🎓 Quick Start: Testing & Monitoring

☑ Everything is Now Running!

Services Status:

- ☑ API (Port 8000) - Model loaded, predictions working
- ☑ Streamlit Dashboard (Port 8501) - Beautiful UI
- ☑ Prometheus (Port 9090) - Metrics collection
- ☑ Grafana (Port 3000) - Visualizations

🚀 Step 1: Make Predictions (Streamlit)

1. **Open Streamlit:**
   URL: http://localhost:8501

2. **Test Different Attack Types:**

   - Select **"DDoS Attack"** from dropdown
   - Click **" 🔍 Analyze Traffic"**
   - You should see:
     - 🚨 **ATTACK DETECTED** (red card)
     - Gauge showing 99.99% confidence
     - Probability bar chart
     - Expert weights pie chart

3. **Try All 6 Patterns:**

   - DDoS Attack → Should detect Attack
   - Port Scan → Should detect Attack (~53%)
   - Web Attack → Should detect Attack
   - Brute Force → Should detect Attack
   - Normal Traffic → Should detect Normal
   - Normal HTTPS → Should detect Normal

4. **Make 10-15 predictions** to generate good metrics data!

📊 Step 2: View Metrics (Prometheus)

1. **Open Prometheus:**
   URL: http://localhost:9090

2. **Try These Queries** (copy-paste into query box):

Query 1: Total Predictions

```
predictions_total
```

- Click **"Execute"**
- Click **"Graph"** tab
- Shows: How many Attack vs Normal predictions

Query 2: Attack Detection Rate (per second)

```
rate(predictions_total{prediction="Attack"}[1m])
```

- Shows: Attacks detected per second
- Graph shows trend over time

Query 3: Expert Weights

```
expert_gating_weight
```

- Shows: FT-Transformer vs CNN contribution
- Example: FT-Transformer usually 98%, CNN 2%

Query 4: API Request Rate

```
rate(api_requests_total[1m])
```

- Shows: Requests per second
- Useful: See API load

Query 5: Average Prediction Time

```
rate(prediction_duration_seconds_sum[5m]) /
rate(prediction_duration_seconds_count[5m])
```

- Shows: Average latency in seconds
- Example: 0.018 = 18ms

Query 6: 99th Percentile Latency

```
histogram_quantile(0.99, rate(prediction_duration_seconds_bucket[5m]))
```

- Shows: 99% of predictions faster than this
- Performance monitoring!

🎨 Step 3: Beautiful Dashboards (Grafana)

1. **Open Grafana:**
   URL: http://localhost:3000

2. **Login:**
   - Username: `admin`
   - Password: `admin`
   - (Skip password change if prompted)

3. **Add Prometheus Data Source:**
   - Click ⚙️ **Configuration** → **Data Sources**
   - Click **"Add data source"**
   - Select **"Prometheus"**
   - URL: `http://prometheus:9090`
   - Click **"Save & Test"** (should see green ✓)

4. **Create Your First Panel:**
   - Click **+** → **Dashboard** → **Add visualization**
   - Select **Prometheus** data source

   Panel 1 - Total Predictions (Stat):
   - Query: `sum(predictions_total)`
   - Visualization type: **Stat**
   - Title: "Total Predictions"
   - Click **Apply**

   Panel 2 - Attack vs Normal (Pie Chart):
   - Click **Add** → **Visualization**
   - Query: `predictions_total`
   - Visualization type: **Pie chart**
   - Legend: `{{prediction}}`
   - Title: "Predictions Distribution"
   - Click **Apply**

   Panel 3 - Request Rate (Time Series):
   - Click **Add** → **Visualization**
   - Query: `rate(api_requests_total[1m])`
   - Visualization type: **Time series**
   - Legend: `{{endpoint}}`

- Title: "API Request Rate"
- Click **Apply**

5. **Save Dashboard:**
   - Click 💾 **Save dashboard** (top right)
   - Name: "MoE Cybersecurity Monitoring"
   - Click **Save**
6. **Auto-Refresh:**
   - Top right: Click refresh interval dropdown
   - Select **5s** or **10s**
   - Now it updates in real-time!

🔥 Live Demo Exercise

Do this to see everything work together:

Part 1: Generate Traffic

1. Open **Streamlit** (http://localhost:8501)
2. Make these predictions:
   - 3x DDoS Attack
   - 2x Web Attack
   - 2x Normal Traffic
   - 1x Port Scan

Part 2: Watch Prometheus

1. Open **Prometheus** (http://localhost:9090)
2. Query: `predictions_total`
3. Click **Execute**
4. You should see:

```
predictions_total{prediction="Attack"} 6
predictions_total{prediction="Normal"} 2
```

Part 3: See Grafana Update

1. Open **Grafana** (http://localhost:3000)
2. If you created the pie chart, watch it update!
3. The pie should show:
   - ~75% Attack (red)
   - ~25% Normal (green)

📈 Cool Things to Notice

In Prometheus:

- **Metrics update every 10 seconds** (scrape interval)
- **Graphs show trends** - you can see spikes when you make predictions
- **PromQL is powerful** - you can do math, aggregations, percentiles

In Grafana:

- **Auto-refresh** - dashboards update live
- **Beautiful visualizations** - much prettier than Prometheus
- **Multiple panels** - combine different metrics in one view
- **Alerting** - can set up alerts (e.g., if attacks > 50%)

In Streamlit:

- **Each prediction** triggers:
  - API call to `/predict`
  - Prometheus metrics updated
  - Counters incremented
  - Gauges set to new values

🎯 Understanding the Flow

```
1. You click "Analyze Traffic" in Streamlit
   ↓
2. Streamlit sends features to API: POST /predict
   ↓
3. API makes prediction with MoE model
   ↓
4. API updates Prometheus metrics:
   - predictions_total{prediction="Attack"} +1
   - prediction_confidence = 0.9999
   - expert_gating_weight{expert="Tabular Expert"} = 0.984
   ↓
5. Streamlit shows beautiful results (gauge, charts)
   ↓
6. Prometheus scrapes API /metrics endpoint (every 10s)
   ↓
7. Grafana queries Prometheus (every 5s if auto-refresh)
   ↓
8. You see updated dashboards!
```

🐛 Troubleshooting

"No data in Prometheus"

- Make predictions first!
- Wait 10-15 seconds for scrape
- Check: http://localhost:8000/metrics (should see metrics)

"Grafana shows no data"

- Make sure Prometheus data source is configured
- URL must be: `http://prometheus:9090` (not localhost!)
- Make predictions to generate data
- Check time range (top right) - set to "Last 15 minutes"

"Streamlit button doesn't work"

- Check API is running: http://localhost:8000/health
- Should see: `"model_loaded": true`
- Check browser console for errors (F12)

🎓 What You're Learning

☑ **Real-time ML Monitoring** - Track model behavior in production
☑ **Prometheus** - Industry-standard metrics collection
☑ **PromQL** - Query language for time-series data
☑ **Grafana** - Professional dashboarding tool
☑ **Metrics Types**:

- **Counter**: Always increases (predictions_total)
- **Gauge**: Goes up/down (confidence, expert weights)
- **Histogram**: Distribution (latency buckets)

🚀 Next: Advanced Monitoring

Try these next:

1. **Set up alerts** - Email when attack rate > threshold
2. **Add more panels** - Feature importance, data drift
3. **Long-term trends** - Daily/weekly attack patterns
4. **Performance tuning** - Optimize based on latency metrics

You're now monitoring like a pro! 🎉

📚 Quick Reference

Service	URL	Credentials
Streamlit	http://localhost:8501	-

Service	URL	Credentials
API Docs	http://localhost:8000/docs	-
Prometheus	http://localhost:9090	-
Grafana	http://localhost:3000	admin/admin
API Metrics	http://localhost:8000/metrics	-
API Health	http://localhost:8000/health	-

Happy Monitoring! 🌓 📊