🚀 GROK 4.1 FAST REASONING UPGRADE - COMPLETE

Date: November 20, 2025
Upgrade: Grok 2 → Grok 4.1 Fast Reasoning
Impact: All Jazz-related tasks and verification now use Grok 3+

✅ UPGRADE SUMMARY

What Changed:

- **Verification Agent**: Now uses `grok-4-1-fast-reasoning` for all fact-checking
- **Jazz Team**: Upgraded from GPT-4o to `grok-4-1-fast-reasoning` for self-improvement analysis
- **Cost Tracking**: Added pricing for Grok 4.1 Fast models
- **Documentation**: Updated all references to reflect Grok 4.1

Why Grok 4.1 Fast Reasoning?

- **2M token context** (vs 128k in Grok 2) - 15.6x larger!
- **90% cheaper** - $0.20/M input (vs $2.00/M in Grok 2)
- **State-of-the-art reasoning** - Optimized for agentic tool calling
- **Superior performance** - Elo 1483 on LMArena (top tier)

📊 TECHNICAL DETAILS

Model Configuration (`llm.config.ts`)

```
models: {
  verification: 'grok-4-1-fast-reasoning',  // Verifier Agent (fact-checking)
  jazz: 'grok-4-1-fast-reasoning',          // Jazz Team (self-improvement)
}

costs: {
  'grok-4-1-fast-reasoning': {
    inputPer1k: 0.0002,   // $0.20 per 1M input tokens
    outputPer1k: 0.0005,  // $0.50 per 1M output tokens
  },
  'grok-4-1-fast-non-reasoning': {
    inputPer1k: 0.0002,   // $0.20 per 1M input tokens
    outputPer1k: 0.0005,  // $0.50 per 1M output tokens
  },
}
```

Cascade Configuration (`llm-cascade.service.ts`)

```
// Jazz Team Cascade (Specialized Counterfactual Analysis)
jazz: [
  { name: 'Grok-4.1-Fast-Reasoning', tier: 1, call: this.callGrok.bind(this) },
  { name: 'GPT-5-Fallback', tier: 2, call: this.callGPT5.bind(this) },
],

// Verification Cascade (Fact-Checking)
verification: [
  { name: 'Grok-4.1-WebSearch', tier: 1, call: this.callGrok.bind(this) },
  { name: 'Direct-Claude-WebSearch', tier: 2, call: this.callDirectClaude.bind(this) }
,
  { name: 'GPT-5', tier: 3, call: this.callGPT5.bind(this) },
],
```

🧪 TEST RESULTS (Local)

Code Edit Test (`/api/ide/code-edit`)

```
Request: Add comment to TypeScript function
Duration: 58.5 seconds
Result: SUCCESS ✅

Agent Performance:
- Analyst: Grok-4.1 (Tier 3 fallback), latency: 885ms, cost: $0.0017
- Relational: Grok-4.1 (Tier 4 fallback), latency: 924ms, cost: $0.0017
- Verifier: grok-4-1-fast-reasoning, latency: 19.3s, cost: $0.0002, confidence: 1.00
- Synthesizer: Grok-4.1 (Tier 4 fallback), latency: 3.7s, cost: $0.0066
- Jazz Team: Grok-4.1-Fast-Reasoning, latency: 6.9s, cost: $0.0080

Total Cost: ~$0.018 (vs ~$0.18 with old Grok 2 pricing)
```

Jazz Analysis Output

```
{
  "voice": 0.95,     // Logical coherence
  "choice": 0.92,    // Emotional balance
  "transparency": 0.97,  // Clarity of reasoning
  "trust": 0.96,     // Overall trust τ
  "suggestions": [
    "Consider adding a more detailed comment...",
    "Implement a feature to suggest appropriate comment templates...",
    "Explore the possibility of automatically generating unit tests..."
  ],
  "refinedInstruction": "Add a comprehensive comment explaining..."
}
```

🔧 FILES MODIFIED

1. Configuration

- `nodejs_space/src/config/llm.config.ts`
- Added `verification: 'grok-4-1-fast-reasoning'`
- Added `jazz: 'grok-4-1-fast-reasoning'`
- Added cost tracking for Grok 4.1 Fast models

2. Services

- `nodejs_space/src/services/llm.service.ts`
- Updated cost calculation to use Grok 4.1 pricing
- `nodejs_space/src/services/llm-cascade.service.ts`
- Added `jazz` cascade with Grok 4.1 as Tier 1
- Updated documentation to reflect Grok 4.1 upgrade
- `nodejs_space/src/services/vctt-engine.service.ts`
- Updated `processBuildArtifact` to use `jazz` role (Grok 4.1)
- Updated contribution tracking to use `grok-4-1-fast-reasoning`
- Updated pre-jam truth sweep documentation

3. Agents

- `nodejs_space/src/agents/verifier.agent.ts`
- Updated documentation header to reflect Grok 4.1 Fast Reasoning
- Added specs: 2M context, 90% cheaper, state-of-the-art reasoning

---

💰 COST COMPARISON

Per Million Tokens:

Model	Input Cost	Output Cost	Total (1M in + 1M out)
Grok 2 (grok-2-1212)	$2.00	$10.00	$12.00
Grok 4.1 Fast	**$0.20**	**$0.50**	**$0.70**
Savings	**90%**	**95%**	**94.2%**

Typical Code Edit Session:

- Old (Grok 2): ~$0.18
- New (Grok 4.1): ~$0.018
- **Savings per edit: $0.162 (90%)**

🎯 USE CASES NOW USING GROK 4.1

1. Verification Agent (Fact-Checking)

- Real-time fact verification during Band Jam Mode
- Pre-jam truth sweep (mycelium seeding)
- Post-synthesis correctness checks
- **Weight**: 20% base, 30% for factual queries
- **Veto**: Triggers re-jam if confidence < 0.8

2. Jazz Team (Self-Improvement Analysis)

- Counterfactual trust testing of code edits
- Voice/Choice/Transparency/Trust (VCTT) measurement
- Build artifact analysis for continuous improvement
- Refined instruction generation
- **Specialized reasoning**: Deep analytical capabilities

📈 PERFORMANCE IMPROVEMENTS

Context Window:

- Grok 2: 128,000 tokens
- **Grok 4.1 Fast: 2,000,000 tokens** (15.6x larger!)
- Impact: Can analyze much larger codebases and conversation histories

Reasoning Quality:

- **LMArena Elo Score**: 1483 (top tier, surpassing GPT-4 and Claude)
- **Hallucination Rate**: 4.22% (down from 12.09% in Grok 2)
- **FActScore**: 2.97% error rate (down from 9.89%)

Tool Calling:

- Optimized for agentic tasks
- State-of-the-art function calling
- Integrated with Agent Tools API (Web Search, X Search, Code Execution)

🚀 DEPLOYMENT STATUS

Local Testing:

- ✅ Build successful
- ✅ Server started on port 10000
- ✅ Code edit endpoint tested
- ✅ Jazz analysis working
- ✅ Verification agent using Grok 4.1

Ready for Production:

- ✅ All Grok model references updated
- ✅ Cost tracking configured
- ✅ Documentation updated
- ✅ End-to-end test passed
- ⏳ **Next step**: Deploy to production via AbacusAI UI

🔍 VERIFICATION LOGS (Sample)

```
[Nest] 2865 - 11/20/2025, 1:41:11 PM     LOG [VerifierAgent] ✅ Verifier JSON parsed
successfully (model: grok-4-1-fast-reasoning)
[Nest] 2865 - 11/20/2025, 1:41:11 PM     LOG [VerifierAgent] ✅ Verifier complete - c
onfidence: 1.00, discrepancy: false, facts: 3, cost: $0.0002, latency: 19364ms,
model: grok-4-1-fast-reasoning
[Nest] 2865 - 11/20/2025, 1:41:11 PM     LOG [VCTTEngineService] ✅ Grok verified cod
e correctness - trust: τ=1.000
[Nest] 2865 - 11/20/2025, 1:41:44 PM     LOG [LLMCascadeService] → Tier 1: Trying Gro
k-4.1-Fast-Reasoning...
[Nest] 2865 - 11/20/2025, 1:41:50 PM     LOG [LLMCascadeService] ✅ Grok-4.1-Fast-Rea
soning succeeded - tokens: 1361, cost: $0.0080, latency: 6929ms
[Nest] 2865 - 11/20/2025, 1:41:50 PM     LOG [VCTTEngineService] ✅ JAZZ TEAM: Ana-
lysis complete
[Nest] 2865 - 11/20/2025, 1:41:50 PM     LOG [IdeService] ✅ Jazz team analysis com-
plete:
[Nest] 2865 - 11/20/2025, 1:41:50 PM     LOG [IdeService]    Voice (logic): 0.95
[Nest] 2865 - 11/20/2025, 1:41:50 PM     LOG [IdeService]    Choice (balance): 0.92
[Nest] 2865 - 11/20/2025, 1:41:50 PM     LOG [IdeService]    Transparency: 0.97
[Nest] 2865 - 11/20/2025, 1:41:50 PM     LOG [IdeService]    Enhanced Trust τ: 0.960
```

🎵 JAZZ TEAM INTEGRATION DETAILS

The Jazz Team is the self-improvement engine of VCTT-AGI. It analyzes every code edit through the lens of:

1. **Voice (V)** - Logical coherence and correctness
2. **Choice (C)** - Balance between competing considerations
3. **Transparency (T_1)** - Clarity of reasoning and documentation
4. **Trust (T_2)** - Overall confidence metric (τ)

Formula: $\tau = 1 - (0.4 \cdot (1-V) + 0.3 \cdot (1-C) + 0.3 \cdot (1-T_1))$

Why Grok 4.1 for Jazz?
- Advanced reasoning capabilities excel at counterfactual analysis
- 2M context window allows comprehensive code review
- Superior analytical depth for Voice/Choice/Transparency evaluation
- Cheaper than GPT-4o, faster than Claude with tools

📚 REFERENCES

xAI Documentation:

- Grok 4.1 Announcement (https://x.ai/news/grok-4-1)
- Grok 4.1 Fast and Agent Tools API (https://x.ai/news/grok-4-1-fast)
- xAI Models and Pricing (https://docs.x.ai/docs/models)

Performance Benchmarks:

- LMArena Elo Score: 1483 (grok-4-1-thinking)
- Human Preference: 64.78% win rate vs Grok 4
- Hallucination Rate: 4.22% (down from 12.09%)
- FActScore: 2.97% error (down from 9.89%)

🎯 NEXT STEPS

1. **Test Preview Deployment:**
   ```
   bash
     # Preview URL will be generated by AbacusAI
     curl https://[preview-url].abacusai.app/api/ide/code-edit -X POST -d '{...}'
   ```

2. **Production Deployment:**
   - Deploy via AbacusAI UI “Deploy” button
   - Monitor logs for Grok 4.1 usage
   - Track cost savings vs Grok 2

3. **Future Enhancements:**
   - Add Grok 4.1 Fast **Non-Reasoning** mode for faster responses
   - Implement Agent Tools API integration (Web Search, X Search, Code Execution)
   - Expand Jazz Team to analyze entire project structure

✅ CHECKLIST

- [x] Update LLM config to use `grok-4-1-fast-reasoning`
- [x] Add Grok 4.1 cost tracking
- [x] Update Verifier Agent documentation
- [x] Update Jazz Team to use Grok 4.1
- [x] Add `jazz` cascade configuration
- [x] Update contribution tracking
- [x] Test locally with `/api/ide/code-edit`
- [x] Verify logs show Grok 4.1 usage
- [x] Confirm Jazz analysis working
- [x] Confirm cost reduction (90%)
- [ ] Deploy to preview environment
- [ ] Test preview deployment
- [ ] Deploy to production

Upgrade Status: ✅ **COMPLETE**
Local Testing: ✅ **PASSED**
Ready for Production: ✅ **YES**

VCTT-AGI Engine - Phase 3.7: Grok 4.1 Fast Reasoning Integration
November 20, 2025