Free Models AI Negotiation System Setup

🎯 What Changed

I've completely redesigned the system to use **free, open-source models** instead of OpenAI:

- **Qwen3** (Alibaba's latest, excellent reasoning)
- **Llama 3.3** (Meta's latest, great for conversations)
- **Mistral** (Strong reasoning and multilingual support)

All models run locally through **Ollama** - no API keys needed!

🚀 Quick Setup (5 Minutes)

Step 1: Install Ollama

```bash
# Visit https://ollama.com/download and install for your OS
# Or use these quick commands:

# macOS
brew install ollama

# Linux
curl -fsSL https://ollama.com/install.sh | sh

# Windows: Download from website
```

Step 2: Start Ollama Server

```bash
ollama serve
# Keep this running in a separate terminal
```

Step 3: Pull Recommended Models

```bash
```

```
# Get the best models for negotiation (choose 1-2)
ollama pull qwen3:8b        # 🏆 Best reasoning (5GB)
ollama pull llama3.3:8b     # 🏆 Best conversation (5GB)
ollama pull mistral:7b      # 🏆 Balanced performance (4GB)

# For lower-end hardware:
ollama pull qwen3:4b        # Smaller but still great (2.5GB)
ollama pull llama3.2:3b     # Very fast (2GB)
```

Step 4: Install Python Dependencies

```
bash

pip install ollama requests asyncio
```

Step 5: Check Everything Works

```
bash

python negotiation_agents.py --check
```

🎛️ Model Recommendations

For Best Results (8GB+ RAM):

- **Agent USA**: `qwen3:8b` (excellent reasoning)
- **Agent Russia**: `llama3.3:8b` (great conversation)

For Balanced Performance (4-8GB RAM):

- **Both Agents**: `mistral:7b` (good all-around)

For Lower-End Hardware (2-4GB RAM):

- **Both Agents**: `qwen3:4b` or `llama3.2:3b`

▶️ Running Negotiations

Basic Run

```
bash

python negotiation_agents.py
```

Check Prerequisites Only

```bash
python negotiation_agents.py --check
```

📊 Expected Output

```
🚀 Starting AI Agent Negotiation System (Free Models)
============================================================
✅ Ollama server is running
📋 Available models: ['qwen3:8b', 'llama3.3:8b']
🤖 Using models: USA=qwen3:8b, Russia=llama3.3:8b
🔄 Initializing RAG system...
🤖 Creating negotiation agents...
📋 Setting up negotiation on Ukraine conflict resolution...

🎯 Starting negotiation...
Topic: Ukraine conflict resolution and future security arrangements
Participants: Agent_USA, Agent_Russia
Models: qwen3:8b vs llama3.3:8b

=== Round 1 ===

Agent_USA [proposal]:
I appreciate this opportunity to discuss a framework for resolving the Ukraine conflict. The United States position centers on several core principles: respect for Ukraine's sovereignty and territorial integrity, adherence to international law, and ensuring that any resolution doesn't reward aggression...

Agent_Russia [counter_proposal]:
Thank you for outlining the US position. Russia's perspective focuses on legitimate security concerns that have driven this conflict. We must address NATO's eastward expansion, establish neutral status for Ukraine, and recognize the security architecture that existed before 2014...

=== Round 2 ===
...
```

🔧 Troubleshooting

Common Issues

1. "Ollama server not running"

```bash
# In a separate terminal:
ollama serve

# Then run your script in another terminal
python negotiation_agents.py
```

2. "No suitable models found"

```bash
# Pull at least one model:
ollama pull qwen3:8b

# Check what's available:
ollama list
```

3. "Models too slow"

```bash
# Use smaller models:
ollama pull qwen3:4b
ollama pull llama3.2:3b

# Or reduce max_rounds in the script
```

4. "Out of memory"

```bash
# Use smaller models
ollama pull qwen3:4b

# Or close other applications
```

5. "Qdrant connection failed"

```bash
```

```
# Make sure your Qdrant databases are running:
docker ps

# Restart if needed:
docker restart <container_id>
```

💡 Performance Tips

Hardware Requirements:

- **Minimum**: 4GB RAM, any CPU
- **Recommended**: 8GB+ RAM, GPU optional
- **Optimal**: 16GB RAM, dedicated GPU

Speed Optimization:

```
bash

# Use quantized models (smaller, faster):
ollama pull qwen3:4b-q4_0
ollama pull llama3.2:3b-q4_0

# Reduce rounds for faster testing:
# Edit max_rounds=3 in negotiation_agents.py
```

Quality vs Speed:

- **Quality**: `qwen3:8b` + `llama3.3:8b`
- **Balanced**: `mistral:7b` for both
- **Speed**: `qwen3:4b` + `llama3.2:3b`

🎯 Key Advantages of Free Models

✅ Benefits:

- **No API costs** - run unlimited negotiations
- **Privacy** - all data stays on your machine
- **Customizable** - modify models and prompts
- **Offline capable** - no internet required after setup
- **Latest models** - Qwen3 and Llama 3.3 are cutting-edge

🔄 vs OpenAI:

- **Cost**: Free vs $$$
- **Speed**: Similar (depends on hardware)
- **Quality**: Very competitive, especially Qwen3
- **Privacy**: Much better (local)
- **Availability**: No rate limits

📈 Model Performance Comparison

Model	Size	Quality	Speed	Memory	Best For
qwen3:8b	5GB	⭐⭐⭐⭐⭐	⭐⭐⭐	8GB+	Reasoning, facts
llama3.3:8b	5GB	⭐⭐⭐⭐⭐	⭐⭐⭐	8GB+	Conversation, strategy
mistral:7b	4GB	⭐⭐⭐⭐	⭐⭐⭐⭐	6GB+	Balanced
qwen3:4b	2.5GB	⭐⭐⭐⭐	⭐⭐⭐⭐	4GB+	Fast reasoning
llama3.2:3b	2GB	⭐⭐⭐	⭐⭐⭐⭐⭐	3GB+	Very fast

🔧 Advanced Configuration

Custom Models Per Agent

```python
# Edit in negotiation_agents.py setup_system():
usa_model = "qwen3:8b"      # Detail-oriented
russia_model = "llama3.3:8b"  # Strategic

# Or use same model for both:
usa_model = russia_model = "mistral:7b"
```

Adjust Model Parameters

```python
```

```
# In OllamaClient.generate_response():
"options": {
    "temperature": 0.7,    # Creativity (0.1-1.0)
    "num_predict": 1000,   # Max response length
    "top_p": 0.9,          # Nucleus sampling
    "repeat_penalty": 1.1  # Avoid repetition
}
```

Custom Negotiation Topics

python

```
# Change in main():
initial_prompt = """
Negotiate a climate change agreement between
major industrial nations...
"""
```

🎯 Next Steps

1. ✅ **Setup Complete** - Run your first negotiation
2. 📊 **Analyze Results** - Check the JSON logs
3. 🔧 **Tune Parameters** - Adjust models and settings
4. 🎭 **Try Different Topics** - Energy, trade, cybersecurity
5. 📈 **Scale Up** - Add more agents or longer negotiations

Ready to watch AI agents negotiate with completely free models! 🎉

No API keys, no costs, no limits - just pure AI negotiation powered by the latest open-source models.