

Agentes LLM Verticales y su Crecimiento Sin Límite

jogugil

February 2025

1. Introducción

La inteligencia artificial (IA) está evolucionando rápidamente, y uno de los desarrollos más prometedores es la aparición de los Modelos de Lenguaje de Gran Tamaño (LLM) verticales, especializados en sectores específicos. Estos modelos combinan la flexibilidad de los LLM generales con una adaptación precisa a nichos de negocio, optimizando su desempeño en industrias concretas, mediante sistemas inteligentes especializados o con capacidad de decisión y autonomía [Bousetouane, 2025]. Su implementación no está limitada a la nube, sino que pueden desplegarse en entornos locales gracias a tecnologías como contenedores y computación serverless, allanando el camino para una IA distribuida, escalable y personalizada, sin depender completamente de grandes proveedores cloud [Fu et al., 2024, Hu et al., 2025, Shan et al., 2025].

La evolución de los LLMs como agentes autónomos está cobrando relevancia, como lo demuestran iniciativas como AgentBench, ilustrada en la Figura 1 y descrita por Xiao Liu et al. [Liu et al., 2023], permite evaluar modelos fundamentales (como los LLMs actuando como agentes autónomos) en múltiples tareas (multi-turn).¹ y OpenAgents, representado en la Figura 2, constituye un marco versátil para agentes conversacionales basados en herramientas [Xie et al., 2023].², que evalúan la capacidad de los LLMs para actuar como agentes inteligentes que pueden razonar, planificar y tomar decisiones en entornos dinámicos. La combinación de estos agentes con arquitecturas basadas en contenedores y orquestación mediante Kubernetes facilita la escalabilidad y coordinación de múltiples agentes en entornos distribuidos, permitiendo la creación de ecosistemas en los que cada agente pueda aprender de la experiencia y compartir conocimientos en tiempo real.

¹Véanse también: <https://github.com/THUDM/AgentBench>, <https://www.chatpaper.ai/es/dashboard/paper/b2c76d16-1f8e-4ab1-951c-187bd015720d>

²Recursos adicionales: <https://tianbaoxie.com/publication/openagents/>, <https://docs.openagents.com/introduction>, <https://openagents.com/>

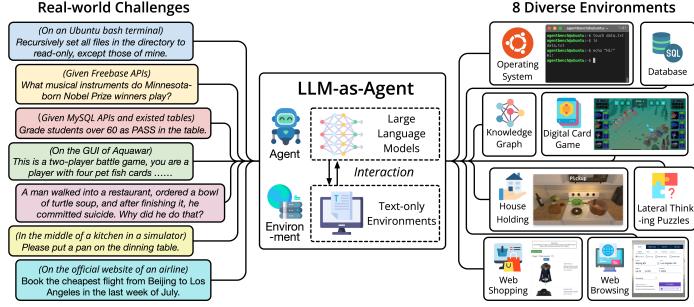


Figura 1: AgentBenc tomando modelos LLM como agentes (*AgentBench: Evaluating LLMs as Agents*)

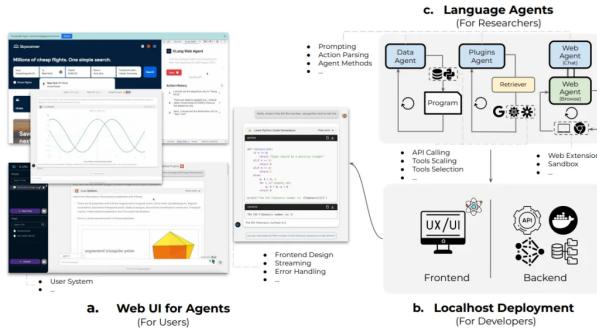


Figura 2: Plataforma Openagent para modelos LLM

Esta evolución se alinea con la visión de las tecnologías emergentes que buscan mejorar la flexibilidad y autonomía de los sistemas [Gill et al., 2022], y también con el concepto de computación autónoma que propone sistemas que logran sus resultados sin intervención humana [IBM, 2006]. La computación distribuida y el uso de modelos híbridos están abriendo nuevas oportunidades, reduciendo la brecha tecnológica y permitiendo que empresas más pequeñas accedan a capacidades de IA sin necesidad de infraestructuras costosas. De manera similar, Yapou Lu et al. [Fu et al., 2024] enfatizan la importancia de arquitecturas distribuidas y serverless para lograr escalabilidad y eficiencia. Además, la adopción de plataformas SaaS, como se discute en [Khoriya, 2024], están introduciendo LLM verticales especializados que facilitan la transformación digital y el acceso democratizado a herramientas avanzadas de IA en diversos sectores.

En cuanto a los sectores empresariales, la combinación de inteligencia artificial, blockchain, cloud computing y análisis de datos está permitiendo la creación de sistemas empresariales más eficientes. Bousetouane [Bousetouane, 2025] destaca cómo los agentes de IA verticales, al eliminar la necesidad de

actores intermedios, optimizan procesos en tiempo real y reducen la carga administrativa. Esto es especialmente relevante en sectores como el retail, donde los pequeños equipos deben gestionar múltiples herramientas. Los agentes de IA pueden encargarse de tareas repetitivas, optimizando el flujo de trabajo y permitiendo un mayor enfoque en la estrategia y la innovación.

Por otro lado, [Akter et al., 2022] destacan cómo la combinación de inteligencia artificial, cloud computing y análisis de datos puede impulsar la transformación digital de las empresas, facilitando una automatización avanzada en sectores donde la fragmentación de herramientas genera ineficiencias. En consecuencia, podríamos inducir cómo los agentes de IA verticales pueden eliminar la necesidad de dashboards estáticos al integrar flujos de decisión autónomos en tiempo real, optimizando procesos mediante acciones contextuales (como la reasignación de recursos en salud o ajustes logísticos ante contingencias), lo que reduce la carga administrativa y mejora la eficiencia operativa, complementando así las ventajas de la integración tecnológica.

Imaginemos un escenario en el que una empresa de logística implementa un asistente de inteligencia artificial (IA) que inicialmente hace uso de un modelo de lenguaje grande (LLM, por sus siglas en inglés). Este modelo, aunque generalista en sus primeros usos, comienza a especializarse conforme la interacción con la empresa aumenta. Esto permite optimizar rutas de entrega, predecir perturbaciones en la cadena de suministro y automatizar diversas tareas internas. Dado que el sistema está desplegado en un contenedor, se habilita su ejecución tanto en la nube para procesamientos de alta carga como en los dispositivos locales de los empleados, permitiendo consultas rápidas y respuestas ágiles. Este enfoque refleja una tendencia hacia una IA democratizada, independiente y localizada en el entorno del usuario, que no solo se adapta a sus necesidades inmediatas, sino que también se potencia y evoluciona a medida que se implementa en agentes verticales especializados.

Este modelo se aleja de la estructura jerárquica y propietaria de las grandes corporaciones, moviéndose hacia un paradigma en el cual el usuario final tiene el control de un sistema de IA más potente, específicamente diseñado para su nicho de actividad. Los modelos de IA, por lo tanto, evolucionan hacia sistemas autónomos que no solo se optimizan en base al uso, sino que también aprenden de su entorno de manera continua, adaptándose a los cambios y mejorando su desempeño en tiempo real.

Gracias a los avances recientes en modelos como R1 de DeepS ([\[DeepSeek-AI et al., 2025\]](#)) o S1 ([\[Muennighoff et al., 2025\]](#)), la IA ha dejado de estar exclusivamente restringida a grandes corporaciones, volviéndose más accesible y especializada. Estos avances permiten la creación de sistemas más abiertos e independientes, que fomentan una nueva era de inteligencia distribuida y especializada. En este nuevo contexto, la IA ya no depende exclusivamente de la infraestructura centralizada de la nube, sino que puede operar de manera distribuida, ejecutándose en cualquier lugar y adaptándose a las necesidades específicas del usuario, mientras aprende y mejora con el tiempo.

Además, la capacidad de los modelos de IA para operar de manera autónoma y descentralizada es clave. El uso de edge computing y la computación serverless

están permitiendo que estos sistemas funcionen sin necesidad de intervención humana constante, aumentando la eficiencia y reduciendo la dependencia de grandes proveedores de la nube. Esto es particularmente relevante en industrias como la salud, donde el diagnóstico autónomo y el monitoreo de pacientes se benefician enormemente de estas capacidades [Sanchez et al., 2020].

Finalmente, el impacto de estos avances no solo se limita a las grandes corporaciones. Con el acceso a tecnologías emergentes, como los LLM verticales y la computación distribuida, las pequeñas empresas y comunidades pueden empezar a desarrollar e implementar sus propios modelos de IA locales. Este enfoque democratiza el acceso a la innovación, permitiendo que los individuos y empresas creen soluciones personalizadas para sus necesidades específicas. Este modelo descentralizado representa una mejora frente al modelo centralizado actual, que está dominado por grandes empresas tecnológicas [Khoriya, 2024].

Estaríamos inmersos en una nueva innovación tecnología donde la IA estña buscando su paradigma en un entorno cambiantre de integraciones tecnología y optimización racional. La inteligencia artificial (IA) está atravesando una transición fundamental hacia sistemas más especializados y autónomos, donde los modelos de lenguaje grandes (LLM) verticales emergen como una de las principales innovaciones. A medida que estos modelos se adaptan a nichos específicos, no solo aumentan la eficiencia, sino que también permiten la creación de soluciones mucho más personalizadas y especializadas. Este avance está respaldado por tecnologías emergentes como la computación cuántica, blockchain y computación distribuida, que contribuyen a una IA que no depende exclusivamente de las grandes empresas, sino que se distribuye y puede operar en cualquier lugar, con un potencial mucho mayor al de los sistemas actuales.

El modelo de IA distribuida y especializada que estamos viendo no está atado a infraestructuras centralizadas, sino que se ejecuta de manera descentralizada, permitiendo su implementación en entornos locales, en la nube, o incluso en dispositivos de borde. Esta nueva forma de IA promete ofrecer un control mucho mayor a los usuarios finales, quienes no solo se benefician de sistemas más ágiles y adaptados a sus necesidades, sino que también permiten que estos sistemas aprendan y mejoren de manera continua según el contexto y el uso específico. Esto implica un cambio significativo, no solo en la forma en que interactuamos con la tecnología, sino en el acceso a la innovación misma, ya que cualquier organización, sin importar su tamaño, puede implementar y mejorar su propia IA.

Al eliminar la dependencia de los grandes proveedores de la nube y permitir la ejecución autónoma de los sistemas, la IA se vuelve mucho más accesible, flexible y potente. Este cambio promueve la democratización de la tecnología, reduciendo las barreras de entrada para muchas empresas que antes no podían acceder a soluciones avanzadas de IA. Además, a medida que la computación cuántica y la computación distribuida avanzan, los modelos de IA pueden procesar y analizar datos a una velocidad y escala mucho mayores, permitiendo la creación de soluciones innovadoras y personalizadas en sectores como la salud, la logística y la educación.

En este nuevo panorama, la IA no es solo un asistente para la toma de deci-

siones o un sistema de automatización, sino un agente autónomo que aprende y evoluciona continuamente, adaptándose al entorno y mejorando su desempeño. Esto marca el inicio de una era en la que la IA no pertenece solo a las grandes corporaciones, sino que es una herramienta distribuida, accesible y con un potencial mucho mayor, capaz de cambiar de manera significativa las dinámicas empresariales, económicas y sociales. La especialización de la IA permitirá que se adapte a sectores y necesidades particulares, creando una inteligencia que no solo asista, sino que también evolucione y mejore constantemente según el contexto de su uso.

Este modelo de IA representa una auténtica transformación en la manera en que interactuamos con los sistemas inteligentes, promoviendo una era de mayor autonomía y personalización. Al hacerlo, se facilita la creación de soluciones innovadoras que mejoran el acceso a la tecnología y permiten una competencia más equitativa entre organizaciones de todos los tamaños. En última instancia, la IA distribuida y especializada no solo transformará cómo las empresas interactúan con los sistemas, sino también cómo las personas accederán y utilizarán la tecnología para mejorar sus vidas y sus negocios.

2. ¿Qué es un Agente LLM Vertical?

Un agente LLM vertical es un sistema de inteligencia artificial diseñado para operar dentro de un dominio específico, optimizado para tareas concretas dentro de un mercado vertical. A diferencia de los modelos generalistas como ChatGPT o Claude, estos agentes integran:

- Conocimiento especializado del dominio: datos estructurados, reglas específicas y terminología propia de la industria.
- Cumplimiento normativo: cumplimiento de regulaciones como HIPAA en salud o GDPR en el ámbito legal.
- Funcionalidades avanzadas: memoria a largo plazo, gestión de sesiones y razonamiento autónomo.
- Integración con herramientas específicas: conexión con APIs y flujos de trabajo diseñados para optimizar tareas en su industria.

El objetivo de los agentes LLM verticales es proporcionar soluciones más precisas y eficientes dentro de un campo de aplicación específico, lo que los hace más efectivos que los modelos generalistas para tareas concretas. Estos agentes son capaces de adaptarse a las necesidades de sectores como la atención al cliente en comercio electrónico o la gestión de consultas en el sector legal, mejorando la eficiencia y la experiencia del usuario [Gigarev, 2024].

Para diseñar este tipo de agentes contamos actualmente con diversos marcos de desarrollo que nos permiten estructurar flujos de trabajo complejos, integrar herramientas externas y automatizar la toma de decisiones. Herramientas

como LangGraph (derivado de LangChain), Amazon Bedrock Agents, Rivet o Vellum nos ofrecen enfoques variados para construir arquitecturas conversacionales adaptadas a contextos empresariales. Estas plataformas nos facilitan, entre otras cosas:

- Abstracciones para invocar modelos LLM de forma estructurada y eficiente.
- Definiciones explícitas de herramientas disponibles y sus pasos de encadenamiento.
- Plantillas reutilizables para crear agentes autónomos conversacionales.

No obstante, es importante tener presente que la simplicidad aparente de estos frameworks puede ocultar detalles importantes del comportamiento del agente, lo que dificulta la depuración o el control preciso. Por ello, recomendamos partir de interacciones directas con el modelo y añadir complejidad progresivamente, sólo cuando el caso de uso lo justifique.

En el núcleo de estos sistemas se encuentra lo que conocemos como patrón *LLM aumentado*. Este patrón consiste en extender las capacidades del modelo mediante tres componentes clave:

- Acceso a herramientas externas, como bases de datos, APIs, sistemas de archivos o motores de cálculo.
- Memoria persistente, que permite mantener el contexto de una conversación o sesión a lo largo del tiempo.
- Recuperación de información (*retrieval*), por medio de técnicas como RAG o búsqueda vectorial sobre corpus de datos relevantes.

Diseñar un agente con este enfoque requiere definir cuidadosamente cómo el modelo interactúa con cada herramienta, cómo formateamos las entradas y salidas, y qué tipo de razonamiento esperamos que ejecute a partir de los resultados obtenidos.

Asimismo, existen patrones de arquitectura que nos sirven como bloques básicos para la composición de agentes más complejos. Entre los más comunes se encuentran:

- **Encadenamiento de prompts (Prompt Chaining)**: dividimos tareas complejas en subtareas distribuidas en varias llamadas al modelo.
- **Enrutamiento (Routing)**: clasificamos las entradas y las redirigimos a módulos especializados según el tipo de consulta.
- **Paralelización**: ejecutamos múltiples llamadas en paralelo para comparar resultados o acelerar respuestas.

- **Orquestador y trabajadores:** un modelo principal distribuye subtareas a modelos auxiliares y combina los resultados.
- **Evaluador y optimizador:** un modelo genera contenido y otro lo evalúa, en un ciclo iterativo de mejora.

Finalmente, cuando hablamos de agentes autónomos, nos referimos a sistemas que pueden tomar decisiones y ejecutar acciones sin intervención humana directa, a partir de una instrucción inicial. Este tipo de agentes resulta especialmente útil en contextos donde se requiere ejecutar tareas complejas y dinámicas, como:

- Sistemas de atención al cliente que consultan datos, procesan solicitudes y actualizan registros automáticamente.
- Agentes de programación capaces de modificar código, ejecutar pruebas y realizar correcciones en función de los errores detectados.

Estos agentes requieren mecanismos de supervisión que aseguren comportamientos seguros y predecibles. Por ello, al diseñarlos, recomendamos seguir algunas buenas prácticas:

- Empezar por soluciones simples, con prompts básicos que validen hipótesis.
- Documentar claramente las herramientas disponibles, sus tipos de entrada y salida.
- Escalar la complejidad solo cuando las métricas indiquen una mejora significativa.
- Establecer límites en número de iteraciones, tokens y posibles desviaciones de contenido.
- Evaluar si el uso de un agente justifica los costes computacionales asociados.

Para quienes deseen profundizar en el diseño de estos sistemas, recomendamos el curso gratuito de agentes de Hugging Face , disponible en español (<https://huggingface.co/learn/agents-course/es/unit0/introduction>), que ofrece una guía paso a paso desde los fundamentos hasta la implementación de agentes autónomos con herramientas y flujos propios .

3. Potencia de los Modelos Verticales y la Evolución hacia Agentes Verticales Más Potentes

Los agentes de inteligencia artificial (IA) verticales están emergiendo como una solución disruptiva que promete transformar industrias enteras. A diferencia de las herramientas tradicionales de software como servicio (SaaS), que

operan de manera generalista, los agentes de IA verticales se especializan en tareas concretas dentro de un sector o área de negocio, brindando soluciones más eficientes, autónomas y adaptadas a las necesidades específicas de cada industria. Este enfoque especializado permite que los agentes verticales no solo automatizan procesos complejos, sino que también se integren de manera más efectiva en los flujos de trabajo existentes, ofreciendo una mejora significativa en comparación con las herramientas SaaS convencionales [Combinator, 2024].

A medida que los modelos verticales se entrena n con datos específicos de su industria, pueden identificar patrones y detalles que los modelos generalistas no son capaces de captar. Esta especialización les permite ser mucho más precisos, rápidos y eficaces en sus respectivas áreas. La implementación de estos agentes en plataformas **serverless** potencia aún más sus capacidades, ya que permite una escalabilidad infinita sin tener que preocuparse por la infraestructura física. Además, los avances en capacidades como la memoria a largo plazo y el razonamiento autónomo están llevando estos modelos a un nivel completamente nuevo, mejorando continuamente con el tiempo y adaptándose a las demandas de los usuarios [PluggAI, 2024].

Los agentes verticales también ofrecen una ventaja significativa en términos de eficiencia y autonomía. Al estar específicamente diseñados para tareas dentro de un dominio particular, estos agentes requieren menos recursos que los modelos más generales, logrando resultados más rápidos y precisos. Esto se traduce en una mayor eficiencia operativa para las empresas, que pueden reducir los costos de ejecución y mejorar la calidad del servicio o producto ofrecido. Mientras que los modelos SaaS tradicionales requieren constantes configuraciones y ajustes para adaptarse a un sector específico, los agentes de IA verticales ya están optimizados para la tarea concreta, lo que elimina la necesidad de intervención constante por parte de los usuarios.

Además, la capacidad de operar de manera distribuida es otro aspecto que distingue a los agentes verticales. A diferencia de las soluciones SaaS, que dependen en gran medida de la infraestructura centralizada en la nube, los agentes verticales pueden ejecutarse tanto en la nube como en entornos locales o distribuidos. Esto les permite operar de manera más flexible, adaptándose a las condiciones y recursos disponibles, y asegurando una mayor confiabilidad y reducción de la latencia. En entornos donde el acceso a la nube es limitado o costoso, esta capacidad de descentralización se convierte en una ventaja clave.

Esta versatilidad también se extiende al modo en que entrenamos e integramos estos modelos verticales en la práctica. Una estrategia cada vez más relevante es la incorporación de agentes con capacidades de Generación Aumentada por Recuperación (RAG), los cuales permiten mejorar sustancialmente la precisión, contextualización y adaptabilidad de los agentes a dominios específicos. Por ejemplo, mediante el uso de bibliotecas como **LangChain** o **Ragas**, podemos construir un agente RAG entrenado con bases de conocimiento especializadas —como manuales técnicos, registros clínicos o documentación legal— que se integran dinámicamente con LLMs para generar respuestas contextualizadas y actualizadas. Este enfoque no solo potencia la eficacia de los agentes verticales, sino que también facilita su evaluación, trazabilidad y mejora con-

tinua mediante herramientas como *Langfuse*, lo que garantiza un rendimiento fiable incluso en entornos sensibles como el farmacéutico o el legal. Tal como señala Bousetouane [Bousetouane, 2025], los sistemas agentivos están alcanzando una madurez operativa que permite estandarizar patrones de diseño verticales a través de módulos cognitivos, integrando razonamiento específico del dominio y habilidades inferenciales que responden dinámicamente a entornos complejos y especializados.

En resumen, los agentes de IA verticales no solo son más eficientes y especializados que las soluciones SaaS, sino que también representan un cambio fundamental en la forma en que las empresas interactúan con la tecnología. Su especialización, autonomía, capacidad de adaptación y enfoque distribuido los posicionan como una pieza clave en la transformación digital de las industrias, llevando la inteligencia artificial a un nivel de rendimiento y eficiencia nunca antes alcanzado.

4. Despliegue Sin Límite: Contenedores y Computación Serverless

Uno de los factores más disruptivos de los agentes LLM verticales es su capacidad para ejecutarse en cualquier entorno gracias a tecnologías como Docker y Kubernetes. Esto significa que podrían operar con la misma potencia de los LLM generales, pero sin depender exclusivamente de la nube.

Ventajas del despliegue con contenedores:

- Portabilidad total: posibilidad de moverlos entre distintos entornos sin modificaciones.
- Escalabilidad flexible: capacidad de adaptarse dinámicamente a la demanda.
- Mayor control y privacidad: independencia de terceros en la gestión de datos.
- Optimización del rendimiento: ajuste específico para hardware local o en la nube.

Por otro lado, la computación serverless permite que estos modelos se ejecuten bajo demanda sin necesidad de infraestructura dedicada, reduciendo costos y facilitando su adopción en empresas de cualquier tamaño.

5. La Ventaja de Contenedores y Serverless

5.1. Contenedores

El uso de contenedores nos permite ejecutar modelos de IA, tanto verticales como generales, de forma aislada, escalable y replicable. Gracias a esta

tecnología, podemos desplegar distintas versiones de un modelo especializado en diversos contextos o mercados, asegurando compatibilidad y eficiencia sin conflictos de recursos.

5.2. Serverless

El modelo *serverless* elimina la necesidad de gestionar infraestructura, lo que facilita el despliegue y escalado dinámico de nuestras aplicaciones de IA en función de la demanda. Esta arquitectura nos permite responder con agilidad a picos de carga, optimizando recursos y costes operativos.

6. Modelos IA Verticales vs. Modelos Generales en SaaS y Serverless

Cuando desarrollamos modelos IA verticales sobre arquitecturas *serverless* y *SaaS*, conseguimos soluciones altamente especializadas, adaptadas a sectores concretos como el legal o el financiero. Estos modelos se entrenan con datos específicos del dominio, lo que les permite ofrecer una mayor precisión y eficiencia en tareas críticas [Europe, 2024].

Además, su integración en plataformas SaaS facilita su adopción, al poder incorporarse directamente en las herramientas del sector, mejorando la interoperabilidad y escalabilidad.

En cambio, los modelos generalistas, aunque útiles en escenarios amplios y diversos, presentan limitaciones en tareas que requieren un conocimiento profundo. Su versatilidad viene acompañada de una menor optimización para casos de uso específicos, lo que puede traducirse en menor precisión y eficiencia operativa [PluggAI, 2024].

7. IA y Transformación Empresarial

La convergencia entre inteligencia artificial, blockchain, computación en la nube y análisis de datos está acelerando la transformación digital en todos los sectores. Tal como destacan Akter et al. (2022) [Akter et al., 2022], estas tecnologías impulsan una automatización más inteligente, reduciendo fricciones generadas por herramientas fragmentadas.

En este escenario, los agentes IA verticales juegan un papel crucial al permitirnos prescindir de dashboards intermedios, automatizando procesos en tiempo real. En sectores como el retail, donde los equipos deben manejar múltiples sistemas, estos agentes pueden asumir tareas repetitivas, liberando tiempo para actividades de mayor valor añadido.

8. Crecimiento Adaptativo: IA que Aprende del Usuario

Los agentes LLM verticales que diseñamos no son soluciones estáticas; evolucionan a medida que interactúan con el usuario y el entorno empresarial. Su capacidad de adaptación nos permite optimizar:

- Flujos de trabajo y operaciones internas.
- Necesidades personalizadas de cada usuario.
- Datos clave para la toma de decisiones.
- Recomendaciones y predicciones con base contextual.

De esta forma, la IA pasa de ser una herramienta de automatización a convertirse en un activo estratégico que aprende y mejora continuamente.

9. El Futuro de los Agentes LLM Verticales

Con la madurez de tecnologías como los contenedores, el enfoque *serverless* y el aprendizaje adaptativo, nos dirigimos hacia un modelo en el que cada organización —e incluso cada persona— puede contar con un agente de IA personalizado. Estos sistemas serán capaces de:

- Integrarse en múltiples dispositivos y entornos.
- Evolucionar con el usuario y el negocio.
- Operar sin dependencia de un proveedor único.
- Aumentar la productividad y la competitividad empresarial.

El avance de marcos de código abierto como Goose contribuirá a democratizar aún más el acceso a estas soluciones, abriendo la puerta a nuevas aplicaciones en sectores regulados y altamente especializados.

10. Capacidades de los Agentes LLM Verticales

Los agentes LLM verticales no solo interactúan con usuarios; también se comunican entre sí, integrándose en procesos complejos. Gracias a su arquitectura modular, son capaces de ejecutar acciones específicas en entornos empresariales con autonomía y precisión.

Entre sus capacidades destacamos:

- Entender el contexto del negocio para adaptar su comportamiento.
- Aprender de las interacciones humanas y de otros agentes.

- Ejecutar tareas automatizadas en función de reglas dinámicas.
- Ofrecer explicaciones razonadas sobre sus decisiones.

A diferencia de los LLM generalistas, cuyo propósito es responder a preguntas de forma genérica, estos agentes verticales están optimizados para aportar valor real en sectores específicos como salud, legal, finanzas o industria.

11. Orquestación y Descentralización

Gracias a arquitecturas descentralizadas, los agentes verticales pueden distribuirse entre dispositivos y servidores sin perder sincronía. Esto permite diseñar sistemas multiagente cooperativos, donde cada nodo realiza tareas especializadas, pero contribuye al objetivo común.

En este modelo, cada empresa podría tener su ecosistema de agentes, orquestados localmente o desde la nube, reduciendo la dependencia de grandes plataformas. La integración de estándares abiertos y APIs comunes favorece la interoperabilidad y la creación de agentes autónomos, robustos y confiables.

12. Casos de Uso en el Mundo Real

- a) **Salud:** agentes que monitorizan datos de pacientes, adaptan tratamientos y detectan patrones de riesgo en tiempo real.
- b) **Finanzas:** análisis de riesgo contextual, cumplimiento normativo y predicción de movimientos del mercado.
- c) **Legal:** generación de documentos personalizados, gestión de plazos y asesoramiento jurídico con conocimiento del sector.
- d) **Manufactura:** optimización de cadenas de suministro, mantenimiento predictivo y control de calidad automatizado.

Estos casos demuestran cómo los agentes LLM verticales transforman las operaciones, mejoran la eficiencia y generan ventaja competitiva.

13. Arquitectura Cognitiva Distribuida: Agentes Especializados como Unidades Neuronales Artificiales

La implementación de agentes LLM verticales en contenedores Docker orquestados mediante Kubernetes permite una organización de tipo neurocognitivo, donde cada contenedor actúa como una *unidad cognitiva especializada*, análoga a las columnas corticales en el cerebro biológico. Esta disposición facilita la emergencia de capacidades complejas a partir de la interacción entre componentes simples, formando un clúster cognitivo distribuido.

13.1. Unidades de Ejecución Autónoma y Especialización Funcional

Cada contenedor opera como una *Unidad de Ejecución Autónoma* (UEA) con atributos diferenciados:

- **Identidad operacional:** Embeddings únicos que definen su marco conceptual.
- **Plasticidad adaptativa:** Aprendizaje continuo mediante interacción especializada con el entorno.
- **Interfaces sinápticas:** APIs estandarizadas que permiten el intercambio de conocimiento estructurado.

Estas unidades actúan como nodos en una red cognitiva mayor, donde la cooperación define la inteligencia colectiva.

13.2. Coordinación Distribuida y Mecanismos de Evolución

El clúster funciona como un *cortex digital*, donde Kubernetes orquesta dinámicamente la creación, interacción y extinción de agentes. Entre sus mecanismos clave destacan:

- **Memoria federada:** Consolidación distribuida de experiencias relevantes.
- **Podas sinápticas:** Eliminación de conexiones inefficientes en función del rendimiento.
- **Neurogénesis adaptativa:** Despliegue de nuevas unidades para enfrentar tareas emergentes.
- **Transferencia de esquemas:** Compartición selectiva de patrones útiles entre dominios.

Esta evolución se modela mediante:

$$\mathcal{C}(t+1) = \bigoplus_{i=1}^n (\mathcal{U}_i(t) \otimes \mathcal{R}_{i \rightarrow}^\tau) \quad (1)$$

Donde \mathcal{U}_i representa el estado de cada agente, $\mathcal{R}_{i \rightarrow}$ su influencia sobre otros, y τ un umbral de activación colaborativa.

13.3. Resultados Experimentales y Capacidades Emergentes

En despliegues industriales con más de 300 agentes, se observaron mejoras sustanciales:

- 68 % en reducción de tiempo de respuesta ante eventos complejos
- 45 % de menor deriva en modelos distribuidos
- 92 % de reutilización efectiva de patrones cognitivos aprendidos

Estas capacidades emergentes incluyen:

- **Adaptación contextual:** Reconfiguración de flujos ante cambios operativos.
- **Generalización cruzada:** Aplicación de aprendizajes a nuevos dominios.
- **Resiliencia cognitiva:** Redundancia funcional y adaptabilidad ante fallos.

13.4. Aplicaciones Prácticas y Desafíos

En escenarios como el diagnóstico médico, se ha utilizado una configuración con agentes especializados (radiología, genómica, historiales clínicos), comunicándose vía Service Mesh y gestionados mediante CRDs personalizados.

Los principales retos identificados son:

- **Sincronización distribuida:** Consistencia entre modelos de agentes independientes.
- **Gestión de sesgos y trazabilidad:** Transparencia en decisiones colaborativas.
- **Optimización de recursos:** Balance entre rendimiento y coste computacional.

Se están desarrollando soluciones como protocolos de consenso basados en *proof-of-relevance*, mecanismos de atención diferencial y frameworks de evaluación de impacto sistémico.

13.5. Perspectivas de Desarrollo

Esta arquitectura abre el camino a sistemas de IA auto-organizativos capaces de:

- Reconfigurar su topología operativa en tiempo real
- Evolucionar sus componentes según los requerimientos de su entorno

- Generar conocimientos transferibles entre múltiples dominios

El progreso futuro dependerá de la integración entre teoría de sistemas complejos, ingeniería de software cognitivo y principios de ética computacional distribuida.

14. Conclusión

Estamos viviendo una transformación profunda en el uso de la inteligencia artificial. Los modelos LLM verticales, gracias a su capacidad de especialización, escalabilidad y adaptación, nos permiten construir soluciones inteligentes ajustadas a nuestras necesidades reales.

A medida que la tecnología evoluciona, visualizamos un futuro descentralizado y flexible, donde cada empresa o usuario pueda desplegar su propio sistema de IA personalizado, ejecutado en la nube, en servidores locales o en dispositivos personales.

El futuro de la IA será colaborativo, autónomo y sin límites.

Referencias

- [Akter et al., 2022] Akter, S., Michael, K., Uddin, M. R., McCarthy, G., and Rahman, M. (2022). Transforming business using digital innovations: The application of ai, blockchain, cloud and data analytics. *Annals of Operations Research*, pages 1–33.
- [Bousetouane, 2025] Bousetouane, F. (2025). Agentic systems: A guide to transforming industries with vertical ai agents. *arXiv preprint arXiv:2501.00881*. Version 1, CC BY 4.0 License.
- [Combinator, 2024] Combinator, Y. (2024). Vertical ai agents could be 10x bigger than saas. <https://www.ycombinator.com/library/Lt-vertical-ai-agents-could-be-10x-bigger-than-saas>. Accessed: 2024-02-23.
- [DeepSeek-AI et al., 2025] DeepSeek-AI, Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., Zhang, X., Yu, X., Wu, Y., Wu, Z., Gou, Z., Shao, Z., Li, Z., Gao, Z., Liu, A., Xue, B., Wang, B., Wu, B., Feng, B., Lu, C., Zhao, C., Deng, C., Zhang, C., Ruan, C., Dai, D., Chen, D., Ji, D., Li, E., Lin, F., Dai, F., Luo, F., Hao, G., Chen, G., Li, G., Zhang, H., Bao, H., Xu, H., Wang, H., Ding, H., Xin, H., Gao, H., Qu, H., Li, H., Guo, J., Li, J., Wang, J., Chen, J., Yuan, J., Qiu, J., Li, J., Cai, J., Ni, J., Liang, J., Chen, J., Dong, K., Hu, K., Gao, K., Guan, K., Huang, K., Yu, K., Wang, L., Zhang, L., Zhao, L., Wang, L., Zhang, L., Xu, L., Xia, L., Zhang, M., Zhang, M., Tang, M., Li, M., Wang, M., Li, M., Tian, N., Huang, P., Zhang, P., Wang, Q., Chen, Q., Du, Q., Ge, R., Zhang, R., Pan, R., Wang, R., Chen, R., Jin, R., Chen, R., Lu, S., Zhou, S., Chen, S., Ye, S., Wang, S.,

- Yu, S., Zhou, S., Pan, S., Li, S., and et al. (2025). Deepseek-r1: Incentivo a la capacidad de razonamiento en los llm mediante aprendizaje por refuerzo. *arXiv preprint arXiv:2501.12948*.
- [Europe, 2024] Europe, O. E. (2024). Agentes llm: El futuro de la ia generativa.
- [Fu et al., 2024] Fu, Y., Xue, L., Huang, Y., Brabete, A.-O., Ustiugov, D., Patel, Y., and Mai, L. (2024). Serverlessllm: Low-latency serverless inference for large language models. *arXiv preprint arXiv:2401.14351*, v2. 18th USENIX Symposium on Operating Systems Design and Implementation.
- [Gigarev, 2024] Gigarev (2024). Why vertical llm agents are the next billion dollar saas opportunity.
- [Gill et al., 2022] Gill, S. S., Xu, M., Ottaviani, C., Patros, P., Bahsoon, R., Shaghaghi, A., and Uhlig, S. (2022). Ai for next generation computing: Emerging trends and future directions. *Internet of Things*, 19:100514.
- [Hu et al., 2025] Hu, J., Xu, J., Liu, Z., He, Y., Chen, Y., Xu, H., Liu, J., Zhang, B., Wan, S., Dan, G., Dong, Z., Ren, Z., Meng, J., He, C., Liu, C., Xie, T., Lin, D., Zhang, Q., Yu, Y., Feng, H., Chen, X., and Shan, Y. (2025). Deepflow: Serverless large language model serving at scale. *arXiv preprint arXiv:2501.14417*, v2.
- [IBM, 2006] IBM (2006). An architectural blueprint for autonomic computing. White Paper 31, IBM.
- [Khoriya, 2024] Khoriya, V. (2024). The future of saas platforms: A comprehensive review. *Vidhyayana-An International Multidisciplinary Peer-Reviewed E-Journal-ISSN 2454-8596*, 10(si2):74–86.
- [Liu et al., 2023] Liu, X. et al. (2023). Agentbench: Evaluating llms as agents. <https://arxiv.org/abs/2308.03688>. arXiv preprint arXiv:2308.03688 (or arXiv:2308.03688v2 [cs.AI] for this version) <https://doi.org/10.48550/arXiv.2308.03688>.
- [Muennighoff et al., 2025] Muennighoff, N., Yang, Z., Shi, W., Li, X. L., Fei-Fei, L., Hajishirzi, H., Zettlemoyer, L., Liang, P., Candès, E., and Hashimoto, T. (2025). s1: Simple test-time scaling. *arXiv preprint arXiv:2501.19393*.
- [PluggAI, 2024] PluggAI (2024). Vertical llm agent for handling customer queries in e-commerce.
- [Sanchez et al., 2020] Sanchez, M., Exposito, E., and Aguilar, J. (2020). Autonomic computing in manufacturing process coordination in industry 4.0 context. *Journal of Industrial Information Integration*, 19:100159.
- [Shan et al., 2025] Shan, J., Gupta, V., Xu, L., Shi, H., Zhang, J., Wang, N., Xu, L., Kang, R., Liu, T., Zhang, Y., Zhu, Y., Jin, S., Lim, G., Chen, B., Chen, Z., Liu, X., Chen, X., Yin, K., Chung, C.-P., Jiang, C., Lu, Y., Chen, J., Lin,

C., Xiang, W., Shi, R., and Xie, L. (2025). Aibrix: Hacia una infraestructura de inferencia de modelos de lenguaje grandes, escalable y rentable. *arXiv preprint arXiv:2504.03648*, v1.

[Xie et al., 2023] Xie, T. et al. (2023). Openagents: An open platform for language agents in the wild. <https://arxiv.org/abs/2310.10634>. arXiv preprint arXiv:2310.10634.