Ollama Model Download Enhancement

Overview

Enhanced the installation script to automatically download and verify all required AI models during the Ollama installation phase. This ensures the application is fully functional immediately after installation.

Changes Made

1. Enhanced `install_ollama()` Function

- **Improved Service Readiness Check**: Replaced simple `sleep 5` with active polling of Ollama API endpoint
- **Timeout Handling**: Added 30-second timeout with proper error handling
- **Model Download Integration**: Calls new `download_ollama_models()` function

2. New `download_ollama_models()` Function

A comprehensive model download function with the following features:

Required Models

The function downloads all four models required by the application:

1. **llama3.2:3b** - Primary model
   - Used by: Enhanced Chat API, Tool Chat API, Log Analysis API
   - Purpose: Main AI capabilities for chat and analysis

2. **phi3:mini** - Lightweight model
   - Used by: General Chat API
   - Purpose: Fast responses for general chat interface

3. **llama2** - Backup model
   - Used by: Device diagnostics (as per install-local-ai.sh)
   - Purpose: Fallback for device-related queries

4. **mistral** - Fast model
   - Used by: Quick queries (as per install-local-ai.sh)
   - Purpose: Rapid responses for simple queries

Key Features

Progress Indicators
- Shows current model number (e.g., “Model 2 of 4”)
- Displays model name and purpose
- Visual separators for clarity
- Real-time download status

Retry Logic
- 3 attempts per model
- 5-second delay between retries

- Clear retry attempt messages
- Graceful handling of network failures

Verification
- Checks if model already exists before downloading
- Verifies successful download with `ollama list`
- Double-checks model availability after pull
- Tracks failed models for reporting

Error Handling
- Non-blocking failures (installation continues even if models fail)
- Detailed troubleshooting guidance
- Lists all failed models at the end
- Provides manual recovery steps

User Experience
- Clear, colorful output with visual separators
- Estimated time information (10-30 minutes)
- Summary of installed models
- Feature-to-model mapping in success message

Technical Details

Sequential Downloads

Models are downloaded one at a time to:
- Avoid resource contention
- Ensure reliable downloads
- Provide clear progress tracking
- Simplify error handling

Verification Process

Each model goes through:
1. Pre-check: Skip if already installed
2. Download: Pull with retry logic
3. Verification: Confirm with `ollama list`
4. Reporting: Add to success or failure list

Error Recovery

If models fail to download:
- Installation continues (non-blocking)
- User receives clear warning
- Troubleshooting steps provided
- Models can be downloaded manually later

Benefits

1. **Complete Installation**: All AI features work immediately after installation
2. **Reliability**: Retry logic handles temporary network issues
3. **Transparency**: Users see exactly what’s being downloaded and why
4. **Robustness**: Graceful handling of failures with recovery guidance

5. **User-Friendly**: Clear progress indicators and helpful messages

Testing Recommendations

1. **Fresh Install**: Test on system without Ollama
2. **Existing Install**: Test with Ollama already installed
3. **Partial Models**: Test with some models already present
4. **Network Issues**: Test retry logic with simulated failures
5. **Disk Space**: Test behavior with insufficient space

Future Enhancements

Potential improvements for future versions:
- Parallel downloads with proper resource management
- Model size information before download
- Download progress bars (if Ollama API supports it)
- Configurable model list via environment variables
- Optional model downloads (core vs. optional)

Related Files

- `install.sh` - Main installation script (lines 341-523)
- `install-local-ai.sh` - Original model installation reference
- `src/app/api/chat/route.ts` - Uses phi3:mini
- `src/app/api/ai/enhanced-chat/route.ts` - Uses llama3.2:3b
- `src/app/api/ai/tool-chat/route.ts` - Uses llama3.2:3b
- `src/app/api/ai-assistant/analyze-logs/route.ts` - Uses llama3.2:3b

Commit Message

```
feat: enhance Ollama installation with complete model downloads

- Add download_ollama_models() function for comprehensive model management
- Download all 4 required models: llama3.2:3b, phi3:mini, llama2, mistral
- Implement retry logic (3 attempts) for network resilience
- Add verification checks for each downloaded model
- Improve service readiness check with active polling
- Provide clear progress indicators and troubleshooting guidance
- Ensure non-blocking failures with helpful error messages

This ensures all AI features are immediately functional after installation,
eliminating the need for manual model downloads.
```