FICHA TÉCNICA CORPORATIVA

CyberAgente Enterprise

Plataforma Corporativa de Agentes Autónomos Gobernados, Ejecución Segura On-Premise y Voz en Tiempo Real

Categoría	Appliance empresarial de agentes autónomos gobernados
Despliegue	On-Premise / Air-Gapped **o** en la nube, según la necesidad del cliente
Modelos	Locales en LAN (Gemma 4 31B · Qwen 3 VL 30B) **o** en la nube (Gemini)
Gobernanza	Zero-Trust · Human-in-the-Loop · separación razonamiento / acción

1 Descripción General

CyberAgente es una plataforma tecnológica de nivel empresarial diseñada para dotar a organizaciones y corporativos de una **fuerza de trabajo digital de agentes autónomos de Inteligencia Artificial**, operando bajo estrictos estándares de gobernanza corporativa, mínimo privilegio (*Zero-Trust*), supervisión humana en el bucle (*Human-in-the-Loop*) y privacidad total de la información.

A diferencia de los chatbots convencionales de consumo masivo, CyberAgente está concebido como un **Appliance Empresarial** capaz de operar **100% On-Premise y en modo aislado de internet (*Air-Gapped*)**, conectado a servidores dedicados de inferencia en la red local (con modelos de última generación como **Gemma 4 31B** y **Qwen 3 VL 30B Multimodal**). Esto asegura que ningún secreto comercial, dato financiero o información sensible de clientes abandone la infraestructura privada de la empresa.

El sistema funciona bajo el principio de **Separación Estricta entre Razonamiento y Acción**: los modelos de IA razonan y proponen soluciones, pero ninguna acción que modifique el estado del negocio (transacciones, modificaciones en CRM, cotizaciones o correos formales) se ejecuta sin la emisión de tokens criptográficos de un solo uso y la autorización explícita del personal humano responsable.

2 Arquitectura del Sistema

Componente	Función Principal
Enrutador de Intenciones (CyberRouter)	Analiza las solicitudes en lenguaje natural y las canaliza al agente idóneo bajo mínimo privilegio y abstención segura (*Abstain-by-Default*).
Orquestador de Voz Corporativa (VoiceOrchestrator)	Detecta la palabra de activación (**«Cyber»**), segmenta el texto para streaming sin latencia y cancela procesos si el usuario interrumpe (*Barge-in*).
Fábrica de Agentes (Agent Factory & Compiler)	Entorno asistido para diseñar y compilar nuevos trabajadores digitales, generando manifiestos inmutables `AGENT.yaml` con digest SHA-256.
Arnés de Ejecución (Execution Harness)	Controla el ciclo de vida de los turnos agénticos, frena bucles con límites de pasos (*Step Limits*) y pausa cuando se requiere aprobación humana.
Motor de Políticas Determinista (Policy Engine - PDP)	Evalúa cada intención contra reglas RBAC/ABAC emitiendo dictámenes auditables (`ALLOW`, `DENY`, `REQUIRE_APPROVAL`, `ESCALATE`).

Componente	Función Principal
Protocolo de Acciones Hermes (Action Protocol)	Blinda la ejecución de herramientas con hashes inmutables, previene ataques de repetición y bloquea la auto-autorización agéntica.
Tokens de Capacidad Criptográficos	Llaves de ejecución efímeras firmadas con HMAC SHA-256, de un solo uso, que se destruyen tras ejecutar la herramienta autorizada.
Centro de Misiones y War Room (A2A)	Orquesta flujos donde múltiples agentes colaboran en tareas con dependencias, control de calidad (DoD) y salas de crisis en tiempo real.
Base de Conocimiento Vectorial RAG	Almacena y recupera información en PostgreSQL con `pgvector`, aplicando filtros de seguridad previos (*Pre-Retrieval ACL*) según el rol.
Puerta de Enlace Multi-Modelo (ModelGateway + FinOps)	Enruta a modelos locales en LAN o en la nube verificando límites de presupuesto mensual y clasificación de confidencialidad.
Base de Datos Multi-Tenant con RLS Forzado	Repositorio PostgreSQL 16 con aislamiento estricto por cliente a nivel de motor (*Row-Level Security* forzado).
Consola Web de Operaciones & Gateway HTTPS	Interfaz Dark Glass en los puertos `5173` y `8443` (HTTPS TLS) para chat, dictado por micrófono, configuración y monitoreo.

3 Módulos Principales

Módulo 1 — Consola Conversacional y Control por Voz

- Canal unificado de chat y audio bidireccional con activación instantánea al pronunciar la palabra clave **«Cyber»**.
- Selector en tiempo real del motor de inferencia: **Gemma 4 31B Local LAN**, **Qwen 3 VL 30B Multimodal Local** o **Google Gemini Cloud**.
- Soporte de micrófono físico en vivo gracias al Gateway HTTPS nativo con certificados TLS en el puerto `8443`.
- Transmisión de respuestas estructuradas en Markdown con desglose del agente asignado y el motivo de enrutamiento.

Módulo 2 — Fábrica y Compilador de Agentes Especializados

- Entrevista interactiva para especificar identidad, rol, responsabilidades, herramientas autorizadas y límites presupuestarios del nuevo agente.
- Compilación determinista del manifiesto formal `AGENT.yaml` sellado con firma digital SHA-256.
- Configuración de restricciones absolutas (*Never rules*) y definición de criterios de aceptación (*Definition of Done - DoD*).
- Activación y registro en caliente en el catálogo de `CyberRouter` sin necesidad de reiniciar servidores.

Módulo 3 — Arnés de Seguridad y Protocolo de Acciones Hermes

- Supervisión estricta de cada paso del agente: intercepción obligatoria antes de tocar cualquier base de datos o servicio externo.
- **Human-in-the-Loop:** ante riesgo financiero o legal, el arnés congela la tarea y genera una solicitud de aprobación en la consola del supervisor.
- Blindaje **Anti-Tampering:** si un agente intenta evadir la política o aprobarse permisos, se genera el evento `security.tamper-attempt` y se bloquea la sesión.
- Trazabilidad forense mediante identificador universal de rastreo (`trace_id`).

Módulo 4 — Misiones Multiagente y War Room de Incidentes

- Despliegue de misiones estratégicas compuestas por múltiples subtareas asignadas a agentes especializados o colaboradores humanos.
- Grafo de dependencias: una tarea no inicia hasta que sus requisitos previos cumplan con la *Definition of Done*.
- Freno de emergencia anti-bucles: corte automático si el intercambio entre agentes supera los 20 pasamanos (*max handoffs*).
- Sala de crisis (*War Room*) para el monitoreo de incidentes críticos, registro de bitácora y publicación de eventos en tiempo real.

Módulo 5 — Memoria Corporativa y RAG con Pre-Retrieval ACL

- Ingesta y fragmentación de manuales, políticas internas y catálogos de servicios con generación de embeddings vectoriales.
- Búsqueda semántica por similitud de coseno sobre la extensión `vector` de PostgreSQL.
- Filtro previo de control de acceso (*Pre-Retrieval ACL*): el motor solo busca en los documentos que el rol del usuario tiene autorizados.
- Memoria estructurada de hechos y preferencias con clasificación de procedencia (*User-Stated*, *Inferred*, *Verified*).

Módulo 6 — Operación de Infraestructura y CLI Maestro (`cyber`)

- Utilidad unificada de línea de comandos en `/usr/local/bin/cyber` para la administración integral del appliance.
- Diagnóstico de 20 puntos de control con `cyber doctor` (host, Docker, PostgreSQL, RLS, modelos de IA y guardrails).
- Verificación formal de preparación para producción con `cyber production-check`.
- Respaldo transaccional consistente y restauración verificada con `cyber backup` y `cyber restore`.

4 Funcionalidades Clave

- **Operación 100% Air-Gapped:** funcionamiento en entornos totalmente desconectados de internet, garantizando soberanía de datos.
- **Cero Alucinaciones en Tareas No Autorizadas:** el enrutador emite una abstención segura (`ABSTAIN · Sin Inventar`) en lugar de respuestas inciertas.
- **Tokens Criptográficos Efímeros:** cada herramienta recibe un token HMAC SHA-256 válido por 300 segundos que se destruye tras su uso único.
- **Control Presupuestario FinOps por Agente:** límite mensual de gasto y tokens en tiempo real; el sistema frena la llamada si se excede.
- **Interrupción Fluida por Voz (*Barge-in*):** el usuario puede interrumpir a Cyber mientras habla, reiniciando la atención de inmediato.
- **Aislamiento Multi-Tenant Estricto:** separación de datos forzada en PostgreSQL con Row-Level Security (`current_tenant_id()`).
- **Doble Modo de Acceso Web:** puerto HTTP `5173` para navegación estándar y HTTPS `8443` con TLS propio para uso seguro del micrófono.

5 Beneficios Operativos

- **Privacidad Absoluta y Cumplimiento Normativo:** alineado con LFPDPPP, GDPR e ISO 27001 al retener la información sensible en infraestructura propia.
- **Ahorro en Costos de Inferencia (FinOps):** reducción de hasta un **90%** en facturas de APIs cloud mediante GPUs locales con modelos abiertos (**Gemma 4 31B**, **Qwen 3 VL**).
- **Mayor Velocidad de Ejecución Comercial:** calificación de oportunidades B2B y borradores de cotización en menos de **5 segundos**.
- **Eliminación de Errores y Riesgos de Fraude:** ningún agente compromete presupuestos ni modifica datos críticos sin validación de un directivo humano.
- **Continuidad Operativa 24/7:** despliegue bajo `systemd` y Docker con auto-reinicio ante fallos y monitoreo con `cyber doctor`.
- **Escalabilidad Horizontal:** creación ilimitada de agentes para distintas divisiones (Ventas, Operaciones, Legal, Finanzas, TI) sobre el mismo appliance.

6 Requisitos Técnicos

- **Servidor Central del Appliance (Host Linux):**
  - **Sistema Operativo:** Ubuntu Server 24.04 LTS o Debian 12 (64 bits).
  - **Procesador (CPU):** mínimo 2 núcleos físicos (recomendado 4 a 8 vCPUs en entornos virtualizados Proxmox VE / VMware).
  - **Memoria RAM:** mínimo 6 GB (el sistema base consume ~700 MB, dejando más de 4.5 GB libres).
  - **Almacenamiento:** mínimo 40 GB en disco SSD / NVMe.
  - **Software Base:** Node.js v22.14 LTS+, Docker Engine v27+, Docker Compose v2.32+.
- **Servidor de Inferencia Local (GPU Cluster en LAN):**
  - **Hardware Gráfico:** 1x o 2x NVIDIA RTX 3090 / 4090 (24 GB VRAM) o aceleradores empresariales (A5000 / A6000 / H100).
  - **Software de Inferencia:** vLLM v0.21+ u Ollama con interfaz HTTP OpenAI-Compatible.
  - **Red Local:** conectividad LAN Gigabit (1 Gbps o 10 Gbps) con latencia menor a 1 ms hacia el servidor central.
- **Puertos de Red Requeridos:**
  - `3000 / TCP`: API REST de Producción CyberAgente.
  - `5173 / TCP`: Consola Web SPA (HTTP).
  - `8443 / TCP`: Gateway Seguro con TLS para Reconocimiento de Voz (HTTPS).
  - `5432 / TCP`: PostgreSQL 16 con extensión pgvector (red interna de datos).
  - `8003 / TCP`: Endpoint de inferencia de Gemma 4 31B en el servidor GPU LAN.
  - `8000 / TCP`: Endpoint de inferencia multimodal de Qwen 3 VL en el servidor GPU LAN.

7 Casos de Uso

- **Dirección General y Consultoría Ejecutiva:** asistencia estratégica con el **Cyber Asistente Ejecutivo**, redactando memorándums, analizando escenarios y coordinando reportes interdepartamentales.

- **Fuerza de Ventas Consultivas B2B:** el agente **Cyber Ventas** califica prospectos bajo ICP y BANT/MEDDIC, maneja objeciones y genera propuestas listas para firma humana.
- **Monitoreo de Infraestructura y Operaciones:** el agente **Cyber Operaciones** diagnostica servidores, supervisa colas, define matrices de escalación y reporta KPIs.
- **Gestión de Crisis e Incidentes de TI:** apertura de **War Rooms** automáticos ante eventos críticos, coordinando soporte y ciberseguridad con bitácora forense.
- **Consulta Inteligente de Bases de Datos (Text-to-SQL):** agentes de reportería que traducen preguntas de directivos en consultas SQL seguras de solo lectura (`SELECT`).

8 Escalabilidad y Seguridad

- **Arquitectura de Red Zero-Trust:** 3 redes Docker segmentadas (`cyberagente-edge`, `cyberagente-application`, `cyberagente-data`) que impiden el acceso directo a la base de datos desde internet.
- **Protección contra Inyección de Prompts:** el enrutador analiza la intención antes del LLM, y las llamadas a herramientas se aíslan por contratos tipados inmutables.
- **Cifrado de Comunicaciones:** canales TLS/HTTPS con certificados dedicados para la privacidad de los paquetes de voz y datos.
- **Prevención de Ataques de Repetición:** cada `CapabilityToken` tiene TTL de 300 s y un estado atómico que pasa de `ACTIVE` a `CONSUMED` en una sola transacción.
- **Aislamiento Criptográfico de Políticas:** los paquetes de políticas se compilan con hash inmutable; cualquier modificación no autorizada los invalida al instante.

9 Supervisión, Métricas y Analítica

- **Auditoría Forense Append-Only:** intenciones, decisiones del motor de políticas, aprobaciones y eventos se guardan en `audit_events` sin sobreescritura ni borrado.
- **Diagnóstico en Tiempo Real con `cyber doctor`:** salida visual o JSON (`--json`) para integración con SIEM y tableros (Datadog, Grafana, Zabbix).
- **Supervisión de Turnos y Pasos:** el arnés registra pasos de razonamiento, pensamientos intermedios, herramientas invocadas y latencias por turno.
- **Control de Pasamanos A2A:** contador visible de transferencias entre agentes para prevenir ciclos recursivos y acotar la duración de las misiones.

10 Mantenimiento, Actualización y Resiliencia

- **Instalación Idempotente en 1 Clic:** script `install.sh` que aprovisiona el entorno, valida dependencias y configura servicios de arranque automático.
- **Respaldo Atómico y Verificado:** `cyber backup` genera un paquete con la base de datos, certificados, configuraciones y manifiesto de integridad.
- **Restauración Transaccional con Pre-Check:** `cyber restore` valida la firma del respaldo antes de aplicarlo, con RTO menor a **3 minutos**.
- **Auto-Recuperación de Servicios:** `systemd` con reinicio automático (`Restart=always`, `RestartSec=3s`) y health-check continuo vía `/health`.

11 Diferenciadores Clave de CyberAgente

- **Gobernanza Nativa vs IA Sin Control:** no es un simple *wrapper* de APIs, sino un sistema operativo agéntico con separación de funciones, contratos normativos y frenos de emergencia.
- **Soberanía On-Premise y Air-Gapped:** diseñado desde su primera línea para correr en servidores propios y redes locales cerradas, sin depender de la nube.
- **Inferencia Local Híbrida:** compatible con clústeres locales de GPU que ejecutan modelos de 31 mil millones de parámetros con costo marginal cero por token.
- **Interacción Natural por Voz con Wake Word:** experiencia ejecutiva diciendo **«Cyber»** desde cualquier navegador autorizado con HTTPS.
- **Fábrica de Agentes Estandarizada:** cualquier área crea sus propios especialistas digitales sin escribir código, bajo el estándar formal OpenSpec.

12 Integración Empresarial

- **Conectividad REST Estándar:** API HTTP documentada con OpenAPI 3.1 y errores normalizados según **RFC 9457 Problem Details**.
- **Integración con Bases de Datos y ERPs:** conexión con PostgreSQL, MySQL, SQL Server, Oracle y sistemas ERP/CRM mediante adaptadores gobernados de solo lectura o escritura condicionada.
- **Compatibilidad con Hipervisores:** optimizado para Proxmox VE, VMware ESXi, Nutanix o bare-metal en Dell PowerEdge, HP ProLiant o Lenovo ThinkSystem.
- **Formatos de Intercambio Seguros:** mensajería JSON estructurada, eventos firmados y exportación de auditorías para Cumplimiento y Ciberseguridad.