# SCM-Arena: Supply Chain Management LLM Benchmark Platform

## 🎯 What is SCM-Arena?

SCM-Arena is the \*\*first standardized benchmark platform\*\* for evaluating Large Language Models on supply chain management coordination tasks. Unlike existing benchmarks that focus on individual optimization or knowledge recall, SCM-Arena tests models' ability to coordinate effectively in multi-agent supply chain scenarios.

### Core Value Proposition

- \*\*For Researchers\*\*: Compare your SCM AI systems against industry standards

- \*\*For Industry\*\*: Identify the best LLM solutions for supply chain applications

- \*\*For Community\*\*: Track progress in supply chain AI coordination capabilities

## 🏗️ Platform Architecture

### What SCM-Arena Hosts

```

SCM-Arena Infrastructure:

├── 🎮 Beer Game Engine (Python)

├── 📊 Evaluation Pipeline (FastAPI + Celery)

├── 🗄️ Results Database (PostgreSQL)

├── 📈 Leaderboard Website (Next.js)

└── ⚡ Admin Dashboard (Model submissions)

```

### What Researchers Host

```

Researcher Infrastructure:

├── 🤖 Their Model/RAG System

├── 🌐 API Endpoint

├── 💰 Their Own Compute Costs

└── 🔧 Their Own Tech Stack

```

### Evaluation Flow

```

1. Researcher submits API endpoint → SCM-Arena

2. SCM-Arena calls their API with beer game scenarios

3. Model responds with supply chain decisions

4. SCM-Arena runs full multi-agent simulation

5. Results stored & leaderboard updated

6. Public can view performance analytics

```

## 🎮 First Benchmark: Multi-Agent Beer Game

### Why the Beer Game?

- \*\*50+ years\*\* of validation in operations research literature

- \*\*Known failure patterns\*\* (bullwhip effect, coordination breakdowns)

- \*\*Clear performance metrics\*\* (cost, service level, variance amplification)

- \*\*Anti-gaming properties\*\* (emergent coordination can't be memorized)

### Benchmark Scenarios

```

Core Evaluation Matrix:

├── 📊 Information Visibility (3 conditions)

│ ├── Local Only: Own inventory + orders

│ ├── Neighbor Visibility: Adjacent agents' states

│ └── Full Chain: Complete supply chain transparency

├── 🧠 Memory Conditions (2 types)

│ ├── Cold Start: No historical memory

│ └── Warm Start: 3-round memory window

└── 📈 Demand Patterns (4 types)

├── Classic: 4-4-4-4-8-8-8-8-4-4-4-4

├── Random: Normal distribution (μ=6, σ=2)

├── Shock: Stable with periodic spikes

└── Seasonal: Cyclical with trend

Total: 3 × 2 × 4 = 24 unique scenarios per model

Runs: 10 games per scenario = 240 total evaluations

```

### Evaluation Metrics

```

🎯 System Performance:

├── Total Supply Chain Cost (primary metric)

├── Bullwhip Ratio (variance amplification)

├── Service Level (orders fulfilled %)

└── Inventory Efficiency (turns ratio)

🤝 Coordination Quality:

├── Convergence Time (rounds to stability)

├── Information Utilization Efficiency

├── Coordination Coefficient

└── Decision Consistency

🛡️ Robustness:

├── Shock Recovery Time

├── Performance Variance

└── Degradation Under Heterogeneity

```

## 🌐 Website Structure

### Public Website (scm-arena.com)

#### Homepage/Leaderboard

```

🏆 SCM-Arena Leaderboard

┌─────────────────────────────────────────────────────────────────┐

│ Rank │ Model │ Score │ Cost │ Bullwhip │ SL% │

├──────┼───────────────────────────┼───────┼───────┼──────────┼─────┤

│ 🥇 │ GPT-4-Supply-Chain-RAG │ 87.3 │ $234 │ 1.2 │ 94% │

│ 🥈 │ Claude-3-SCM-Expert │ 84.1 │ $267 │ 1.4 │ 92% │

│ 🥉 │ Llama-70B-Fine-tuned │ 79.8 │ $289 │ 1.6 │ 89% │

│ 4 │ Custom-Research-RAG │ 76.4 │ $312 │ 1.8 │ 87% │

│ 5 │ Mistral-7B-SCM │ 71.2 │ $345 │ 2.1 │ 84% │

└─────────────────────────────────────────────────────────────────┘

📊 Categories:

├── 🟢 Base LLMs (pure models)

├── 🔵 RAG Systems (retrieval augmented)

├── 🟡 Fine-tuned Models (domain specific)

└── 🟣 Hybrid Systems (multi-component)

```

#### Model Detail Pages

```

🔍 GPT-4-Supply-Chain-RAG - Performance Deep Dive

📈 Key Metrics:

├── Overall Score: 87.3/100 (Rank #1)

├── Total Cost: $234 (Best in class)

├── Bullwhip Ratio: 1.2 (Excellent coordination)

├── Service Level: 94.2% (High reliability)

└── Convergence Time: 3.4 rounds (Fast adaptation)

📊 Interactive Visualizations:

├── [Plotly] Order patterns across game rounds

├── [Plotly] Inventory levels by supply chain tier

├── [Plotly] Performance across visibility conditions

├── [Plotly] Bullwhip effect comparison

└── [Plotly] Decision consistency heatmap

🎯 Scenario Breakdown:

├── High Visibility Scenarios: 91.2/100

├── Low Visibility Scenarios: 83.4/100

├── Stable Demand: 89.7/100

├── Volatile Demand: 84.9/100

└── Shock Recovery: 88.1/100

💬 Model Details:

├── Organization: OpenAI Research

├── Type: RAG System

├── Base Model: GPT-4 Turbo

├── Knowledge Sources: Supply chain papers, case studies

└── Last Evaluated: 2 days ago

```

#### About/Documentation

```

📚 About SCM-Arena

🎯 Mission:

Advancing AI coordination capabilities in supply chain management through

standardized, rigorous evaluation of LLM-based systems.

🎮 The Beer Game Foundation:

├── Classic 4-tier supply chain simulation

├── Retailer → Wholesaler → Distributor → Manufacturer

├── Tests coordination under information asymmetry

├── Measures emergent system behavior

└── 50+ years of research validation

🔬 Evaluation Philosophy:

├── ✅ Multi-agent coordination (not individual optimization)

├── ✅ Anti-gaming design (can't memorize solutions)

├── ✅ Real-world relevance (actual supply chain scenarios)

├── ✅ Statistical rigor (multiple runs, significance testing)

└── ✅ Open methodology (reproducible evaluations)

📊 Benchmark Categories:

├── Information visibility effects

├── Memory and learning capabilities

├── Demand volatility handling

├── Cross-agent coordination quality

└── System robustness and recovery

```

### Admin Dashboard (Restricted Access)

#### Model Submission Interface

```

➕ Submit New Model for Evaluation

📝 Model Information:

├── Model Name: [ ]

├── Organization: [ ]

├── Model Type: [RAG System ▼]

├── Description: [ ]

│ [ ]

├── Paper/Blog: [https://... ]

├── Contact Email: [ ]

└── Expected Evaluation Time: ~2-3 hours

🔌 API Configuration:

├── Endpoint URL: [https://api.company.com/v1/chat]

├── Authentication:

│ ├── Type: [Bearer Token ▼]

│ └── Value: [sk-... ]

├── API Format: [OpenAI Compatible ▼]

├── Rate Limits: [100 req/min ]

└── Timeout: [30 seconds ]

🧪 API Test:

├── [Test Connection] ← Validates endpoint works

├── [Sample Request] ← Sends test beer game scenario

└── Status: ✅ API responding correctly

[Submit for Evaluation]

```

#### Evaluation Queue Monitor

```

⚡ Active Evaluations

🔄 Currently Running:

├── Claude-3-SCM-Expert

│ ├── Progress: ████████░░ 80% complete

│ ├── Scenario: 19/24 (High visibility, seasonal demand)

│ ├── ETA: 25 minutes remaining

│ └── Status: ✅ All API calls successful

│

└── Custom-Research-RAG

├── Progress: ███░░░░░░░ 30% complete

├── Scenario: 7/24 (Low visibility, random demand)

├── ETA: 1.2 hours remaining

└── Status: ⚠️ 3 timeouts (still within limits)

📋 Queue:

├── 1. Llama-70B-Supply-Chain (Submitted 15 min ago)

├── 2. GPT-4-Custom-RAG (Submitted 8 min ago)

└── 3. Mistral-Fine-tuned-SCM (Submitted 2 min ago)

✅ Recent Completions:

├── GPT-4-Supply-Chain-RAG: Score 87.3 (2 hours ago)

├── Academic-Research-Model: Score 71.8 (6 hours ago)

└── ❌ Failed-API-Model: Connection errors (1 day ago)

```

## 🚀 Implementation Plan

### Phase 1: Core Platform (Weeks 1-4)

```

Week 1-2: Backend Development

├── 🎮 Beer game engine implementation

├── 🔄 Model API integration framework

├── 📊 Core evaluation metrics

├── 🗄️ PostgreSQL schema design

└── ⚡ Celery task queue setup

Week 3: Frontend Development

├── 📈 Leaderboard display

├── 🔍 Model detail pages

├── 📊 Interactive Plotly visualizations

├── 📱 Responsive design

└── 🎨 Professional styling

Week 4: Integration & Testing

├── 🧪 End-to-end evaluation pipeline

├── 🔧 Admin dashboard completion

├── 📝 API documentation

├── 🚀 Deployment setup

└── 🧹 Bug fixes and polish

```

### Phase 2: Launch & Validation (Weeks 5-8)

```

Week 5-6: Alpha Testing

├── 🤝 Partner with 3-5 research groups

├── 📝 Submit initial model baselines:

│ ├── GPT-4 (baseline)

│ ├── Claude-3 (baseline)

│ ├── Llama-70B (baseline)

│ └── Simple algorithmic agents

├── 📊 Validate benchmark results

└── 🔧 Platform improvements

Week 7-8: Public Launch

├── 🌐 scm-arena.com domain setup

├── 📢 Announce to AI/OR communities

├── 📰 Blog post explaining benchmark

├── 📊 Publish initial leaderboard

└── 📧 Outreach to model creators

```

### Phase 3: Growth & Expansion (Months 3-6)

```

Month 3-4: Community Building

├── 🎓 Academic conference presentations

├── 📝 Research paper submission

├── 🤝 Industry partnerships

├── 📈 Monthly leaderboard updates

└── 📊 Usage analytics and insights

Month 5-6: Platform Enhancement

├── 🔧 Docker container submission support

├── 📊 Additional supply chain benchmarks

├── 🌍 Multi-language beer game scenarios

├── 📱 Mobile-optimized interface

└── 🔍 Advanced filtering and comparison tools

```

## 💰 Resource Requirements

### Technical Infrastructure

```

Development (Weeks 1-4):

├── 👨‍💻 1 Full-stack developer (you)

├── ☁️ Cloud development environment

├── 🔧 Standard dev tools and services

└── 💵 Est. cost: $500-1000

Production Hosting:

├── ⚡ FastAPI backend (Railway/Render: ~$20/month)

├── 🗄️ PostgreSQL database (Supabase: ~$25/month)

├── 🌐 Frontend hosting (Vercel: Free)

├── 📊 Redis cache (Upstash: ~$10/month)

└── 💵 Total: ~$55/month + usage

```

### API Evaluation Costs

```

Model Testing Budget:

├── 🧪 Initial baselines (GPT-4, Claude): ~$200

├── 📊 Ongoing evaluations: ~$100/month

├── 🔬 Research partnerships: Variable

└── 💵 Total: ~$300 setup + $100/month

```

### Marketing & Outreach

```

Community Building:

├── 🌐 Domain registration: $15/year

├── 📧 Email service (ConvertKit): $29/month

├── 📊 Analytics (PostHog): Free tier

├── 🎨 Design assets: $200 one-time

└── 💵 Total: ~$250 setup + $30/month

```

## 🎯 Success Metrics

### Technical Success (Month 3)

- [ ] ✅ Platform evaluates models reliably

- [ ] 📊 20+ models submitted and evaluated

- [ ] 🚀 <30 second average response time

- [ ] 🔧 99% uptime for evaluations

- [ ] 📈 Clear performance differences between models

### Community Success (Month 6)

- [ ] 🎓 5+ research groups actively using platform

- [ ] 🏢 2+ companies submitting commercial models

- [ ] 📰 Coverage in AI/OR publications

- [ ] 🌟 100+ GitHub stars on open-source components

- [ ] 💬 Active discussions in Discord/forums

### Research Impact (Year 1)

- [ ] 📝 Research paper accepted at top venue

- [ ] 📊 Cited by other benchmarking efforts

- [ ] 🎯 Standard reference for SCM AI evaluation

- [ ] 🤝 Adopted by model creators for development

- [ ] 🌍 International research collaborations

## 🔮 Future Vision

### Extended Benchmark Suite

```

Year 1 Expansion:

├── 🚛 Vehicle Routing Problems

├── 📦 Inventory Optimization Scenarios

├── 🏭 Production Planning Challenges

├── 🌐 Global Supply Chain Simulations

└── 🤖 Multi-modal (text + data) evaluations

Year 2+ Advanced Features:

├── 🎮 Real-time competitive leagues

├── 🔄 Continuous learning benchmarks

├── 🌍 Industry-specific leaderboards

├── 🤝 Human-AI collaboration scenarios

└── 📊 Benchmark-as-a-Service APIs

```

### Research Platform Evolution

```

Long-term Goals:

├── 🎯 De facto standard for SCM AI evaluation

├── 🌍 Global research community hub

├── 📊 Industry adoption for vendor selection

├── 🎓 Educational tool for OR/AI courses

└── 🚀 Foundation for next-gen supply chain AI

```

---

\*\*Contact & Next Steps:\*\*

- 📧 Get started: Begin Phase 1 implementation

- 🤝 Partnerships: Reach out to research groups

- 💬 Community: Set up Discord for discussions

- 📰 Documentation: Create detailed API specs

- 🚀 Launch: Target public release in 6-8 weeks