NebulaByte: LLM API Integration

NebulaByte LLM API Integration Guide

Working with LLM APIs

Modern applications leverage Large Language Models through APIs provided by various vendors. Key considerations:

Major Providers:
1. OpenAI: GPT-4, GPT-3.5 with strong reasoning
2. Anthropic: Claude with long context windows
3. Google: Gemini with multimodal capabilities
4. Groq: Ultra-fast inference speeds
5. Cohere: Specialized for enterprise use

API Integration Best Practices:
- Use environment variables for API keys
- Implement rate limiting and retries
- Handle errors gracefully
- Monitor usage and costs
- Cache responses when appropriate

Prompt Engineering:
Effective prompts include clear instructions, relevant context, examples when needed, and structured output formats. Consider system prompts for consistent behavior.

Cost Optimization:
- Choose appropriate model sizes
- Use streaming for real-time responses
- Batch requests when possible
- Implement caching strategies
- Monitor token usage

Security:
Never hardcode API keys, use secret management systems, validate inputs, sanitize outputs, and implement usage limits.