CONFIDENTIAL PRODUCT DESCRIPTION FOR IP COUNSEL

Product Name: LLMscope
Version: v0.6.0-dev
Developer: Brandan Baker (BLB3D Labs, Sole Proprietorship)

Purpose & Function

LLMscope is a self-hosted AI performance monitoring platform designed for developers and organizations using large language models (LLMs) such as OpenAI, Anthropic, or locally hosted models like Ollama. It measures response latency, reliability, and cost trends — applying Statistical Process Control (SPC) and Nelson Rules to detect abnormal model behavior in real time.

Core Innovation

Unlike generic observability tools, LLMscope integrates statistical quality-control methods (μ, σ, control limits, Nelson Rules 1–8) directly into AI latency analytics. It provides real-time anomaly detection for model reasoning strain ('cognitive load') — a novel metric class for LLMs. The system is fully self-hosted, preserving privacy, and deploys in under 10 minutes using Docker Compose.

Emergent Discovery: Cognitive Load Analysis

During early testing of LLM latency variance, the developer identified a repeatable correlation between model reasoning intensity and response-time deviation — termed **Cognitive Load Latency**. This represents a measurable, statistically verifiable dimension of AI reasoning strain. LLMscope integrates this finding into its analytics engine to quantify cognitive load in real time through SPC-based anomaly detection. This concept and its implementation appear to be **novel and proprietary**, with potential applicability for patent or trade secret protection.

Architecture Overview

• **Backend (FastAPI + SQLite):** Receives telemetry data, calculates SPC metrics, stores historical latency and token data, and exposes REST endpoints (/api/log, /api/stats/spc).
• **Frontend (React + Plotly):** Visual dashboard rendering SPC charts, Nelson rule violations, and system telemetry.
• **Monitor Service (Python asyncio):** Periodically queries AI models, measures latency and token usage, and posts results to the backend.

Key Differentiators

• Statistical engine tailored for AI latency, not manufacturing data.
• Cross-provider comparison layer (OpenAI, Anthropic, Ollama).
• Lightweight local deployment for privacy-sensitive users.
• Visual SPC dashboard optimized for accessibility (control bands, violations, Cp/Cpk).

Deliverables & Current Status

• Fully functional prototype under local Docker orchestration.
• Frontend served via Nginx (port 8081); backend via FastAPI (port 8000).
• SQLite data persistence and API key authentication.
• Target beta release: Q4 2025.

Requested Legal Guidance

1. Protecting statistical analysis and SPC visualization logic (potential utility/design patent).
2. Registering copyright for original codebase (Python backend, React frontend, monitor scripts).
3. Trade secret protection for 'cognitive load' analytics and anomaly detection logic.
4. IP assignment/licensing strategy prior to commercialization (self-hosted SaaS/on-prem).

Business Entity Status

The developer currently operates as a **sole proprietorship** under the name BLB3D Labs. No LLC or corporation has been established yet. Guidance is requested on whether forming an LLC or other entity would provide stronger IP and liability protection prior to product launch.