**✅ Files Ready to Use:**

* complete\_file\_structure.md - Good documentation
* conda\_environment.txt - Good for conda users
* conda\_env\_variants.txt - Good alternatives
* makefile\_conda.txt - Good for conda operations
* requirements\_structure.txt - Good requirements templates
* setup\_script.sh - Good setup script

**⚠️ Files with Minor Encoding Issues:**

* command\_reference.md - Has corrupted emojis (ðŸ"Š → 📊)
* auto\_setup.sh - Has corrupted characters
* setup\_readme.md - Has corrupted emojis

**But these encoding issues won't stop the system from working!**

**🚀 Super Simple Quick Start (3 Steps)**

**Step 1: Create Missing Pieces**

bash

*# Create directories*

mkdir -p results dataset\_cache sampled\_datasets logs

*# Create requirements.txt (copy from requirements\_structure.txt)*

cat > requirements.txt << 'EOF'

torch>=2.5.0

transformers>=4.46.0

vllm>=0.6.4

fastapi>=0.115.0

uvicorn[standard]>=0.32.0

promptbench>=0.0.4

datasets>=2.19.0

pandas>=1.3.5

numpy>=1.26.0

pyyaml>=6.0.0

requests>=2.32.0

tqdm>=4.66.0

click>=8.0.0

rich>=13.0.0

EOF

**Step 2: Install & Configure**

bash

*# Install Python packages*

pip install -r requirements.txt

*# Edit config.yaml - PUT YOUR REAL MODEL PATH HERE!*

vim config.yaml

*# Change this line:*

*# path: "/path/to/llama-3.1-70b-Instruct"*

*# To your actual path like:*

*# path: "/home/ubuntu/models/llama-3.1-70b-Instruct"*

**Step 3: Test It!**

bash

*# Terminal 1: Start server*

python model\_manager.py --server

*# Terminal 2: Run quick test*

python bulk\_tester.py --suite quick\_test

*# OR use Makefile (easier):*

make quick

**🎯 That's It! You're Done!**

**What You Should See:**

* Server starts on port 8000 ✅
* Model loads (takes 1-2 minutes) ✅
* Tests run on datasets ✅
* Results appear in results/ folder ✅
* HTML report generated ✅

**If Something Goes Wrong:**

1. **"Module not found"** → Run: pip install [missing\_module]
2. **"Model path not found"** → Check your path in config.yaml
3. **"CUDA error"** → Run without GPU: pip install torch --index-url https://download.pytorch.org/whl/cpu
4. **"Port 8000 in use"** → Kill it: lsof -ti:8000 | xargs kill -9

**Verify Everything Works:**

bash

*# Quick check*

python -c "import model\_manager, bulk\_tester, llm\_cli; print('✓ All modules OK')"

*# If that works, you're golden! 🏆*

**Bottom Line:** Your files are good enough to run! Don't worry about the encoding issues in documentation files - they won't affect the actual code execution. Just follow the 3 steps above and you'll be testing LLMs in 5 minutes! 🚀