GLOSARIO DE TÉRMINOS TÉCNICOS

Proyecto: C.O.L.E.G.A. (Language Copilot for Group and Administration Environments)
Versión del Documento: 1.0
Fecha: 03/12/2025
Referencia: Anexos I, II y III

Este documento recopila y define los términos técnicos, acrónimos y conceptos avanzados utilizados en la documentación del sistema C.O.L.E.G.A., abarcando administración de sistemas, DevOps, arquitectura de software e Inteligencia Artificial.

A

ALSA (Advanced Linux Sound Architecture)

Componente del kernel de Linux que proporciona controladores de dispositivo para tarjetas de sonido. En C.O.L.E.G.A., se utiliza para la captura de audio de baja latencia, evitando capas intermedias como PulseAudio cuando es posible para reducir el retardo en el VAD.

AppArmor

Módulo de seguridad del kernel de Linux que permite al administrador restringir las capacidades de un programa mediante perfiles. Se utiliza para "enjaular" los procesos del asistente, limitando su acceso al sistema de archivos y red.

Artifact (Artefacto)

En el contexto de CI/CD (GitHub Actions), se refiere a los archivos generados por un paso del pipeline (ej. binarios compilados, logs de test) que se guardan para ser usados en pasos posteriores o para su descarga.

B

Backpropagation (Retropropagación)

Algoritmo fundamental para el entrenamiento de redes neuronales. Calcula el gradiente de la función de pérdida con respecto a los pesos de la red, permitiendo ajustar estos pesos para minimizar el error. Usado por FANN en el módulo de NLU.

Barge-in

Capacidad del sistema para detectar que el usuario está hablando mientras el propio sistema está emitiendo audio (TTS), permitiendo interrumpir la respuesta actual para atender la nueva solicitud. Requiere cancelación de eco acústico (AEC).

Batching

Técnica de optimización en inferencia de LLMs que consiste en procesar múltiples solicitudes (prompts) simultáneamente en lugar de secuencialmente, mejorando el throughput (rendimiento) general del sistema.

C

Chain-of-Thought (CoT)

Técnica de ingeniería de prompts que induce al LLM a descomponer problemas complejos en pasos intermedios de razonamiento antes de dar la respuesta final. Mejora significativamente el rendimiento en tareas de lógica y matemáticas.

CI/CD (Continuous Integration / Continuous Deployment)

Práctica de desarrollo de software donde los cambios en el código se construyen, prueban y despliegan automáticamente. C.O.L.E.G.A. utiliza GitHub Actions para

validar cada commit y generar imágenes Docker.

Context Window (Ventana de Contexto)

El límite máximo de tokens (texto) que un LLM puede procesar en una sola interacción, incluyendo el prompt del sistema, el historial de la conversación y la nueva consulta del usuario. Para Gemma-2B, es de 2048 tokens.

D

Docker

Plataforma que utiliza virtualización a nivel de sistema operativo para entregar software en paquetes llamados contenedores. Garantiza que el asistente se ejecute de la misma manera en cualquier entorno (desarrollo, staging, producción).

Docker Compose

Herramienta para definir y ejecutar aplicaciones Docker multi-contenedor. Se usa para orquestar los servicios del asistente (Core, MQTT Broker, Base de Datos) en un solo archivo YAML.

E

Embedding

Representación vectorial (numérica) de un texto, imagen o audio en un espacio multidimensional. Los embeddings capturan el significado semántico, permitiendo que conceptos similares tengan vectores cercanos matemáticamente.

Epoch (Época)

En aprendizaje automático, una pasada completa de todo el conjunto de datos de entrenamiento a través del algoritmo de aprendizaje.

F

Fail2Ban

Framework de prevención de intrusiones que escanea archivos de log (como `/var/log/auth.log`) y banea direcciones IP que muestran signos maliciosos (ej. múltiples intentos fallidos de contraseña SSH).

FANN (Fast Artificial Neural Network)

Librería de redes neuronales ligera y escrita en C, utilizada por Padatious para la clasificación de intenciones. Ideal para sistemas embebidos por su bajo consumo de recursos.

FFT (Fast Fourier Transform)

Algoritmo matemático que transforma una señal del dominio del tiempo (audio crudo) al dominio de la frecuencia (espectro). Es el primer paso para extraer características de audio como los MFCCs.

Fine-Tuning (Ajuste Fino)

Proceso de tomar un modelo pre-entrenado (como Llama-3) y entrenarlo adicionalmente con un conjunto de datos específico de un dominio (ej. administración de sistemas) para mejorar su desempeño en esa área.

Fuzzy Logic (Lógica Difusa)

En el contexto de C.O.L.E.G.A., se refiere al uso de algoritmos de coincidencia de cadenas aproximada (como Levenshtein) para entender comandos de usuario que pueden contener errores tipográficos o variaciones ligeras.

G

GGUF (GPT-Generated Unified Format)

Formato de archivo binario para guardar modelos de lenguaje cuantizados, optimizado para una carga rápida y ejecución eficiente en CPU y GPU (vía `llama.cpp`).

Guardrails

Mecanismos de seguridad y control implementados en el sistema de IA para prevenir comportamientos indeseados, como respuestas tóxicas, ejecución de comandos peligrosos sin confirmación o alucinaciones críticas.

H

Hallucination (Alucinación)

Fenómeno donde un LLM genera información que parece plausible y coherente pero es factualmente incorrecta o inventada. Se mitiga mediante System Prompts estrictos y baja temperatura.

I

IaC (Infrastructure as Code)

Práctica de gestionar y aprovisionar la infraestructura informática a través de archivos de definición legibles por máquina (como Dockerfiles o manifiestos de Kubernetes), en lugar de configuración física de hardware o herramientas interactivas.

In-Context Learning

Capacidad de un LLM para aprender nuevas tareas o formatos simplemente viendo ejemplos en el prompt (few-shot prompting), sin necesidad de actualizar sus pesos (reentrenamiento).

Intent (Intención)

En NLU, representa lo que el usuario quiere conseguir con su frase. Por ejemplo, en la frase "¿Qué tiempo hace?", el intent podría ser `weather_query`.

K

Kubernetes (K8s)

Sistema de orquestación de contenedores de código abierto para automatizar el despliegue, el escalado y la gestión de aplicaciones en contenedores. Utilizado en despliegues de alta disponibilidad de C.O.L.E.G.A.

KV Cache (Key-Value Cache)

Técnica de optimización en Transformers que almacena los vectores de claves y valores de tokens anteriores para no tener que recalcularlos en cada paso de generación, acelerando drásticamente la inferencia.

L

Latency (Latencia)

El tiempo que transcurre entre una solicitud (ej. usuario termina de hablar) y la respuesta del sistema (ej. empieza a sonar el audio). En sistemas de voz, una latencia baja (<500ms) es crítica para la naturalidad.

Levenshtein Distance

Métrica para medir la diferencia entre dos secuencias de caracteres. Es el número mínimo de ediciones de un solo carácter (inserciones, eliminaciones o sustituciones) requeridas para cambiar una palabra en la otra.

LLM (Large Language Model)

Modelo de aprendizaje profundo entrenado en inmensas cantidades de texto que puede generar, resumir, traducir y comprender contenido en lenguaje natural.

LoRA (Low-Rank Adaptation)

Técnica eficiente de fine-tuning que congela los pesos del modelo pre-entrenado e inyecta matrices de rango bajo entrenables en cada capa del Transformer, reduciendo enormemente los requisitos de memoria para el entrenamiento.

M

MFCC (Mel-Frequency Cepstral Coefficients)

Representación del espectro de potencia a corto plazo de un sonido, basada en una transformación lineal de coseno de un espectro de potencia logarítmico en una escala de frecuencia mel no lineal. Es la característica estándar para reconocimiento de voz.

MQTT (Message Queuing Telemetry Transport)

Protocolo de mensajería ligero basado en el modelo publicación-suscripción, ideal para conectar dispositivos remotos con una huella de código pequeña y ancho de banda de red mínimo. Es la "columna vertebral" de comunicación de C.O.L.E.G.A.

N

NER (Named Entity Recognition)

Subtarea de extracción de información que busca localizar y clasificar entidades nombradas mencionadas en texto no estructurado en categorías predefinidas como nombres de personas, organizaciones, ubicaciones, expresiones de tiempo, cantidades, etc.

NLU (Natural Language Understanding)

Rama de la IA que se ocupa de la comprensión de lectura automática para que las máquinas comprendan la estructura y el significado del lenguaje humano.

O

OOM Killer (Out of Memory Killer)

Proceso del kernel de Linux que, en situaciones de memoria crítica, selecciona y termina procesos para liberar RAM y evitar que el sistema colapse. Los procesos de IA son candidatos frecuentes debido a su alto consumo.

P

PCM (Pulse Code Modulation)

Método utilizado para representar digitalmente señales analógicas muestreadas. Es el formato de audio estándar sin compresión utilizado internamente por el pipeline de voz.

Perplexity (Perplejidad)

Medida de qué tan bien un modelo de probabilidad predice una muestra. En LLMs, una perplejidad baja indica que el modelo está menos "sorprendido" por el texto y, por lo tanto, lo entiende/genera mejor.

Prompt Engineering

Arte de elaborar entradas (prompts) para guiar a los modelos de lenguaje generativo (LLMs) para que produzcan las salidas deseadas.

Q

Quantization (Cuantización)

Proceso de reducir la precisión de los números utilizados para representar los parámetros de un modelo (ej. de 16-bit float a 4-bit integer). Reduce el tamaño del modelo y el uso de memoria con una pérdida mínima de precisión.

R

RAG (Retrieval-Augmented Generation)

Técnica que mejora la precisión y fiabilidad de los modelos de IA generativa con datos obtenidos de fuentes externas (como una base de datos vectorial o documentos locales) durante la generación.

ReAct (Reasoning + Acting)

Paradigma donde el LLM genera trazas de razonamiento y acciones específicas de la tarea de manera intercalada. Permite al modelo interactuar con herramientas externas (APIs, bases de datos).

S

STT (Speech-to-Text) / ASR (Automatic Speech Recognition)

Tecnología que convierte el lenguaje hablado en texto escrito. C.O.L.E.G.A. utiliza Vosk (offline) o Whisper para esta tarea.

Swappiness

Parámetro del kernel de Linux que controla la agresividad con la que el sistema mueve procesos de la memoria física (RAM) al espacio de intercambio (Swap).

System Prompt

Instrucción inicial oculta dada a un LLM que define su comportamiento, personalidad y restricciones antes de que comience la interacción con el usuario.

T

Temperature (Temperatura)

Hiperparámetro en la generación de LLMs que controla la aleatoriedad de las predicciones. Una temperatura alta (ej. 1.0) produce respuestas más variadas y creativas; una baja (ej. 0.2) produce respuestas más deterministas y conservadoras.

Token

La unidad básica de texto que un LLM procesa. Puede ser una palabra, parte de una palabra o un carácter. Aproximadamente, 1000 tokens son 750 palabras en inglés.

Transformer

Arquitectura de red neuronal introducida en 2017 ("Attention Is All You Need") que se basa en mecanismos de auto-atención, permitiendo procesar secuencias de datos completas en paralelo. Es la base de todos los LLMs modernos.

TTS (Text-to-Speech)

Tecnología que convierte texto escrito en salida de audio hablada.

V

VAD (Voice Activity Detection)

Técnica utilizada en el procesamiento del habla en la que se detecta la presencia o ausencia de habla humana. Es crucial para saber cuándo empezar y dejar de escuchar al usuario.

Vector Database (Base de Datos Vectorial)

Base de datos optimizada para almacenar y consultar vectores (embeddings). Permite búsquedas de similitud semántica ultra-rápidas.

W

Wake Word (Palabra de Activación)

Palabra o frase clave (ej. "Oye Colega") que activa el asistente desde un estado de reposo de bajo consumo a un estado de escucha activa.

WAL (Write-Ahead Logging)

Modo de journal de SQLite que mejora significativamente la concurrencia, permitiendo que los lectores no bloqueen a los escritores y viceversa. Esencial para el rendimiento de la memoria a largo plazo.

WER (Word Error Rate)

Métrica común para evaluar el rendimiento de un sistema de reconocimiento de voz o traducción automática. Se calcula como (Sustituciones + Inserciones + Eliminaciones) / Número total de palabras.

Fin del Glosario Técnico