EcoMind: Technical Documentation

Team beszketnyky | BHL Hackathon 2025

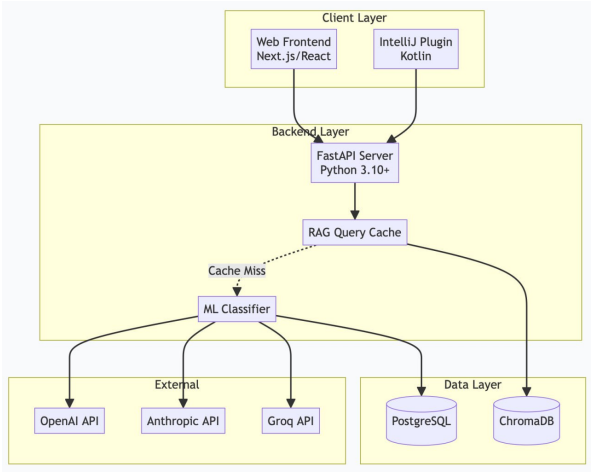
1. Overview

EcoMind is an intelligent AI model routing system that optimizes cost and carbon emissions by automatically selecting the most efficient LLM for each task based on complexity analysis. The system achieves **60-90% cost savings** and **70-95% CO2 reduction** through three-layer optimization: RAG caching, complexity-based routing, and intelligent model selection.

2. System Architecture

2.1 Architecture Overview

EcoMind follows a layered architecture with clear separation of concerns:



2.2 Request Workflow

1. **RAG Cache Check**: Query ChromaDB for similar queries (similarity threshold: 0.8). If match found, return cached answer (zero cost/emissions).

2. **Complexity Analysis**: On cache miss, analyze query using GPT-4o-mini to classify complexity (1-10 scale).

3. **Model Selection**: Route to optimal model based on complexity:

   - Complexity 1-2: Ultra-efficient (GPT-5-nano, Groq Llama-8B)
   - Complexity 3-4: Lightweight (GPT-3.5-turbo, Claude-3-Haiku)
   - Complexity 5-6: Balanced (GPT-4o-mini, Claude-Haiku-4.5)
   - Complexity 7-8: Advanced (GPT-4o, Claude-Sonnet-4)
   - Complexity 9-10: Premium (Claude-Opus-4.5, GPT-o1)

4. **Response & Caching**: Process query, cache result in ChromaDB, track savings.

3. Technology Stack

3.1 Backend

- **Framework**: FastAPI (Python 3.10+)
- **LLM Integration**: LangChain (Anthropic, OpenAI, Groq)
- **Vector Database**: ChromaDB with HuggingFace embeddings
- **Relational Database**: PostgreSQL
- **Embeddings**: sentence-transformers/all-MiniLM-L6-v2
- **Complexity Analysis**: GPT-4o-mini (low-cost classifier)

3.2 Frontend

- **Web**: Next.js 14+ (React 18+, TypeScript), Tailwind CSS, shadcn/ui
- **IDE Plugin**: IntelliJ Platform SDK (Kotlin), Swing UI components

4. Machine Learning Analysis

4.1 Dataset

- **Size**: 1,500 text prompts across 4 task categories
- **Classes**: code_task (40%), text_generation (30%), explanation_task (20%), text_summarization (10%)
- **Preprocessing**: TF-IDF vectorization with bigrams (max_features=3000, ngram_range=(1,2))
- **Split**: 80/20 train-test split with stratification

4.2 Model Selection & Validation

Four models were evaluated: Naive Bayes, Logistic Regression, Random Forest, and SVM. Cross-validation scores aligned with test performance, indicating excellent generalization.

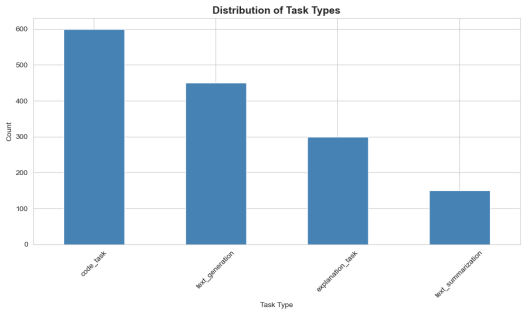
Key Findings:

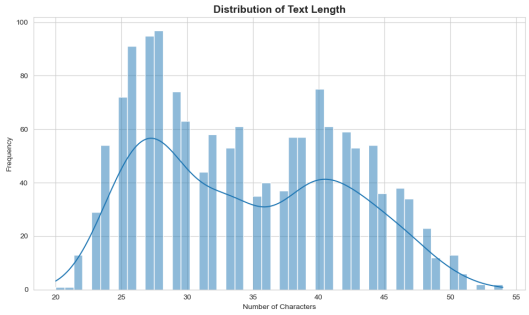
- TF-IDF with bigrams effectively captures task-specific patterns
- Feature matrix: 1,500 samples × 474 features after vectorization
- Zero misclassifications on test set (300 samples)
- Naive Bayes selected as production model (simplest, fastest)

4.3 Visualizations

- **Class Distribution**: Bar chart showing balanced distribution across 4 task types

- **Model Comparison**: Bar charts comparing test accuracy and F1-scores
- **Confusion Matrix**: Perfect diagonal matrix
- **Text Statistics**: Box plots showing text length distribution by task type





5. IntelliJ Plugin (Chatbot Extension)

The **chatbot-extension** provides native IDE integration for EcoMind:

- **Technology**: Kotlin, IntelliJ Platform SDK, Swing UI
- **Features**: Tool window with chat interface, real-time streaming responses from Claude CLI, project context awareness, status indicators (Ready/Thinking/Error)
- **Integration**: Communicates with FastAPI backend via HTTP, supports Claude CLI for direct model access
- **UI**: JetBrains-style design with HTML rendering for message display

6. Performance Metrics

Models Supported	22 AI models across 3 providers
Cost Savings	60-90% reduction in AI spend
Carbon Reduction	70-95% CO2 emission reduction
Response Time	2-5x faster for simple tasks
Cache Hit Rate	Similarity threshold 0.8 (80% similarity required)

7. Key Components

- **ComplexityAgent**: Analyzes query complexity using GPT-4o-mini with conservative rating guidelines
- **PromptRetriever**: Semantic search in ChromaDB using cosine similarity
- **ModelFactory**: Dynamic model selection based on complexity and provider preferences
- **SavingsTracker**: Real-time cost and CO2 emission tracking