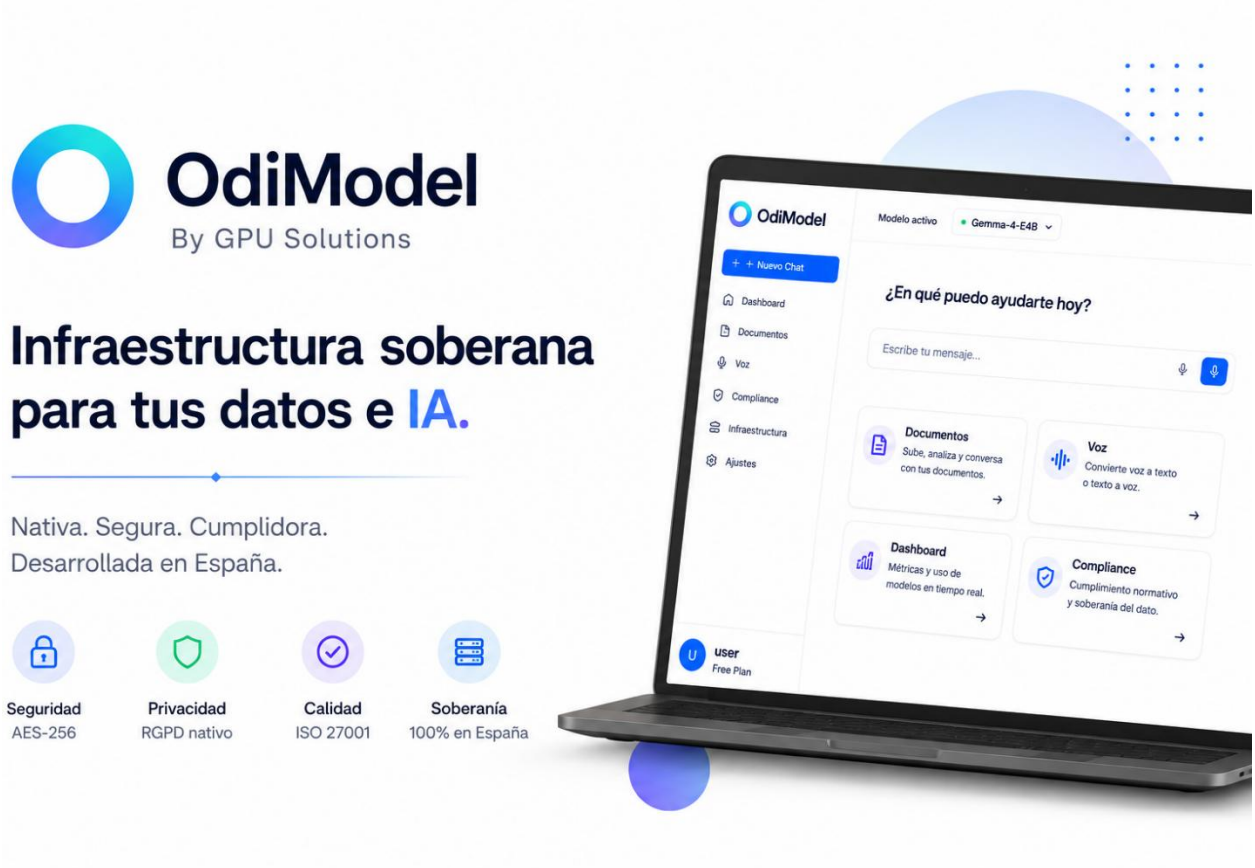

Los mejores modelos de IA a tu alcance, con tus datos siempre protegidos

OdiModel es el asistente de IA de GPU Solutions que reúne los modelos de código abierto más potentes del mercado bajo una única interfaz de chat. Gemma, Qwen, Nemotron, GLM y los que vayan llegando: el usuario elige, en cada conversación, el que mejor se adapta a cada caso (razonamiento, programación, visión, reconocimiento óptico o texto general), con búsqueda web integrada y voz bidireccional nativa, con soporte de síntesis en castellano, inglés y lenguas cooficiales (catalán, euskera y gallego). Y todo corriendo sobre infraestructura NVIDIA HGX B200 dedicada en España, con cifrado AES-256 de extremo a extremo y cumplimiento RGPD nativo: sin transferencias internacionales, sin entrenamiento con las conversaciones del cliente, sin intermediarios.

Qué hace OdiModel

1. Los mejores modelos Siempre actualizados, a tu elección	**2. Chat y voz, también en tu lengua** Castellano, inglés, catalán, euskera y gallego	**3. Tus datos en España** Infraestructura propia, cumplimiento nativo
Accede a los modelos de código abierto más potentes del mercado (Gemma, Qwen, Nemotron, GLM y los que vayan apareciendo), cada uno optimizado para una tarea distinta: texto, razonamiento, programación o visión. Y si necesitas un modelo concreto que no esté, lo desplegamos en 24 horas.	Interfaz de chat con historial persistente, búsqueda web integrada y voz bidireccional nativa en dos modos (conversación y dictado). Síntesis de voz en castellano, inglés y lenguas cooficiales españolas: el único asistente de IA con cobertura completa del Estado.	Infraestructura dedicada NVIDIA HGX B200 en centros de datos españoles, cifrado AES-256 de extremo a extremo y cumplimiento RGPD nativo. Sin transferencias internacionales ni intermediarios.

Objetivos

OdiModel existe para dar a las empresas acceso inmediato a los mejores modelos de IA disponibles, sin la fricción de desplegarlos y sin renunciar al control de sus datos.

- **Los mejores modelos, siempre actualizados.** El código abierto evoluciona a un ritmo que ningún proveedor cerrado replica. OdiModel selecciona, despliega y mantiene lo mejor de cada familia para que el cliente siempre tenga el estado del arte bajo la misma interfaz, y actualiza el catálogo a medida que aparecen modelos nuevos.

- **Si lo quieres tú, lo desplegamos en 24 horas.** La lista de modelos disponibles no es cerrada. Si un cliente necesita un modelo específico que no esté en el catálogo, el equipo técnico de GPU Solutions lo despliega en un plazo de 24 horas, siempre que cumpla condiciones de código abierto.
- **Un asistente utilizable desde el primer minuto.** Regístrate, entra y empieza: sin configuración compleja, sin integraciones previas, sin equipo de MLOps. La interfaz es un chat estándar que funciona igual para equipos técnicos y no técnicos, con voz y búsqueda web integradas.
- **Eliminar la dependencia de proveedor.** El cliente no compra a OpenAI, ni a Anthropic, ni a Google: compra acceso a los mejores modelos de código abierto disponibles en cada momento, operados por un proveedor español con infraestructura propia.
- **Y tus datos, siempre en España.** Cada vez que un empleado escribe en un chat de IA, envía información a un servidor. Con los proveedores dominantes, ese servidor está fuera de la UE y las conversaciones pueden usarse para entrenamiento. OdiModel invierte el contrato: datos en España, en servidores propios, sin entrenar con lo que escribes.

Para quién es

Empresas con datos sensibles	Organizaciones en sectores regulados (administración pública, sanidad, defensa, banca, seguros, jurídico) que no pueden enviar mensajes ni documentos a servicios fuera de la UE.
Administraciones autonómicas y organizaciones plurilingües	Administraciones de Cataluña, País Vasco, Galicia y demás territorios con lengua cooficial, así como empresas, medios y servicios públicos que operan en ellas y necesitan asistente por voz en la lengua del usuario final.
Equipos de producto y negocio	Usuarios no técnicos que necesitan un asistente fiable para redacción, análisis, resúmenes, traducción y búsquedas, sin preocuparse por qué modelo hay detrás.
Equipos técnicos y de I+D	Desarrolladores, data scientists y equipos de ML que quieren comparar modelos de código abierto sobre el mismo mensaje antes de integrarlos en sus pipelines internos.

Clientes y socios de GPU Solutions	Clientes del ecosistema GPU Solutions que quieren añadir una capa de asistente de IA sobre la infraestructura soberana que ya consumen, sin adquirir licencias de proveedores cerrados.

Catálogo de modelos

Estos son algunos de los modelos disponibles hoy en OdiModel. El catálogo es abierto y se amplía de forma continua a medida que aparecen modelos nuevos relevantes. Si un cliente necesita uno específico que no esté disponible, el equipo técnico de GPU Solutions lo despliega en un plazo de 24 horas siempre que cumpla condiciones de código abierto.

Gemma-4-E4B (4B)	Razonamiento, programación y generación de texto general. Modelo ligero por defecto para tareas cotidianas.
Qwen3-14B y Qwen3.5-9B	Razonamiento y texto. Familia Qwen para conversación general y análisis en varios idiomas.
Nemotron-VL-8B (8B)	Visión-lenguaje: reconocimiento óptico, análisis de imágenes, extracción de información desde documentos escaneados o capturas.
Nemotron-30B	Razonamiento complejo, matemáticas y tareas agénticas, para preguntas que requieren cadenas de razonamiento más largas.
GLM-5.1-FP8 (754B)	Razonamiento, programación y tareas agénticas de gran escala. Modelo frontera del catálogo.
Otros y a petición	El catálogo se amplía de forma continua con los modelos de código abierto más relevantes que van apareciendo. Modelos específicos bajo petición, desplegados en 24 horas si cumplen condiciones de código abierto.

Capacidades integradas en la interfaz

Búsqueda web en vivo	Consulta de información actualizada con citación de fuentes dentro de la propia conversación. Activable por mensaje.
Voz bidireccional	Dos modos nativos: hablar con el asistente como si fuera una conversación telefónica o dictar preguntas y leer respuestas. Soporte de voz en lenguas cooficiales (catalán, euskera y gallego). Detalle en la sección dedicada más abajo.
Documentos	El usuario sube documentos e imágenes para analizarlos con los modelos de visión-lenguaje del catálogo, y gestiona su biblioteca desde una pantalla dedicada: ver contenido extraído, buscar por nombre y eliminar definitivamente cualquier archivo en cualquier momento. Control total sobre qué vive en la cuenta y qué no.
Historial y archivado	Conversaciones persistentes por usuario, agrupadas por día, con función de archivado y borrado manual en cualquier momento. Accesibles solo por el titular de la cuenta.
Panel de uso	Dashboard propio donde el usuario ve en tiempo real sus conversaciones, mensajes y consumo de tokens desglosado por modelo, con filtros por día, semana, mes y total. Permite entender qué modelos se están usando más y a qué coste, tanto individualmente como a nivel de equipo.
Multilingüe	Interfaz disponible en español. Los modelos operan nativamente en español, inglés y otros idiomas soportados por cada familia.

Voz: dos modos y cobertura en lenguas cooficiales

OdiModel incorpora voz como ciudadano de primera, no como añadido. El usuario puede elegir entre dos modos según el caso de uso, y la síntesis de voz cubre no solo el castellano y el inglés, sino también las lenguas cooficiales del Estado: un diferenciador pensado para administraciones autonómicas, medios de comunicación y cualquier organización con operación real en Cataluña, País Vasco o Galicia.

Modo conversación	Hablar con el asistente como si fuera una llamada: el usuario habla, el modelo entiende, responde por voz y la interacción fluye sin teclado. Ideal para recorrer coches, moverse por una fábrica o trabajar sin manos.
Modo dictado y lectura	El usuario dicta la pregunta en vez de escribirla, y puede pedir que la respuesta se le lea en voz alta mientras sigue con otra tarea. Perfecto para revisión de textos largos, accesibilidad o sesiones de trabajo prolongadas frente a la pantalla.
Reconocimiento de voz (STT)	whisper-large-v3: referencia de la industria en reconocimiento de voz abierto, con rendimiento sólido en español, inglés y las demás lenguas cubiertas por Whisper. Corre sobre las GPU propias, sin que el audio del usuario salga de la infraestructura.
Síntesis de voz (TTS)	**qwen3-tts**: voz natural para español e inglés en flujos de trabajo generales. **supertonic**: síntesis expresiva pensada para lectura de contenido largo, entrevistas y material de formación. **aHoTTS**: modelo especializado en lenguas cooficiales españolas (catalán, euskera y gallego). Convierte a OdiModel en el único asistente de IA con soporte nativo de voz para las cuatro lenguas oficiales del Estado.
Quién se beneficia	Administraciones autonómicas con obligación lingüística, servicios públicos (sanidad, educación, atención ciudadana), medios de comunicación regionales, cadenas de distribución con operaciones plurilingües y cualquier empresa que atienda clientes en Cataluña, Euskadi, Galicia o el resto de territorios con lengua cooficial.

La infraestructura por debajo

OdiModel corre sobre la infraestructura soberana de GPU Solutions en España. Cada capa está seleccionada para que los datos, el cómputo y las claves permanezcan dentro de la UE de extremo a extremo.

Cómputo	NVIDIA DGX B200: nodos de 8 × B200 con placas HGX, dedicados a clientes empresariales, sin compartición en GPU.
Cifrado	AES-256 de extremo a extremo, en tránsito y en reposo. Claves gestionadas en la UE.
Red	Red troncal de alto ancho de banda dentro del centro de datos, NVLink intranodo para inferencia sobre modelos grandes (GLM-5.1-FP8).
Orquestación	Aislamiento por cuenta a nivel de sesión, cuota y almacenamiento. El historial y los archivos de un cliente nunca tocan el contexto de otro.
Ubicación	Centros de datos certificados en España. Residencia de datos en la UE. Baja latencia para usuarios ibéricos y del sur de Europa.
Cumplimiento	RGPD nativo. ENS (Esquema Nacional de Seguridad). ISO 27001 certificada por EQA. Empresa española bajo legislación europea.
Soporte	Canal directo conectado a Jira. El equipo técnico de GPU Solutions atiende las solicitudes en un plazo de 24 horas hábiles.
Afiliaciones	NVIDIA B200 y NVIDIA Inception.

Estado y salida al mercado

- **Estado.** Beta pública. La plataforma está viva en odimodel.gpusolutions.ai, con alta abierta mediante plan gratuito sin compromiso. Al tratarse de un servicio en beta, las funcionalidades y modelos disponibles pueden cambiar sin previo aviso mientras iteramos con los usuarios.
- **Modelo comercial.** Plan gratuito para pruebas durante la fase beta. Los planes de pago (individual, equipo y empresa) se anunciarán al cerrar la beta, con precios alineados al consumo real sobre la infraestructura HGX B200 en Madrid.
- **Posicionamiento.** La pregunta que OdiModel plantea al mercado («tu empresa ya usa IA, ¿dónde están esos datos?») no es retórica. Los competidores dominantes (OpenAI, Anthropic, Google, Microsoft) operan fundamentalmente fuera de la UE. OdiModel no es una alternativa más barata ni más potente que ellos: es una alternativa que mantiene los datos en casa.
- **Diferenciador lingüístico.** El soporte nativo de voz en castellano, catalán, euskera y gallego abre un nicho que los asistentes de los grandes proveedores no cubren bien: administraciones autonómicas, servicios públicos, medios y empresas con operación real en las cuatro lenguas oficiales del Estado. Para estos clientes, OdiModel es la única opción en el mercado que suma soberanía del dato y cobertura lingüística completa.
- **Distribución.** Directo a empresas españolas y europeas a través de los canales de GPU Solutions. Orientado a organizaciones con requisitos explícitos de soberanía del dato, RGPD o ENS. Abierto a integraciones con socios del ecosistema NVIDIA Inception.
- **Por qué ahora.** Los modelos de código abierto competitivos (Qwen, Gemma, GLM, Nemotron) han alcanzado paridad con los cerrados en muchas tareas. Al mismo tiempo, la presión regulatoria europea (RGPD, Reglamento de IA, ENS) hace que depender de proveedores extracomunitarios sea cada vez más difícil de justificar. OdiModel llega justo en esa intersección.

Cómo se usa: ciclo de una conversación

- **1. Regístrate y entra.** El usuario crea su cuenta y accede al chat. Sin configuración compleja ni integraciones previas: en minutos está listo para lanzar su primera pregunta.
- **2. Selecciona modelo.** Desde el desplegable superior, elige el modelo apropiado para la tarea (razonamiento, programación, visión, uso general) sin cambiar de interfaz.
- **3. Lanza la petición.** Escribiendo, dictando por voz, adjuntando documentos o imágenes, o activando la búsqueda web para contexto actualizado.

- **4. Procesamiento soberano.** La petición se procesa en un clúster NVIDIA HGX B200 con GPU dedicadas en España, cifrada con AES-256 de extremo a extremo, aislada por cuenta. No se almacena para entrenamiento, no se comparte con terceros, no sale de la infraestructura de GPU Solutions.
- **5. Respuesta en tiempo real.** La respuesta se va mostrando en la interfaz conforme se genera, con citas de fuentes web cuando aplica, y puede leerse en voz alta mediante el modelo de texto a voz integrado.
- **6. Historial y continuidad.** La conversación queda archivada en el panel lateral del usuario (accesible solo por él, salvo requerimiento legal) y puede retomarse o archivarse en cualquier momento.

Decisiones de diseño

- **La gravedad del dato se queda en España.** Conversaciones, documentos subidos, peticiones y resultados viven en infraestructura de GPU Solutions en territorio español. Nada transita por un proveedor extracomunitario en ningún punto del ciclo: ni el cómputo, ni el almacenamiento, ni la gestión de claves.
- **Aislamiento por cliente por construcción.** Cada cliente empresarial opera sobre un clúster HGX B200 con GPU dedicadas, con almacenamiento y cuotas aisladas a nivel de cuenta. No hay coubicación en GPU ni en el historial conversacional: la carga de un cliente no afecta la latencia ni la privacidad del resto.
- **Soberanía del dato por defecto, no como opción.** No hay un modo «soberano» que haya que activar: es el único modo que existe. Las garantías (RGPD, ENS, ISO 27001, cifrado AES-256) están en el diseño del producto, no en un plan premium.
- **Agnóstico de proveedor, opinativo con el catálogo.** No casamos al cliente con ningún proveedor. Sí seleccionamos qué modelos de código abierto merecen estar en el catálogo, retirando los que quedan obsoletos y añadiendo los nuevos a medida que maduran: Gemma, Qwen, Nemotron, GLM y los que vengan.
- **Una interfaz, todos los modelos.** Cambiar de modelo no cambia la experiencia. El chat, el historial, la búsqueda web y la voz funcionan igual con un modelo de 4B o de 754B; el usuario solo percibe la diferencia en capacidades y latencia.

- **Transparencia y control del usuario.** El usuario ve en todo momento qué documentos tiene cargados, qué conversaciones ha tenido y cuántos tokens consume por modelo, con panel propio y filtros por periodo. Y puede eliminar cualquier documento o conversación en cualquier momento. Nada queda oculto, nada queda atrapado.
- **Beta visible, cambios esperables.** Comunicamos el estado beta desde la propia interfaz. La iteración rápida con usuarios reales es el motor del producto en esta fase, con compromiso explícito de comunicar cambios relevantes a los clientes empresariales.

Contacto y enlaces

Plataforma	odimodel.gpusolutions.ai
Estado	Beta pública. Alta abierta con plan gratuito, sin compromiso.
Planes	Plan gratuito durante la fase beta. Plan de empresa personalizado según las necesidades de nuestros partners.
Contacto comercial	contact@gpusolutions.ai
Entidad operadora	BIAI Technology Project S.L. (operando como GPU Solutions), España / Europa