or which models to activate if some are experimental. - **Secrets Management:** For secrets (passwords, keys), environment variables are one approach but come with risk if not handled properly. In our deployment, we integrate with container orchestration secrets mechanisms. For example, in Docker Compose, we might use an `.env` file (which is not committed) or Docker secrets. In Kubernetes, we use Secrets objects mounted as env vars. We avoid

storing secrets in plaintext in any repo or image. The rule is: *the code can be made open-source without leaking secrets* <sup>33</sup> – meaning all secrets are injected at runtime, not build time.

We also logically separate config from code. The codebase has default config values for local development (which are safe defaults), but in a production deployment, everything is overridden by env.

#### **Configuration Files:**

In some cases, using environment variables for complex config can be unwieldy. We support config files (like a YAML or JSON) for things like model hyperparameters or ensemble configurations. For example, a `config.yml` might list the models and their weights or endpoints. We treat such files as config, not code – meaning they are not baked into the logic but loaded at startup. These config files can be supplied per environment (like a different one for prod vs test). To avoid the pitfalls of config files (risk of being checked in with secrets, etc.), we often store non-sensitive config in files and sensitive ones in env. Or we use environment variables to point to config files. For instance, an env var `MODEL_CONFIG_PATH=/etc/gateway/models.yml` can point to a mounted ConfigMap in Kubernetes.

#### **Example of Config Management:**

Let's say we have an orchestration setting like how many days of data to use for retraining. Instead of a literal number in the code, we have `RETRAIN_LOOKBACK_DAYS` env. The training workflow reads this value and uses it in a database query. If we find we need to change it, we change the environment variable (via our deployment scripts or platform) and restart – no code change needed.

#### **Secret Example:**

The Overseer LLM might use an API key. In code, we do something like `api_key = os.getenv("LLM_API_KEY")` and then use that for requests. In local dev, the developer might set that in a `.env` file loaded by Docker Compose. In production, we store the API key in a secure vault or secret store, and it gets injected as an env var when the container runs. *Important:* We ensure such env vars aren't accidentally logged. We also ensure to rotate them if needed (the system should be able to handle a secret change by restart or live update).

#### **n8n Credentials:**

n8n has a concept of credentials (for nodes like email, Slack, etc.). Those credentials are encrypted in n8n's database. We either set them via n8n's UI (which is then stored securely on disk using an encryption key from env), or we use env directly in function nodes. For example, an n8n HTTP node could use URL `https://api.slack.com/...?token={{ $env.SLACK_TOKEN }}` (if n8n supports env access). If not, we define the credential in n8n and ensure the n8n instance's encryption key (env `N8N_ENCRYPTION_KEY`) is set to a strong value in production. That way, even if someone dumps the DB, they can't decrypt creds without that key.

#### **Hierarchy and Defaults:**

We implement a hierarchy: default config in code (lowest priority), overridden by environment variables, and sometimes overridden further by command-line args if provided (for flexibility). But mostly, env is the main override. For example, FastAPI might read a config class that first loads defaults from a file (like `config_default.py`), then applies overrides from env. This allows quick tweaking.

#### **Example snippet (Pseudo-code):**

```

class Config:
    N8N_URL = os.getenv("N8N_URL", "http://localhost:5678")
    ANOMALY_ALERT_THRESHOLD = float(os.getenv("ANOMALY_ALERT_THRESHOLD", 0.8))
    LLM_API_KEY = os.getenv("LLM_API_KEY") # No default for secret

    # etc...

```

In production, `N8N_URL` will be set to the actual orchestrator address, `ANOMALY_ALERT_THRESHOLD` might be set differently if needed, and `LLM_API_KEY` will be set. In development, if not set, defaults are used or some safe behavior (like no LLM).

#### **Versioning and Tracking Config:**

One challenge is knowing what config was in place at a given time (for debugging). We address this by logging config values at startup (except secrets). For instance, the FastAPI on start logs “Config: `N8N_URL=x, THRESHOLD=y, ...`”. We also recommend using a git versioned `.env` template (no secrets, just example values) to document what env vars are used.

#### **Safety Measures:**

- We avoid duplicating config in multiple places – single source of truth. E.g., if both FastAPI and n8n need to know a threshold, ideally one is the main user of it. Or we set it in both via env from the same source.
- For secrets in env, we are aware of the risk that environment variables can be seen by process listing or dumped in crash logs. In container setups this risk is lower (process isolated), but still, for extremely sensitive data, consider using a secrets manager that the application can query at runtime. However, that adds complexity. For now, env vars (or k8s secrets) are acceptable practice in many scenarios <sup>32</sup>.
- We ensure that if config is missing or invalid, the system fails fast with a clear error. E.g., if `LLM_API_KEY` is required but not set, the app should throw on startup telling ops that config is incomplete, rather than later failing unpredictably.

#### **Example: Switching off a model via config:**

If we want to temporarily disable, say, the LSTM Autoencoder if it's causing issues, we could have `USE_LSTM_AUTOENCODER=false` in env. The orchestrator workflow would check this (maybe n8n can have an IF node gating that branch based on an env variable value or via a small script). Alternatively, the FastAPI could not call that model's endpoint if disabled. This is simpler than removing code and later adding it back – we just toggle a flag.

In conclusion, Gateway ML Studio treats configuration and secrets with the mantra: “*Config varies, code does not*” <sup>34</sup>. By externalizing environment-specific details to environment variables and secure stores, we achieve flexibility (the same container image can be deployed anywhere with different behavior) and security (no secrets in code repositories). This makes the platform portable across local dev, Docker, and cloud, which leads into the next section on setup instructions.

## **Setup Instructions for Local, Docker, and Cloud Deployment**

Gateway ML Studio is designed to be deployed in various environments, from a developer's local machine to Docker containers, to a scalable cloud setup (e.g., on Kubernetes). In this section, we provide detailed setup

instructions for each scenario, ensuring that both internal developers and external contributors can get the system running and understand the deployment process.

## Local Setup (Development Environment)

For local development (e.g., on a laptop or workstation), the goal is to run all components with minimal friction, possibly without full production-grade dependencies. We assume the developer has Python (for FastAPI and possibly model code) and Node.js (for n8n, if running via source) installed, or use Docker even for local (see next subsection).

**Steps:** 1. **Prerequisites:** Install Python 3.9+ and Node.js 14+ (for n8n) if not using Docker. Also, have Docker installed if you plan to run ancillary services like a local Prometheus/Grafana or if you prefer dockerizing each component. 2. **Clone Repository:** Get the Gateway ML Studio code from GitHub (the repository link will be provided in the final section). Navigate into the project directory. 3. **Environment Setup:** Copy the example environment file: `cp .env.example .env` and fill in any needed values. For local, defaults in `.env.example` might suffice (like using `localhost` for service endpoints, dummy API keys if needed for non-critical use). 4. **Install Dependencies:** For the FastAPI backend and any Python model code, run `pip install -r requirements.txt`. This will install FastAPI, scikit-learn, xgboost, prophet (maybe via fbprophet or cmdstanpy), etc., as listed. If some models (like Prophet or LSTM) need additional system libs (like Prophet requires pystan or cmdstan), refer to README for any extra steps (like installing `cmdstan` for Prophet). 5. **Start n8n:** We provide an n8n workflow export (JSON) in the repo. You can start n8n in dev mode: either `npx n8n` if globally installed or via Docker: `docker run -it --env N8N_EDITOR_BASE_URL=http://localhost:5678 -p 5678:5678 n8nio/n8n`. Once n8n is running (on port 5678 by default), you can import our workflow file through the n8n UI or CLI. The workflow will contain all the orchestration logic. Alternatively, if we provided scripts, run `npm install && npm run start` in an `orchestration/` directory if n8n is part of the repo. 6. **Configure n8n:** In n8n's UI (open `http://localhost:5678` in a browser), set up any necessary credentials (for example, if email nodes exist for alerts, configure an SMTP credential; if Slack nodes, add Slack webhook URL, etc.). Also, ensure the webhooks are accessible: for local dev, you might use n8n's test webhook URLs, which work when you execute the workflow manually via the UI. For integration with FastAPI, use production URLs (which are the same host). 7. **Start FastAPI:** The backend can be started with `uvicorn main:app --reload --env-file .env`. The `--reload` is convenient for dev (auto-reloads on code change). `--env-file .env` ensures your config is loaded. This will run FastAPI on `http://localhost:8000` by default. 8. **Run Support Services:** For local testing, you might not need Prometheus/Grafana if you just want core functionality. But if desired, we have a `docker-compose.monitoring.yml` that can start Prometheus and Grafana. Simply run `docker-compose -f docker-compose.monitoring.yml up -d`. This will start Prometheus (scraping metrics from FastAPI and possibly n8n) and Grafana (accessible at `http://localhost:3000`). 9. **Test the System:** - Open the FastAPI docs at `http://localhost:8000/docs` to see available endpoints. Try a `POST /predict` with sample data (we'll provide an example JSON in docs). You should get a response from the system. In the console, you should see logs from FastAPI as well as actions in n8n (which logs to its console when workflows run). - If using Grafana, log in (default admin/admin for local) and import the provided dashboard JSON (in the repo's monitoring folder) to see metrics. - Check that if you intentionally send an anomaly (if you know a data point that should trigger, or simply manipulate input to be extreme), the system does route accordingly (e.g., triggers an alert workflow or returns an anomaly result). 10. **Iterate:** In dev mode, you can tweak the n8n workflow live via the UI, or edit FastAPI code and see changes (thanks to `--reload`). Most of the heavy models can be mocked or simplified for local dev to speed up (e.g., use smaller trees or fewer data).

For any C extension issues (like if Prophet or others fail to install), refer to documentation; possibly use a virtual environment or conda where those are easier to manage. Alternatively, the Docker approach below can avoid local Python dependency hell by isolating in containers.

## Docker Deployment (Standalone)

We provide Docker images to simplify deployment and ensure consistency. A **Docker Compose** file (`docker-compose.yml`) is included to orchestrate multiple containers. This setup is useful for quickly spinning up the entire stack on a single machine or VM, and is also the basis for cloud deployments.

**Components in Docker-Compose:** - `gateway-api`: The FastAPI app, Dockerfile based on Python (e.g., `python:3.9-slim`). It copies the code, installs requirements, and runs Unicorn. The environment variables for config are passed here in the compose file (or an `.env` file Compose reads). - `n8n`: We use the official n8n image (`n8nio/n8n`). We mount the directory with our workflows JSON so that it can import it on start (or we use an init command to import). We also set environment like `N8N_BASIC_AUTH` (to secure the editor in prod) and `N8N_ENCRYPTION_KEY` for credentials. - `prometheus`: Uses a prometheus image with a config that scrapes the FastAPI (and possibly n8n) metrics endpoints. The config file is mounted. - `grafana`: Uses the official Grafana image. We can provision a dashboard and data source via mounted config so that on startup it knows about Prometheus and maybe loads our ML monitoring dashboard.

Optionally, containers for the LLM or others if they are separate. For simplicity, let's assume LLM is external API so none here; if it was local (like a local model server for a large model), that could be another container.

**Instructions:** 1. Ensure Docker and docker-compose are installed on your system. 2. Prepare a `.env` file for docker-compose with all necessary environment variables (some might be duplicated from dev `.env`). We might provide a `docker-compose.env.example`. 3. Run `docker-compose up -d` (this will build images if not built, e.g., for `gateway-api`). 4. Docker will pull images (for n8n, prometheus, etc.) and start containers. Use `docker-compose logs -f` to monitor startup. 5. Once up, you should have: - FastAPI at `http://localhost:8000` (or other port if mapped differently). - n8n possibly not exposed by default to host (we might expose it on a local-only port or require SSH tunnel if secure), but for testing you could expose `5678` to access the editor. - Grafana at `http://localhost:3000` (if exposed, otherwise maybe only internally). - Prometheus at `http://localhost:9090` if needed. 6. Use `docker-compose ps` to ensure all are running. If something quits (e.g., if n8n didn't find a volume), troubleshoot by reading logs. 7. Test as with local: call the API, see if responses come. The advantage here is everything is running in containers and closely replicates production environment, so issues like dependency versions or OS-level differences are ironed out.

Docker deployment is great for a single-node or small scale. If needed, one can increase replicas of the FastAPI service by modifying compose (though then you'd need a simple load balancer; in dev, one instance is enough).

When new versions of models or code come, you'd rebuild the `gateway-api` image and `docker-compose up` again (or use `docker-compose pull` if images are from a registry). We plan to publish pre-built images to a registry (like Docker Hub or GitHub Packages) for convenience.

## Cloud Deployment (Kubernetes or VM)

For a robust production deployment, Kubernetes (K8s) is a common choice. Alternatively, deploying on cloud VMs or Docker Swarm is possible. We'll outline a Kubernetes approach, which covers the main considerations:

**Kubernetes Deployment:** - We define a Kubernetes namespace (e.g., `gateway-ml-studio`) to contain all resources. - **FastAPI (Gateway API):** Deploy as a Deployment with N replicas (based on load, maybe start with 2 for HA). Use a Kubernetes Service to expose internally, and possibly an Ingress or LoadBalancer to expose externally (complete with TLS termination). Mount a ConfigMap or use `env:` in the manifest for configuration variables. Use a Secret resource for sensitive env (like API keys). The Docker image for FastAPI is pulled from our registry. - **n8n:** Deploy n8n either as a single pod (since it has an internal database by default using SQLite, scaling horizontally is non-trivial). Alternatively, run it with an external database (PostgreSQL) to allow multiple replicas. To keep it simple, we often run n8n as one instance (it can handle a good number of workflows, and our use is not extremely heavy per second). Use a PersistentVolume for n8n to store its database (or use an external DB service). Secure the n8n service – likely we don't expose it publicly. It can be ClusterIP only (accessible to FastAPI and maybe to developers via VPN). In production, likely automation and tests are done, so live editing in n8n is rare; still, we protect it with basic auth. - **Prometheus & Grafana:** Either use an existing monitoring stack or deploy dedicated. We can deploy a Prometheus server with a config to scrape our services (via K8s service discovery). Grafana can be deployed and configured with a Prometheus data source and our dashboards. For production, you'd secure Grafana (auth, possibly behind SSO or at least admin password change). Alternatively, if the company has a central Prom/Grafana, just configure those to monitor this namespace (perhaps by annotating pods with Prometheus scrape configs). - **Scaling considerations:** FastAPI can scale horizontally behind a load balancer (it's stateless). n8n is stateful (holds workflows and possibly execution state), so we typically keep one or use a more complex setup with queue mode if needed. The models that are separate microservices (if any) can scale similarly. The RL Router/Overseer LLM if they were separate could be scaled or even use a serverless approach if integrated. - **CI/CD:** Use Kubernetes manifests or Helm charts to deploy. We might provide a Helm chart for Gateway ML Studio to simplify installation (with values for image tags, replica counts, etc.). In absence of that, a `k8s/` directory in the repo with YAMLS would be present. The deployment pipeline would build the images (for any components we maintain) and push to registry, then update image tags in K8s (via kubectl or a GitOps tool like ArgoCD/Flux). - **Environment Specific Config:** Use different ConfigMaps/Secrets for dev, staging, prod with appropriate values. Possibly utilize K8s `configMapRef` and `secretRef` in Pod spec to load those as env.

### Direct VM Deployment:

If not using containers, one could install services on a VM: - Install Python and run FastAPI via Uvicorn (perhaps as a service using systemd or gunicorn). - Install n8n (via npm or as Docker). - Install Prometheus/Grafana (via their binaries or Docker). - This is less ideal for production since dependency isolation is harder, but for completeness: ensure you configure firewall (expose only needed ports), set up process managers for each service, and perhaps use Nginx as a reverse proxy for FastAPI to handle SSL.

### Docker Swarm or Compose in Cloud:

In a simpler cloud setup, you might just run `docker-compose` on a cloud VM (with environment-specific override file). This can work if high availability is not paramount (since one VM is a single point of failure). For a small organization, running the compose on a managed instance could be an easier starting point. Ensure data volumes (for n8n, any databases) are stored on durable storage or backed up.

### **Networking & Security:**

- Only the FastAPI (and Grafana, if needed) are exposed to the internet (through proper channels). They should be behind an HTTPS endpoint. In K8s, an Ingress with cert-manager can do this. In Docker, maybe use a reverse proxy container (Traefik or Nginx) to route HTTPS to FastAPI container.
- n8n, Prometheus are usually not public. If developers need access to n8n UI, they might port-forward or use a secure connection.
- Apply network policies in K8s to restrict communication (e.g., only allow Prometheus to scrape, etc., if cluster provides).
- Set resource limits for containers so one heavy model doesn't starve others. For instance, limit CPU for the LSTM container or memory for XGBoost if needed.
- Use health probes: FastAPI has a `/health` we implement to return OK if app is running (and maybe check dependent services). K8s can use that for liveness/readiness. n8n might have a health endpoint or we just check the port.

### **Storage and Persistence:**

- Models: We need to ensure the latest model files are accessible by the API. There are a few patterns:
  - Bake models into the Docker image (not great for frequent updates).
  - Mount a volume or use a storage service (like S3 or a model registry) from which the container loads models at startup.
  - Or have the training workflow push new models to a location and instruct running containers to reload (could be via an API trigger or simply they pull periodically). In this version, let's say we use a volume (like an NFS or PVC) where models are saved. FastAPI on startup loads model files from there. The retraining workflow, when run, will save new model files to that volume (since it can be run in a container with the volume mounted or via a Kubernetes Job using same volume). Rolling updates of FastAPI can then pick up new models. Alternatively, FastAPI could watch the files and reload if changed, but that's complex; a redeploy is clearer for new model.
- If using a database for logging or storing results, that should be a managed service or a database container with a persistent volume. We didn't emphasize a DB in design, but one could store all predictions and alerts to query later.

### **Verification:**

After cloud deployment, run end-to-end tests:

- Call the public API endpoint from outside (with proper auth if enabled) to see it works.
- Induce an alert scenario (maybe by feeding a known anomalous input) to see if the alert goes through monitoring and notifications.
- Check Grafana dashboards to ensure metrics from the cluster are coming in.

### **Scaling Out Further:**

If load increases, we can replicate FastAPI pods (and behind a load balancer they share the traffic). For horizontally scaling the models themselves, we might instead separate them (e.g., a deployment for "anomaly service" that FastAPI calls). That microservice architecture can be gradually applied if needed – at the start, simpler is fine (monolith FastAPI that does everything). The RL router might become a microservice if heavy or if we want to train it separately.

In summary, the cloud deployment requires configuring the orchestration platform with our container images and ensuring all needed config/secrets are provided. The provided Docker and K8s configurations in our repository aim to make this as smooth as possible, with sensible defaults that mirror our dev/test setups.

### **Documentation & Scripts:**

We include in our repository:

- A detailed `README.md` with quick start for each environment.
- Dockerfiles and docker-compose files for container setups.
- K8s manifests or a Helm chart (with instructions in a

`deploy/` folder). - Possibly a script (maybe `deploy.sh`) for a simplified deployment on a VM (for those not using K8s). - Also, instructions on how to update models or code in a running system (e.g., “to update, build a new image or trigger the retrain workflow”).

By following these instructions, users and contributors should be able to get Gateway ML Studio up and running locally within minutes, and have a clear path to mirror that in a production environment. The use of Docker ensures consistency, and Kubernetes support ensures scalability for enterprise use. Next, we discuss performance considerations and limitations of the system once deployed.

## Performance Notes, Limitations, and Hardware Requirements

Gateway ML Studio is a powerful platform, but its performance and resource needs depend on the components in use and the volume of data. In this section, we discuss how the system performs, highlight potential bottlenecks and limitations, outline typical hardware requirements for various scales, and provide guidance on optimizing for performance.

**Performance Characteristics:** - *Latency*: The end-to-end latency for a prediction request includes data preprocessing, model inference for each involved model, and routing decision. In a typical case (with, say, 3-5 models and an RL routing step), this might be on the order of tens to a couple hundred milliseconds. For example, logistic regression or random forest are very fast (milliseconds). An anomaly detector like Isolation Forest (if using scikit-learn) on moderate feature size is also quick for a single sample. ARIMA/Prophet forecasting might be a bit slower if computing on the fly (but often we wouldn’t retrain ARIMA per request; forecasts might be precomputed periodically). The LLM overseer is the wildcard – calling a large language model could take seconds, which is not feasible for inline predictions. To mitigate, the LLM is used sparingly (maybe only for generating explanations asynchronously, not blocking the core prediction). In high-throughput scenarios, we might disable the LLM or use a smaller/faster one. - *Throughput*: FastAPI with Uvicorn can handle thousands of requests per second on a decent machine, **if** the underlying processing is efficient. The heavy lifting is in model inference. Many models (like tree ensembles) are CPU-bound. If we expect high throughput, we ensure to either vectorize operations (if many requests can be batched) or scale horizontally. For example, if each request calls XGBoost and an LSTM, those can use CPU and possibly GPU respectively. We might set up separate inference workers for GPU models to not block CPU worker threads. - *Parallelism*: The design using n8n and async FastAPI means that calls to multiple models can be done in parallel. We should ensure that our Python code releases GIL when possible (numpy does for heavy calcs, but if using pure Python loops that could bottleneck). Uvicorn can handle many concurrent connections via async, so waiting on network calls (to microservices or LLM API) doesn’t stall others. n8n by nature will have some overhead (it’s not as low-level as hard-coded concurrency). In extreme performance-critical systems, one might re-implement the orchestration in code to shave overhead, but for moderate loads n8n overhead is acceptable given the productivity it offers. - *RL Router Overhead*: The RL decision is typically a simple computation (a lookup table or a quick neural net or tree based on inputs). So the overhead of the router is negligible compared to the model inference times.

**Potential Bottlenecks & Limitations:** - *n8n Workflow Overhead*: While n8n is convenient, it’s not written in a low-level language – each node adds a bit of overhead in processing. For extremely low-latency needs (<10ms), n8n might be too slow. In such cases, one could refactor the critical path (for example, make the prediction path embedded in FastAPI code directly). But for typical use (100ms+ latency acceptable), n8n overhead is in the few milliseconds range per node perhaps, which is fine. - *Memory Usage*: Loading many models can be memory heavy. For instance, a large Random Forest or XGBoost could be hundreds of MB if

trained on a huge dataset. LSTM models and Prophet aren't too large typically, but if we store entire training data for some reason in memory, that's an issue. We must watch memory – containers should have enough RAM to hold all models plus some overhead for data. If memory is constrained, consider loading models on demand or using smaller versions. Also, Python's memory is not always released to OS; long-running processes might fragmentation – hence monitoring memory usage and maybe periodic restarts of workers could help. - *GPU usage*: If LSTM or other deep models are run on GPU (which is recommended for speed if the model is large), we need to ensure the environment has a GPU and the model code (like TensorFlow/PyTorch) is using it. That adds complexity: drivers, etc. But for high volume time-series forecasting with LSTM or if we embed any deep learning for anomaly detection, GPU can drastically cut inference time. We'd then have to schedule those pods on GPU nodes or run separate service for them. - *Disk I/O*: Not usually a big factor unless we are reading/writing large model files frequently. We try to keep model files loaded in memory. Logging to disk (if any) should be async or non-blocking. If using a database for logs, the I/O goes over network to DB. - *Networking*: If our architecture calls microservices for each model, then network latency adds up. Since our default approach keeps most models within the FastAPI or orchestrator process, we avoid that overhead. But if we scale out, say moving each model to its own container, then each call might add, for example, 1-2 ms network overhead if on same host (or more across hosts). It's usually small compared to model times, but it's something to consider. Using gRPC can cut some overhead vs REST for internal calls. - *Precision of Anomaly Detection*: A limitation to note: unsupervised anomaly detectors can produce false positives. Tuning them is non-trivial. Gateway ML Studio provides the framework but it's not magic – it relies on the models. If underlying data shifts in a way that confuses an anomaly model, it might spam alerts. The ensemble helps reduce this by cross-validating with other models, but users should be aware that periodic review of anomaly model performance is needed. This is why we integrated drift detection and retraining triggers. - *RL Exploration*: The RL router will need to explore to learn. In a production setting, this means occasionally it might route to a suboptimal model to gather experience, which could cause a slight performance dip or a missed detection. We typically handle this by training the RL policy offline on historical data if possible (treating it as a contextual bandit supervised problem first) and then doing minimal exploration online, or doing A/B tests in a shadow mode. This is a complexity that is an "advanced" feature – one could initially run with a simple static ensemble, then gradually enable the RL optimization. - *Oversee LLM Limitations*: Using an LLM for oversight is experimental. Its outputs are not guaranteed to be correct. We do not fully automate actions based on LLM output without human validation (unless in non-critical contexts). So one limitation is that the LLM might identify a pattern or explain something, but it might be wrong. We use it carefully (perhaps just logging its analysis for humans to read). Also, if using an external API for LLM, that introduces dependency on that service's availability and compliance (data might be sent out, which may be a privacy issue; we mitigate by not sending raw sensitive data, maybe only abstracted info). - *Prometheus Metric Cardinality*: We must be careful not to create unbounded metrics (like per user or per transaction labels), as that can degrade Prometheus. We stick to coarse metrics. This is a general monitoring consideration to keep performance of monitoring in check.

**Hardware Requirements:** - *Development*: A standard laptop (8-16 GB RAM) can run a small sample of this system (with smaller models) easily. Using Docker might need a bit more memory allocated if running multiple containers. - *Small Deployment*: For a small scale (say monitoring a few streams, low QPS, few models), a single VM or instance with 4 CPU cores and 16 GB RAM could suffice. This could host all components (with Docker or not). Models like XGBoost can use multiple threads, so having a few cores helps. - *Production (Medium Scale)*: If we expect, e.g., 50 predictions per second and heavy models (like deep learning), we'd want dedicated instances. Possibly: - 2 CPU cores and 4GB RAM for FastAPI app (per replica). - 4 CPU cores and 8GB RAM for a model service if combining, or more if running heavy models in-process. - If LSTM on CPU, maybe more cores or it will be slow; better to use a GPU (like an NVIDIA T4 or similar). - The

RL model (if a small neural net) is negligible, but if we made it fancy (like a reinforcement learning agent with large state) it could use some CPU/GPU. - n8n is lightweight (it can run on 1-2 cores, 1GB RAM easily for orchestrations, though if many workflows or heavy JavaScript in function nodes, give it 2-4GB). - Prometheus/Grafana: at medium scale, 2 CPU and 4GB combined might suffice, unless storing very long retention. - *High Scale*: If expecting thousands of requests/sec, we'd likely deploy multiple FastAPI instances behind a load balancer. Also, we might employ asynchronous inference batching for models (especially if using GPU for LSTM – e.g., batch 32 requests and do one matrix forward pass). We'd also possibly split responsibilities: a separate anomaly detection service that streams data and triggers independently, rather than computing on each request. - Hardware for high scale might involve multiple machines or pods each focusing on part of the workload. For example, a GPU server for all deep learning models, CPU servers for tree models, etc. - Also consider using distributed processing for some tasks (like if doing large scale forecasting on big time series, using Spark or Dask – but that's beyond current scope). - *Storage*: Not huge – model files maybe in the order of MBs to a couple GBs if including large neural nets. Data logs depends; if storing every prediction, that could accumulate but you might pipe them to a data warehouse instead of local disk. Prometheus storage depends on retention and frequency (with modest metrics, default retention 15 days, typically few GB of space). - *Network*: If self-hosting components, ensure low-latency network between them (in same data center/cluster). If your LLM oversight calls a cloud API, you need internet access and note that latency ~100ms+ for that call – again, presumably done asynchronously.

#### **Optimizations:**

- Use of vectorized libraries (NumPy/Pandas) wherever possible in preprocessing and avoid Python loops in critical path.
- Enable XGBoost to use multiple threads (by default it does; ensure not set to 1 unless needed to avoid interference).
- If using PyTorch for LSTM, leverage TorchScript to optimize or even quantize the model if performance is an issue.
- Prophet can be slow if fitting; we only fit Prophet offline and use results for forecasting online (or keep a Prophet model object in memory to compute future points quickly).
- The Router's RL algorithm training should be done with efficient frameworks; but since it might be a small state-space, even a simple implementation is fine. If we do something advanced like deep RL, we'll use libraries (Ray RLlib or stable-baselines) which can use vectorized envs.
- Monitor the event loop in FastAPI: if any synchronous tasks creep in (like a long DB call without async), it could block others. We try to make all I/O async or offload to threads as needed.

**Known Limitations:**

- The system currently assumes a relatively homogeneous input stream. If you have vastly different data types, you might need to deploy separate instances of the studio or significantly modify workflows.
- Not all edge cases of ensemble logic are handled – e.g., if all models abstain or produce extremely contradictory results, human oversight might be needed to resolve.
- Cold start: If using an LLM or some models, initial loading might be slow (e.g., downloading a model on first run). We mitigate by warm-up procedures (the FastAPI startup event could run a dummy request through each model to load it into memory).
- Explainability: Combining many models can reduce interpretability. We provide partial measures like feature importances or the LLM's textual explanations, but rigorous users might need to do their own analysis to trust decisions. This is a trade-off for the complexity and performance gained.

Despite these considerations, Gateway ML Studio is built to be robust and performant for a wide range of use cases. By tuning deployment configurations and hardware allocation, it can be adapted from small-scale testing up to real-time production scenarios that demand both speed and accuracy. We encourage users to start with the provided defaults and profiles, measure the performance under their workload, and incrementally adjust resources and settings as needed.

# GitHub Repository, Roadmap, and Contribution Guidelines

Gateway ML Studio is envisioned as an open and evolving project. In this final section, we point to the GitHub repository (which serves as a placeholder for code and documentation), outline the future roadmap of features and improvements, and provide guidelines for external contributors who wish to collaborate on the project.

## GitHub Repository

The project's source code, examples, and documentation are hosted on GitHub for easy access and version control. The repository (placeholder name: `gateway-ml-studio`) contains the following structure:

- `/src` : The source code for the FastAPI app and any supporting Python code (model wrappers, router logic, etc.).
- `/workflows` : Exported n8n workflow JSON files and possibly scripts to import them.
- `/models` : Either the trained model artifacts (if small and for demo) or code to download/generate them. We might use Git LFS for larger files or provide a separate link.
- `/deploy` : Dockerfiles, Docker Compose files, and Kubernetes manifests/Helm charts for deployment.
- `/docs` : Additional documentation, possibly including this whitepaper, usage guides, architecture diagrams.
- `README.md` : A comprehensive introduction with quick start steps, prerequisites, and basic usage. This is the first point of contact for new users visiting the repo.
- `CONTRIBUTING.md` : A guide specifically for contributors (which we'll summarize below).
- **Issue Tracker and Discussions:** The repository's Issues page is used for bug reports, feature requests, and project tasks. We encourage users to report any problems or suggestions there. The Discussions tab (if enabled) can be used for more open-ended Q&A or brainstorming.

We will tag releases (using semantic versioning, e.g., v1.0.0 for the first stable release). Each release will correspond to a certain set of features and will be documented in the Release notes section on GitHub.

## Roadmap

Gateway ML Studio is a living project, and we plan to extend its capabilities. Our roadmap includes both near-term enhancements and long-term ambitious features:

1. **Improved User Interface & Dashboards:** Currently, interaction is via API and Grafana. We plan to add a simple web dashboard UI for monitoring the models and alerts, possibly as a small React frontend that consumes the FastAPI (for instance, showing recent anomaly alerts, model performance metrics, etc., in a user-friendly way). This can help stakeholders who prefer not to dig into Grafana or raw logs.
2. **Automated Model Management:** Introduce a Model Registry and versioning in a more formal way. This includes tracking datasets, experiment results, and possibly integrating with tools like MLflow. This would facilitate comparing model versions and rolling back if needed.
3. **Online Learning Capabilities:** Extend the RL Router to update in real-time (online learning) when possible, and to incorporate multi-objective optimization (like balancing precision and recall for anomalies via reward function tweaks). Additionally, exploring bandit algorithms for ensemble weighting (a simpler alternative to full RL) for cases with immediate feedback.
4. **Additional Algorithms:** Enrich the model lineup with more options:
5. **Novelty detection algorithms:** e.g., Deep Autoencoders beyond LSTM (for generic data), or Isolation Forest variations like Extended Isolation Forest.

6. **Advanced forecasting models:** e.g., integrate a Facebook Prophet's successor or any probabilistic forecasting models (like NeuralProphet or GluonTS models), and a variety of classical models (maybe dynamic regression, etc.).
7. **Graph-based anomaly detection** if applicable (for example, if data involves networks).
8. Possibly an **online change detection** algorithm like ADWIN for continuously monitored streams.
9. **Edge/Stream Deployment:** Simplify a mode of deployment for edge environments (where you might not run full n8n or heavy containers). This could involve an all-in-one binary or using lightweight orchestrators. The idea is to allow running a subset of the system on a single device or VM for specific tasks (e.g., anomaly detection on IoT gateway).
10. **Better Integration with Data Lakes/Streams:** Right now, ingestion is assumed via n8n or API. We want to integrate with streaming platforms like Kafka or MQTT – so that Gateway ML Studio can subscribe to a data stream, process events, and output anomalies/predictions in real-time to another topic. This can widen the use-case (fitting into existing data pipelines).
11. **LLM Overseer Enhancement:** Explore making the Overseer LLM more interactive. For instance, allow it to take actions in a sandbox (maybe generating natural language reports of anomalies daily for management). Also evaluate open-source smaller LLMs that could be fine-tuned to understand our domain, which can then run locally (removing the dependency on an external API).
12. **Hardening and Security:** Conduct security audits and performance testing. For example, ensure the system is robust against malicious inputs (like injection attacks via input data, which might be a vector if any part passes data to a shell or database, or prompt injection for the LLM). Also implement rate limiting and authentication/authorization best practices so it can be safely exposed in enterprise environments.
13. **Testing & CI Improvements:** Increase automated test coverage, including integration tests that spin up a testing environment with containers to simulate workflows. Possibly integrate with GitHub Actions to run tests on each pull request. Also, add tests for monitoring (simulate drift and see if alert triggers in a test environment).
14. **Documentation & Examples:** Expand documentation with more examples and use-cases (for instance, a tutorial on how to customize the studio for a different domain, or how to add a new model into the ensemble). Possibly provide pre-built examples for common tasks (fraud detection, predictive maintenance, etc.) as separate recipe documents or branches.
15. **Community Engagement:** If interest grows, potentially set up a community Slack/Discord or use GitHub Discussions actively so users can share experiences or ask for support. The roadmap itself will be community-driven; we will label issues that are open to contribution and welcome community suggestions.

We maintain a Roadmap file or project board on GitHub where these items (and others) are tracked with timelines or release targets, subject to change based on priorities.

## Contribution Guidelines

We highly encourage contributions from internal team members and the open-source community. To ensure a smooth collaboration process, we have established the following guidelines (detailed in `CONTRIBUTING.md` in the repo):

- **Project Setup for Contributors:** Instructions to fork the repo, get the development environment up (similar to the local setup above), and best practices to test changes.
- **Issue Tracking:** We use GitHub Issues for tracking bugs and proposed enhancements. If you plan to work on something, it's best to comment on an existing issue or open a new one to discuss your

approach before writing code (especially for larger changes). This avoids duplicate work and ensures alignment with the project's direction.

- **Branching and Pull Requests:** The repository likely uses a branching model where `main` or `master` is stable. Feature development should be done on separate branches (preferably named descriptively, like `feature-add-prophet-intervals` or `bugfix-drift-threshold`). Once your feature or fix is ready, open a Pull Request (PR) to merge into the main branch. We have a PR template which reminds you to detail what was changed, include any relevant screenshots or logs, and reference issues it closes.
- **Coding Standards:** We enforce coding style via linters (for Python, using flake8/Black; for Node/JS in n8n nodes, maybe ESLint). Contributors should run these (we provide configs) to ensure consistency. Also, we follow best practices like writing clear, modular code, adding comments where non-obvious, and keeping functions small. For Python, type hints are appreciated as they make the code easier to understand.
- **Testing:** If you add a new feature, try to add corresponding tests. We use Pytest for Python tests. For n8n workflows, we might include some example input/outputs to test the logic (though automated testing of workflows is tricky; at least maybe via API calls in tests). If a contribution is a bug fix, ideally include a test that fails before the fix and passes after. We run tests in CI for each PR.
- **Documentation:** All significant contributions should also update relevant documentation. If you add a new model or feature, update the README or a relevant section in docs. We have a docs site (maybe via GitHub Pages or just markdown in the repo) – ensure any new public APIs or config options are documented. This helps users quickly adopt your contribution.
- **Commit Messages:** We prefer conventional commit messages (for example, prefix with `feat:`, `fix:`, `docs:`, `chore:`, etc.). This helps in auto-generating changelogs. E.g., “`feat: add support for ARIMA seasonal order configuration`” or “`fix: correct drift detection threshold logic`”.
- **Code Reviews:** The maintainers will review PRs. We aim to be prompt and constructive. As a contributor, be open to feedback – we might request changes or have discussions about approach. It’s part of making sure the code fits well into the project. Similarly, maintainers will be respectful and thankful for your contributions.
- **License and CLA:** Gateway ML Studio is released under an open-source license (to be specified, e.g., MIT or Apache 2.0). By contributing, you agree that your contributions will be under the same license. We may use a CLA (Contributor License Agreement) tool or DCO (Developer Certificate of Origin) sign-off to ensure all legal aspects are clear – check CONTRIBUTING.md for details if applicable.
- **Attribution:** We list significant contributors in the CONTRIBUTORS file or README. We appreciate every contribution, from small typo fixes in docs to major feature additions.

Finally, for those who want to contribute but are not sure where to start, we label some GitHub issues as “good first issue” or “help wanted.” These are typically self-contained tasks that are friendly to new contributors. We encourage you to pick one of those and make your first PR.

#### Contact and Support:

For any questions or support, contributors and users can open issues or join the project’s discussion forums. Maintainers (which currently include the original authors and core team) will try to respond and guide as needed. As the project grows, we might set up scheduled community calls or a chat channel as mentioned in the roadmap.

---

We thank you for reading the Gateway ML Studio whitepaper. By combining a rich set of models, adaptive routing, and robust orchestration, we hope this platform empowers organizations to more easily deploy complex ML-driven analytics. We welcome feedback and collaboration to improve it further. Together, through open development, Gateway ML Studio can continuously evolve into a more powerful, efficient, and reliable system for machine learning operations.

---

1 2 3 4 31 Building a Real-Time Fraud Detection Pipeline with Python, FastAPI, and n8n | by arjun nagathankandy | Sep, 2025 | Medium

<https://medium.com/@arjunac009/building-a-real-time-fraud-detection-pipeline-with-python-fastapi-and-n8n-f8d43250de51>

5 14 17 Demand forecasting algorithms - Supply Chain Management | Dynamics 365 | Microsoft Learn

<https://learn.microsoft.com/en-us/dynamics365/supply-chain/demand-planning/forecast-algorithm-types>

6 7 8 9 Anomaly Detection Without Neural Networks: Isolation Forest, LOF, and Other Useful Techniques | by Jorge Martinez Santiago | Medium

<https://medium.com/@jorgemswork/anomaly-detection-without-neural-networks-isolation-forest-lof-and-other-useful-techniques-385e85e68c26>

10 CUSUM - Wikipedia

<https://en.wikipedia.org/wiki/CUSUM>

11 25 Anomaly Detection in Time Series Data using LSTM Autoencoders | by Zhong Hong | Medium

<https://medium.com/@zhonghong9998/anomaly-detection-in-time-series-data-using-lstm-autoencoders-51fd14946fa3>

12 20 Novelty Detection with Local Outlier Factor (LOF) in Scikit Learn - GeeksforGeeks

<https://www.geeksforgeeks.org/machine-learning/novelty-detection-with-local-outlier-factor-lof-in-scikit-learn/>

13 15 16 ARIMA vs Prophet vs LSTM for Time Series Prediction

<https://neptune.ai/blog/arima-vs-prophet-vs-lstm>

18 19 Stacking Ensemble With XGBoost Meta Model (Final Model) | XGBoosting

<https://xgboosting.com/stacking-ensemble-with-xgboost-meta-model-final-model/>

21 24 Deep-Reinforcement-Learning-Based Dynamic Ensemble Model for Stock Prediction

<https://www.mdpi.com/2079-9292/12/21/4483>

22 What is hybrid anomaly detection?

<https://milvus.io/ai-quick-reference/what-is-hybrid-anomaly-detection>

23 A Governance-First Paradigm for Principled Agent Engineering - arXiv

<https://arxiv.org/html/2510.13857v1>

26 27 28 29 How BasisAI uses Grafana and Prometheus to monitor model drift in machine learning workloads | Grafana Labs

<https://grafana.com/blog/2021/08/02/how-basisai-uses-grafana-and-prometheus-to-monitor-model-drift-in-machine-learning-workloads/>

30 Convert n8n Workflows to Python FastAPI and Self-Host with Ease ...

<https://medium.com/@adityb/convert-n8n-workflows-to-python-fastapi-and-self-host-with-ease-using-n8n2py-me-85072afab1b3>

32 33 34 The Twelve-Factor App

<https://12factor.net/config>