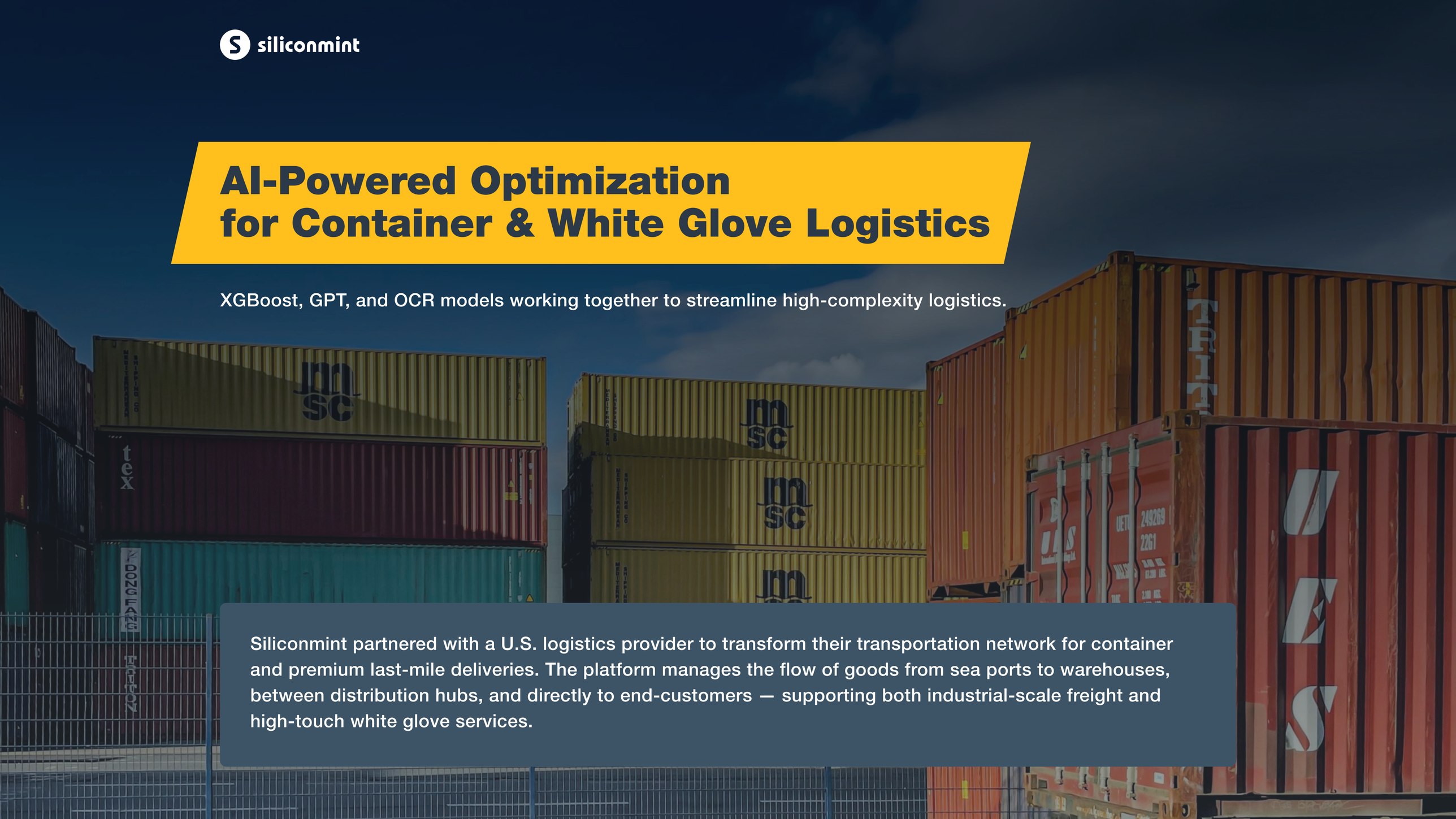
siliconmint

AI-Powered Optimization for Container & White Glove Logistics

XGBoost, GPT, and OCR models working together to streamline high-complexity logistics.



Siliconmint partnered with a U.S. logistics provider to transform their transportation network for container and premium last-mile deliveries. The platform manages the flow of goods from sea ports to warehouses, between distribution hubs, and directly to end-customers — supporting both industrial-scale freight and high-touch white glove services.

siliconmint

Business Challenge

The client faced several critical issues in managing container and specialized deliveries across the U.S.:

To remain competitive, the client required a platform capable of dynamically sourcing available drivers and vehicles, minimizing idle time, and automating complex route and delivery decisions.

Driver and Truck Availability:

A recurring problem was the inability to quickly locate and assign available trucks and drivers for container pickup at sea ports. Containers often remained idle at terminals or port-side warehouses, incurring storage fees and causing downstream delays.

Idle Container Costs:

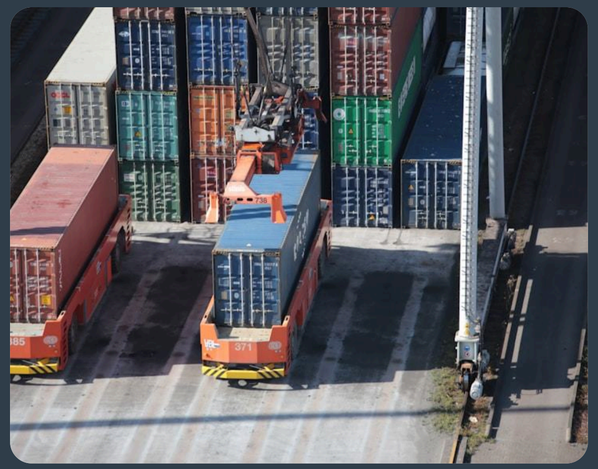
Delays in finding transportation resulted in long dwell times for containers, leading to increased demurrage charges and inefficient asset utilization.

Lack of Real-Time Coordination:

The legacy systems lacked real-time insight into port schedules, traffic, or vehicle availability, which made proactive planning impossible and introduced bottlenecks across the delivery chain.

Diverse Delivery Requirements:

The company needed a unified solution that could serve not only industrial-scale container logistics, but also handle white glove delivery scenarios — where precise timing, handling, and customer experience are essential.



siliconmint

Technical Challenge

The core technical challenge was to build a logistics platform that could intelligently track, predict, and coordinate the availability of freight drivers — effectively orchestrating deliveries in real time. This required solving several deep operational and engineering problems:

Dynamic Driver Availability Tracking:

Identify which drivers are currently active, who will be available later, and who might drop out — using real-time app signals and historical behavior patterns.

Disruption Response & Reallocation:

Automatically handle no-shows, early/late arrivals, and last-minute cancellations — and reassign tasks without human bottlenecks.

AI-Assisted Dispatcher Operations:

Provide dispatchers with AI assistance for interpreting free-text updates from drivers, generating responses, and flagging operational issues.

Document Processing & Validation:

Process and validate proof-of-delivery documents, customs paperwork, and gate passes via AI — with OCR and field extraction.

siliconmint

Model definitions

MODEL NAME	ARCHITECTURE	PROVIDER	PURPOSE	INPUT	OUTPUT	USAGE CONTEXT
DriverPredict-XGB	XGBoost (Gradient Boosting Trees)	Custom (in-house)	Predicts driver availability based on historical behavior and app data	Driver ID, activity history, shifts, app pings	Availability probability for time slots	Used in planning to pre-allocate or reassign deliveries
DispatchGPT	Large Language Model (GPT-4 Turbo)	OpenAI	Interprets driver messages, suggests replies, flags risk	Free-text message, dispatcher context	Suggested response, summarized intent, risk flags	Live assistant in dispatcher UI
DocOCR-Validator	TrOCR + LayoutLMv3	Microsoft / HuggingFace	Processes and validates delivery documents	Scans, PDFs, images	Structured data, validation results, compliance status	Document processing for proof-of-delivery and compliance workflows

Model Runtime Characteristics

MODEL NAME	AVG INFERENCE TIME	RETRAINING FREQUENCY	CONTEXT VOLUME	EXPLAINABILITY	REAL-TIME COMPATIBLE
DriverPredict-XGB	~200 ms	Weekly	~50 variables (driver activity patterns)	✅ SHAP feature importance	✅ Yes
DispatchGPT	~500–700 ms	N/A (prompt-based)	~300–500 tokens (text + metadata)	✅ Natural-language explanation (LLM)	✅ Yes
DocOCR-Validator	~2–3 sec	Rare (base models), rules updated quarterly	~1 document page (image + form structure)	⚠️ Partial (OCR confidence + heuristics)	⚠️ Batch or async preferred

siliconmint

Model Comparison for Dispatcher Assistant

MODEL	RESPONSE TIME	OUTPUT QUALITY	STRENGTHS	WEAKNESSES
GPT-4 Turbo (Chosen)	~500–700 ms	High (natural, contextual, actionable)	Fast, cost-effective, highly contextual	Limited to ~128k tokens; sometimes verbose
GPT-4 (original)	~2–3 sec	Very High (nuanced, detailed, reliable)	Deep understanding, great in edge cases	High latency and cost for frequent UI interactions

DriverPredict Model Comparison

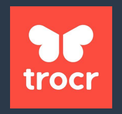
MODEL NAME	ACCURACY	TRAINING TIME	INFERENCE TIME	EXPLAINABILITY	STRENGTHS	WEAKNESSES TIME
XGBoost (Chosen)	High (great for structured behavioral data)	Fast	~200 ms	✅ SHAP feature importance	Interpretable, robust, effective for tabular data	May need manual feature engineering
LightGBM	Comparable to XGBoost	Very fast	~150 ms	✅ SHAP	Fast training, scalable on large datasets	Slightly harder to tune for accuracy
Temporal Fusion Transformer	Very high for time-series	Slow	~1–2 sec	✅ Attention-based insights	Captures time dependencies, interpretable via attention	Complex infra, slow inference, overkill

siliconmint

Tech Stack

GPT-4 Turbo

XGBoost



TrOCR

LayoutLMv3

Node.js

Express js

GraphQL

React

MongoDB

PostgreSQL

AWS
(ECS, Fargate, Lambda)

siliconmint

How Was AI Integrated?

The platform primarily operates as a logistics marketplace, where drivers — either independently or via affiliated companies — browse and accept freight assignments directly through the app. To ensure reliability, especially during peak load times or when certain shipments were not picked up, the company also maintained a pool of internal and contracted drivers who could fill the gaps.

In earlier stages, when a shipment remained unassigned or was at risk of delay, dispatch managers would manually reach out to drivers, remind them to log in, or negotiate urgent coverage. While effective, this workflow placed a heavy burden on the operations team.

To reduce that load, AI was gradually introduced — starting with the most repetitive and scalable processes.

Phase 1

Handling Driver Inquiries and Documents

The first use case for AI was to answer frequently asked questions from drivers, such as:

- “Where do I drop off the container?”
- “What documents are needed at pickup?”
- “Is the terminal open before 7am?”

These questions were common and predictable, though phrased in slightly different ways. Using GPT-4 Turbo, we trained an assistant to handle them based on historical driver-dispatcher conversations. For document-related queries, the AI also leveraged OCR and document understanding models (TrOCR + LayoutLMv3) to extract key data from uploaded scans or images.

This freed up dispatchers from a high volume of routine messages and reduced driver response time significantly.

Phase 2

Predicting Assignment Gaps Before They Happen

Next, we focused on a more strategic application: helping managers detect in advance when a delivery might not be picked up through the marketplace.

A predictive model was trained to identify patterns in regions, load types, and time windows where assignments historically fell through. Instead of waiting until the last moment, the system could now surface shipments with a high likelihood of going unclaimed — even if the deadline hadn't passed.

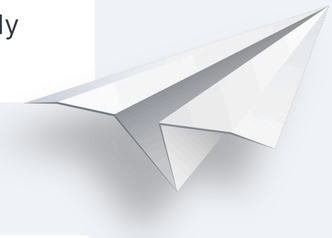
By doing so, the system helped minimize disruptions and gave dispatchers more time and flexibility to act — turning reactive triage into proactive planning.

This enabled the platform to:

- Send proactive reminders to drivers
- Notify a manager to intervene early
- Suggest fallback options from the internal driver pool

With AI embedded into day-to-day workflows, the platform now offloads a large portion of repetitive operational tasks:

- Answers driver questions automatically via AI assistant
- Interprets documents and extracts key delivery details
- Identifies high-risk assignments early based on patterns and rules



This shift allows managers to:

- Focus on exceptions that require human judgment
- Make more strategic decisions, not micromanage the routine
- Trust the system to handle predictable, repeatable tasks

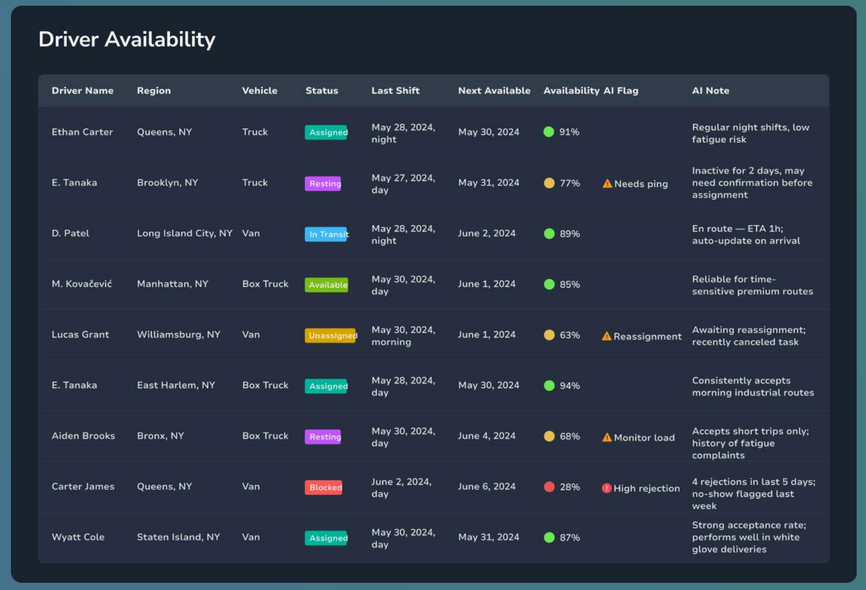
siliconmint

Our Solution

We delivered a modular, AI-driven logistics platform optimized for real-time container and white glove coordination. Key components include:

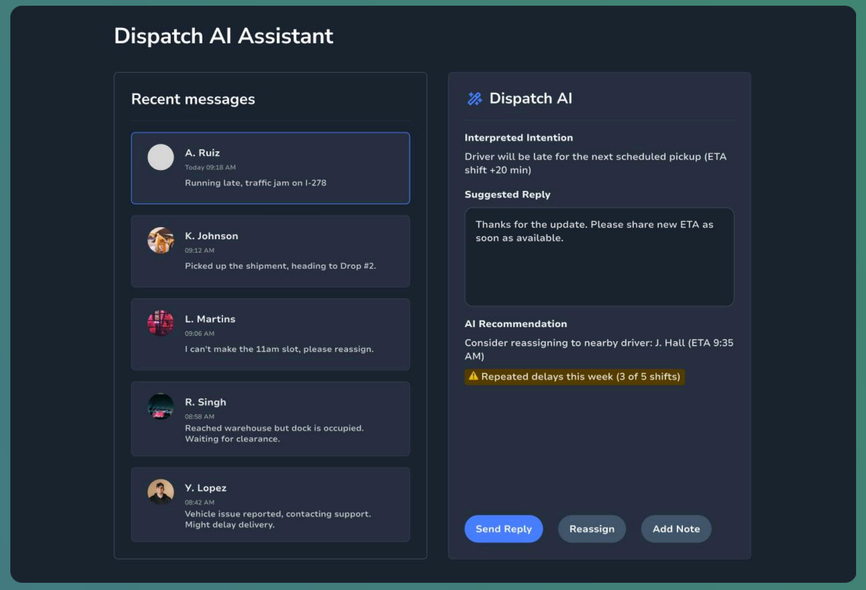
DriverPredict-XGB

A predictive model (XGBoost) that forecasts driver availability using real-time telemetry, app usage, and historical shift patterns.



DispatchGPT (GPT-4 Turbo)

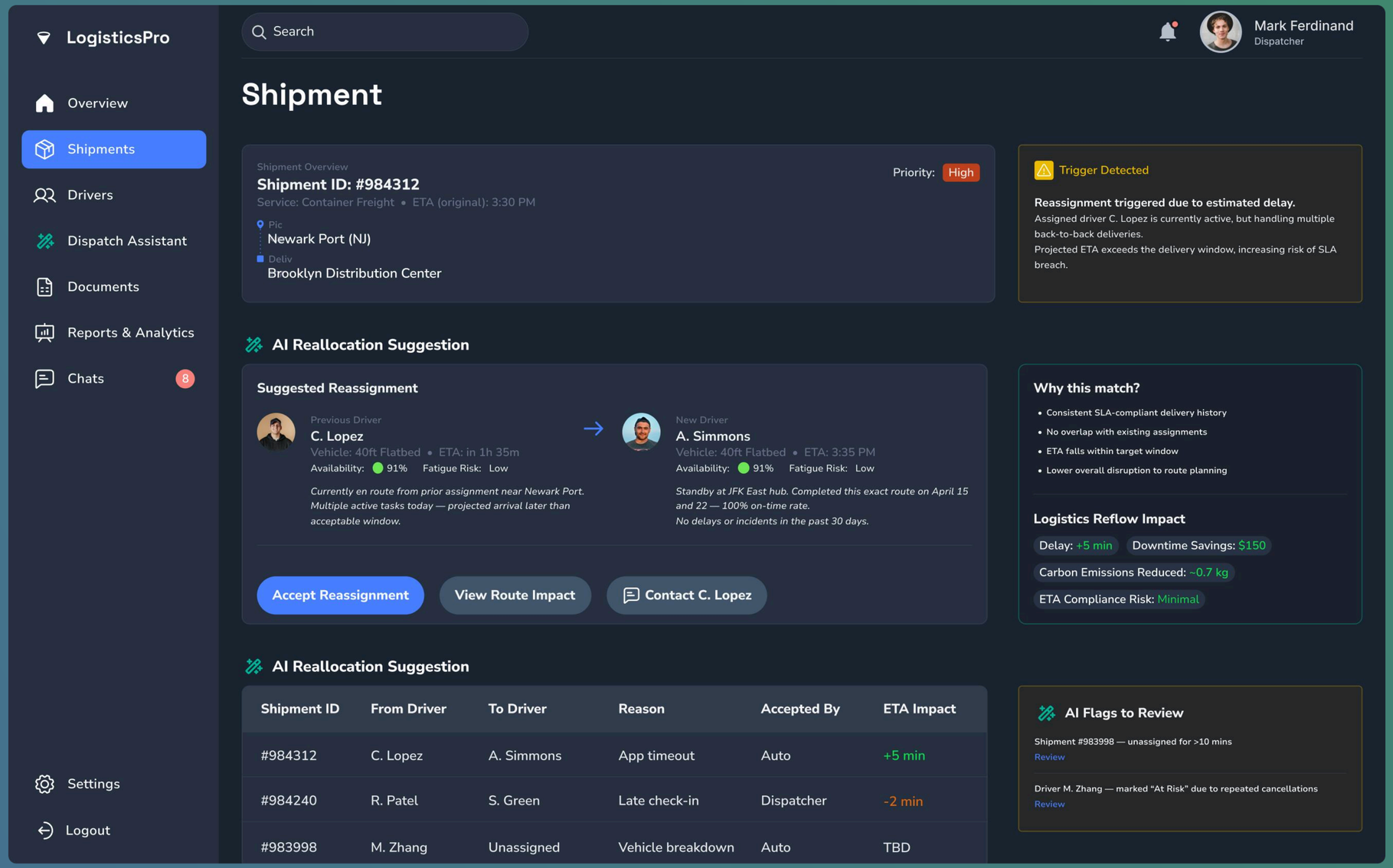
An LLM-powered assistant that helps dispatchers interpret driver messages, suggest responses, and recommend reassignment actions.



siliconmint

Smart Reallocation Engine

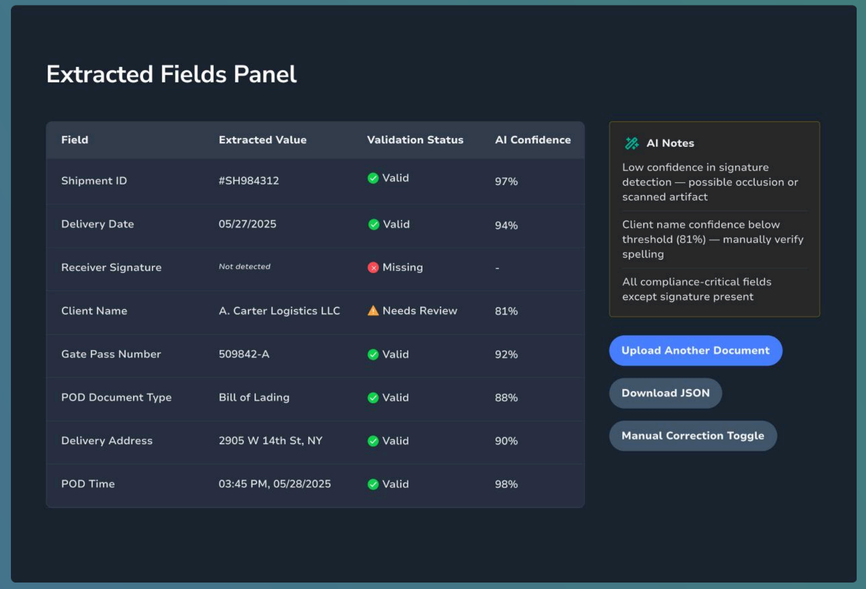
Combines rule-based logic with reinforcement learning to dynamically reassign shipments based on real-time availability and delivery windows.



siliconmint

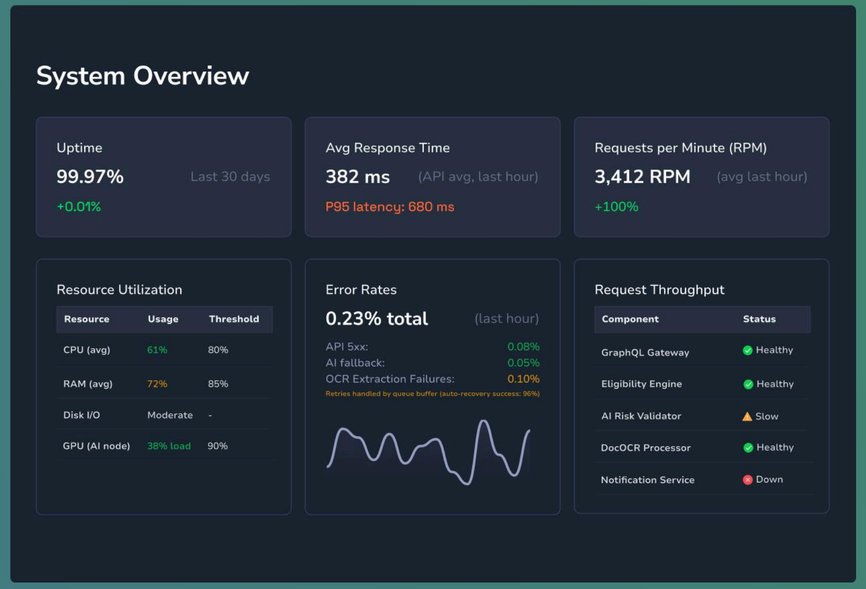
DocOCR-Validator

A document-processing pipeline built on TrOCR + LayoutLMv3 for extracting and verifying fields in delivery documents and compliance forms.



Scalable Architecture

Backend (Node.js, GraphQL, AWS, PostgreSQL + MongoDB) and frontend (React) support large-scale dispatching and operator control.



siliconmint

Impact

Up to 95% automation

of route and assignment planning

99% on-time pickup

and delivery across **500,000+** annual shipments

3x improvement in SLA adherence

for white glove deliveries

70% boost in operational efficiency

for container transport

30% fewer idle/empty miles

reducing logistics overhead and CO_2 emissions

>90% automation

in document verification and compliance handling