OllamaNet Glossary

This glossary provides consistent definitions for terminology used throughout the OllamaNet platform.

Core Concepts

AI/ML Terminology

Term	Definition
AI Model	A machine learning model trained to perform specific tasks, such as text generation or completion.
LLM (Large Language Model)	A type of AI model trained on vast amounts of text data, capable of generating human-like text based on prompts.
Ollama	An open-source framework for running large language models locally.
Prompt	The input text provided to an AI model to generate a response.
Completion	The text generated by an AI model in response to a prompt.
Context Window	The amount of text (tokens) an AI model can consider when generating a response.
RAG (Retrieval-Augmented Generation)	A technique that enhances AI responses by retrieving relevant information from external sources before generating a response.
Embedding	A numerical representation of text that captures semantic meaning, used for vector search in RAG.
Vector Database	A specialized database for storing and querying vector embeddings efficiently.
Semantic Search	Finding information based on meaning rather than exact keyword matching.
Document Chunking	The process of breaking documents into smaller pieces for processing and embedding.
Relevance Score	A numerical value indicating how relevant a piece of information is to a query.

Platform Components

Term	Definition
Gateway Service	The API gateway that routes requests to appropriate microservices.
AuthService	The service responsible for user authentication and authorization.
ConversationService	The service that manages conversations and chat interactions.
AdminService	The service providing administrative capabilities for platform management.

Term	Definition
ExploreService	The service enabling discovery and browsing of available AI models.
Ollama_DB_layer	The shared database access layer used by all services.
InferenceEngineConnector	The component that connects to the Ollama API for model interactions.

Conversation Management

Term	Definition
Conversation	A series of messages between a user and an AI model.
Message	A single text exchange within a conversation, either from the user or the AI.
Chat History	The complete record of messages within a conversation.
Folder	An organizational unit for grouping related conversations.
Note	User-created annotations associated with conversations.
Feedback	User evaluations of AI responses for quality assessment.
Streaming Response	Real-time delivery of AI responses as they are generated.
Server-Sent Events (SSE)	A technology for streaming data from server to client in real-time.

User Management

Term	Definition
User	An individual with an account on the platform.
Role	A set of permissions defining what actions a user can perform.
Admin	A user with administrative privileges.
JWT (JSON Web Token)	A secure token used for authentication and authorization.
Refresh Token	A long-lived token used to obtain new JWTs without re-authentication.
Claims	Pieces of information about a user contained in a JWT.

Model Management

Term	Definition
Model	An AI model available on the platform.
Tag	A label for categorizing and filtering models.
Model Metadata	Information about a model such as size, capabilities, and parameters.
Model Installation	The process of making a model available for use.
Model Uninstallation	The process of removing a model from availability.

Term	Definition
Progress Streaming	Real-time updates on long-running operations like model installation.

Document Processing

Term	Definition
Document	A file uploaded to the platform for context enhancement.
Document Processing	The extraction and preparation of text from documents.
Text Extraction	The process of obtaining plain text from various file formats.
Document Processor	A component that handles a specific document format.
Attachment	A document associated with a conversation.

Technical Concepts

Term	Definition
Microservice	An architectural approach where an application is composed of small, independent services.
API Gateway	A service that acts as an entry point for client requests to backend services.
Cache	A high-speed data storage layer that stores a subset of data for faster access.
TTL (Time to Live)	The duration for which cached data remains valid.
Repository Pattern	A design pattern that mediates between the domain and data mapping layers.
Unit of Work	A design pattern that maintains a list of objects affected by a business transaction.
Dependency Injection	A technique where one object supplies the dependencies of another object.
Options Pattern	A pattern for configuring .NET applications using strongly-typed settings.
Circuit Breaker	A design pattern that prevents cascading failures in distributed systems.
Retry Pattern	A pattern that enables an application to retry an operation in anticipation of it eventually succeeding.

Infrastructure Components

Term	Definition
Redis	An in-memory data structure store used for caching.
SQL Server	A relational database management system used for persistent storage.

Term	Definition
Entity Framework Core	An object-relational mapper (ORM) for .NET.
Ocelot	A .NET API Gateway.
Pinecone	A vector database used for semantic search in RAG.
RabbitMQ	A message broker used for service discovery and configuration updates.
Upstash	A cloud provider for Redis services.
ngrok	A service that exposes local servers to the internet through secure tunnels.

Performance Metrics

Term	Definition
Latency	The time delay between request initiation and response reception.
Throughput	The rate at which a system processes requests.
Cache Hit Ratio	The percentage of requests that are served from cache.
Response Time	The total time taken to respond to a request.
Concurrent Users	The number of users simultaneously using the system.
Uptime	The percentage of time a service is available and operational.
Resource Utilization	The amount of system resources (CPU, memory, etc.) being used.