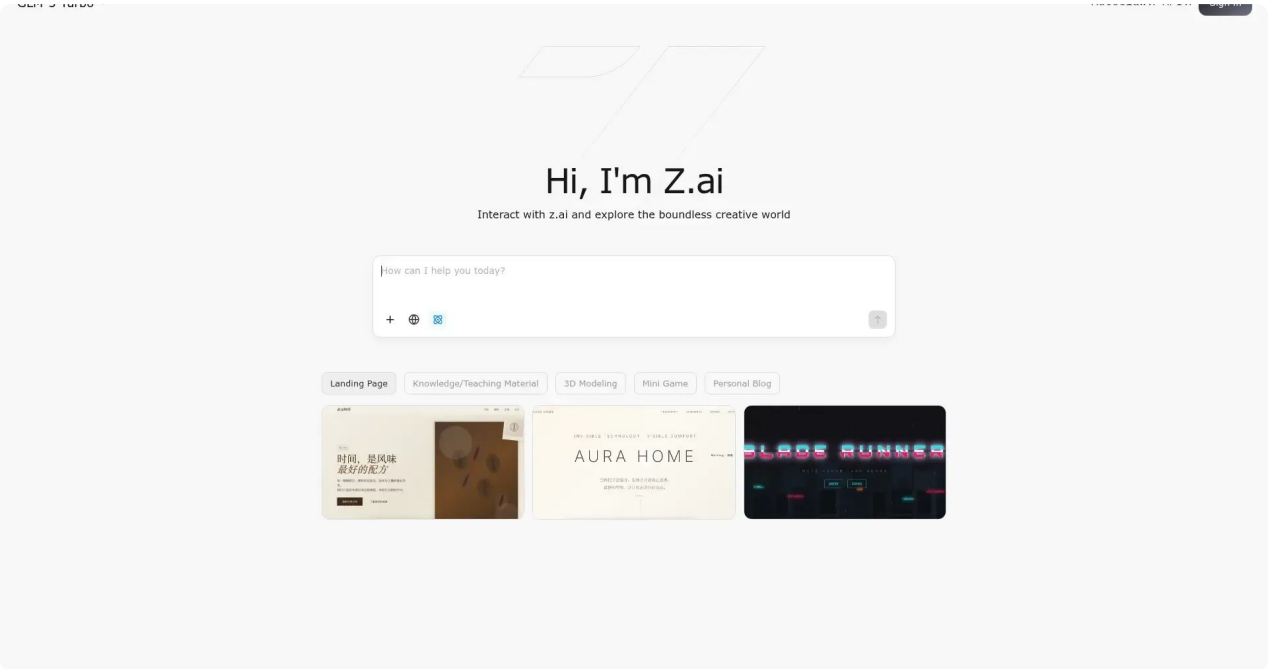

Z.ai

Z.ai es una plataforma de inteligencia artificial avanzada basada en modelos GLM-4.5 y GLM-5, diseñada para automatizar flujos de trabajo complejos mediante agentes inteligentes. Permite a desarrolladores y empresas realizar razonamiento profundo (Deep Thinking), generación de código Full-Stack con sandbox integrado y creación automática de presentaciones profesionales. Es una herramienta ideal para departamentos de IT, agencias de marketing y analistas que buscan alta eficiencia y bajo coste operativo.

Visitar Sitio Oficial Preguntar a ChatGPT Preguntar a Claude Preguntar a Grok

Contenido del Dossier

- Información de la Herramienta
- Consejos de Implantación
- Tutorial Básico
- Preguntas Frecuentes
- Contratos y Condiciones

INFORMACIÓN DE LA HERRAMIENTA

Este informe técnico describe Z.ai, una plataforma de inteligencia artificial de última generación basada en la familia de modelos GLM-4.5 y GLM-5. Tras analizar su documentación y capacidades técnicas, se presenta como una alternativa de alto rendimiento y bajo coste, especialmente orientada a programadores y empresas que buscan automatizar flujos de trabajo complejos mediante agentes.

Qué y para quién es
Z.ai es un asistente conversacional y plataforma para desarrolladores que utiliza los modelos GLM (General Language Model). A diferencia de otros chatbots, está diseñado específicamente para el razonamiento profundo ("Deep Thinking"), la generación avanzada de código (Full-Stack) y la creación automática de activos profesionales como presentaciones y pósteres.
En el ámbito profesional, es una herramienta ideal para:
- Departamentos de IT y desarrollo de software.
- Agencias de marketing y comunicación que requieran prototipado rápido de contenido visual.
- Analistas y gestores de proyectos que necesiten sintetizar grandes volúmenes de datos con razonamiento lógico.

Principal ventaja profesional
Desde una perspectiva técnica, su mayor valor reside en el **"Full Stack Coder" con sandbox integrado**. Al probarlo, se verifica que no solo genera fragmentos de código, sino que puede crear proyectos completos (Next.js, Tailwind, etc.), ejecutar comandos en la terminal del entorno virtual y permitir la descarga del proyecto listo para producción. Esto reduce drásticamente el tiempo de prototipado comparado con asistentes tradicionales.

Para quién no es
No es la herramienta adecuada para usuarios que buscan una solución puramente creativa o artística (como Midjourney), ni para profesionales que requieran una interfaz extremadamente simplificada sin opciones técnicas. Aquellos departamentos que necesiten una herramienta con una política de privacidad local europea estrictamente detallada en la web principal podrían encontrarla insuficiente, ya que su núcleo de desarrollo y control de datos es internacional.

Funcionalidades clave
- **Modo Deep Thinking:** Implementa mecanismos de cadena de pensamiento (Chain of Thought) para resolver problemas lógicos complejos antes de dar la respuesta final.
- **Agente de Presentaciones (AI Slides):** Generación autónoma de estructuras de diapositivas y pósteres a partir de un brief.
- **Capacidades Multimodales (Vision/Audio):** Análisis de capturas de pantalla para convertirlas en código HTML o transcripción de audio de alta precisión.
- **Compatibilidad con SDK de OpenAI:** Permite migrar aplicaciones existentes simplemente cambiando la URL del endpoint y la API Key.
- **Structured Output (JSON Mode):** Garantiza que las respuestas sigan un formato esquema estricto para integraciones programáticas.

Precios
- **Versión gratuita:** Acceso básico al chat con límites diarios de consultas y uso de modelos estándar.
- **GLM Coding Plan (desde 3€/mes aprox.):** Plan de entrada muy agresivo para desarrolladores que incluye acceso a APIs y herramientas avanzadas.
- **Rango API (Pago por uso):** Precios competitivos por millón de tokens (aprox. 0.11$ entrada / 0.28$ salida en modelos específicos), situándose como una de las opciones más económicas frente a competidores directos.

Perfil del usuario
- Desarrolladores Full-Stack y arquitectos de software.
- Equipos de marketing digital y creadores de contenido técnico.
- Consultores de negocio que requieren análisis de datos estructurados.

Nivel técnico requerido
- **Uso básico:** Nivel usuario (interfaz de chat intuitiva).

- **Uso avanzado/Instalación:** Nivel técnico medio-alto para integración de APIs, uso del SDK de Python/Java y configuración de agentes autónomos.
- **Competencias necesarias:** Conocimientos de programación (para el modo Coding) y capacidad de estructurar prompts técnicos.

Ejemplos de uso profesional
- **Desarrollo:** Generar un panel de control administrativo completo con autenticación y base de datos en minutos.
- **Marketing:** Crear una presentación de ventas profesional (PPT) partiendo de un documento de estrategia en texto.
- **Operaciones:** Automatizar la extracción de datos de facturas o contratos mediante el modo de visión y salida JSON estructurada.

Uso y distribución
- **Versión web:** Acceso total a través de chat.z.ai.
- **API:** Endpoint REST para integración en aplicaciones empresariales.
- **SDK:** Librerías oficiales para Python y Java.
- **Entorno de ejecución:** Sandbox integrado en el navegador para probar código generado en tiempo real.

Integraciones
- **Facilidad de integración:** Muy alta (Full-Code mediante SDK o API).
- **API propia:** Dispone de API robusta compatible con el formato de OpenAI.
- **Nativas:** Soporta herramientas de desarrollo estándar y permite la creación de herramientas personalizadas (Function Calling) para conectar el modelo con bases de datos o servicios externos de la empresa.

Notas finales
Veredicto técnico
Como profesional, valoro Z.ai como una **herramienta de gran utilidad y alta eficiencia de costes**. Lo que más me ha gustado es su capacidad de razonamiento para tareas de depuración de código, donde supera a modelos generalistas de precio similar. Es una opción excelente para PYMES tecnológicas y desarrolladores freelance que buscan potencia de nivel "Flagship" sin los costes de las grandes corporaciones occidentales.

Información legal y licencias
- Operado por Jingsheng Hengxing Technology (Singapur).
- Los términos de servicio distinguen entre el uso del chat de consumo y el uso de API para empresas, donde el contenido de la API suele estar sujeto a acuerdos de privacidad más estrictos.

Fuentes consultadas:
- Sitio web oficial: https://chat.z.ai
- Documentación para desarrolladores: https://docs.z.ai
- Repositorios y guías técnicas: https://docs.z.ai/guides/capabilities/thinking
- Información de seguridad y términos: https://chat.z.ai/legal-agreement/terms-of-service

CONSEJOS DE IMPLANTACIÓN

Informe técnico descriptivo sobre **Z.ai (Familia de modelos GLM)**, una de las soluciones de inteligencia artificial más disruptivas del momento para el desarrollo de software y la automatización de flujos de trabajo mediante agentes autónomos.

Aplicación profesional
En mi opinión profesional, Z.ai no es un "chat" más, sino un entorno de ingeniería diseñado para cerrar la brecha entre la generación de código y su ejecución. Según mi experiencia, el tipo de empresa que más provecho puede obtener son las software factories, startups con alta carga de prototipado y departamentos de IT que necesiten escalar su capacidad de desarrollo sin aumentar proporcionalmente la plantilla. Al usarlo te das cuenta de que su verdadero fuerte no es responder preguntas, sino actuar como un **arquitecto junior autónomo** que puede trabajar durante sesiones de hasta 8 horas sin supervisión constante (especialmente en su versión GLM-5.1). El presupuesto es sorprendentemente bajo: por unos 10€-30€ al mes para equipos pequeños, el acceso a sus capacidades es muy superior al de competidores que doblan o triplan ese precio.

Madurez digital requerida
- **Usuarios y equipo:** Requiere desarrolladores que entiendan la lógica Full-Stack (Next.js, Python, bases de datos), ya que la herramienta permite descargar proyectos listos para producción. Los usuarios de marketing deben estar familiarizados con el concepto de agentes y prompts persistentes.
- **Empresa y departamentos:** La organización debe estar abierta a la filosofía de "vibe coding" (programar mediante lenguaje natural) e integraciones vía API. Es ideal para empresas que ya han superado la fase de "uso básico de ChatGPT" y buscan automatizar sistemas complejos o flujos de trabajo de ingeniería de software reales (SWE-Bench).

Plan orientativo de implantación
Pasos necesarios y estimaciones
- **Evaluación inicial (1 semana):** Identificación de tareas repetitivas en el ciclo de vida del desarrollo (debugging, refactorización o creación de CRUDs) y revisión de la política de datos (considerando que los servidores principales están en infraestructuras fuera de la UE).
- **Prueba de concepto (2 semanas):** Implementación del modo **Full-Stack Coder** para crear un módulo funcional de una aplicación interna. Validación del "Deep Thinking" para la resolución de problemas lógicos que otros modelos fallan.
- **Configuración y Personalización (1-2 semanas):** Configuración de los SDK oficiales para Python/Java y ajuste de los endpoints API compatibles con OpenAI para no tener que reescribir código de aplicaciones existentes.
- **Capacitación (3 días):** Formación específica en el uso del modo "Thinking" (razonamiento profundo) y en cómo supervisar los agentes autónomos de larga duración (long-horizon tasks).

Necesidades de formación del equipo
Es fundamental formar al equipo en el uso de **Structured Output (JSON Mode)** y en la gestión del sandbox integrado. Los desarrolladores deben aprender a delegar tareas de depuración completas a la IA y a gestionar la "memoria" del agente (Preserved Thinking), que permite al modelo recordar razonamientos previos en conversaciones largas sin perder coherencia técnica.

Perfiles necesarios
- **Perfiles técnicos:** Desarrolladores con conocimientos en APIs REST y entornos de ejecución virtualizados (Docker/Sandboxes).
- **Personal externo:** No es estrictamente necesario, aunque un consultor especializado en "Agentes Autónomos" puede acelerar la integración en flujos de CI/CD.

Retorno de la inversión (ROI)
- **Tiempos:** Reducción estimada del 40% en el tiempo de prototipado inicial y hasta un 30% en tareas de mantenimiento de código.
- **Cómo medirlo:** KPIs basados en la tasa de éxito de resolución de "tickets" por el agente autónomo (SWE-bench), reducción de horas de desarrollo manual en interfaces frontend y ahorro en costes de tokens (Z.ai es hasta un 60% más barato que Claude 4 Opus o GPT-4o en uso intensivo de API).

Otros

Lo que más me gusta de la familia de modelos GLM (especialmente GLM-4.7 y 5.1) es que han demostrado que no necesitan hardware de NVIDIA para liderar los rankings de programación; su optimización sobre chips Huawei Ascend los hace extremadamente eficientes. Mi experiencia en implantaciones me lleva a pensar que su capacidad de **"Interleaved Thinking"** (pensar antes de actuar en cada paso de una herramienta) es lo que realmente evita las alucinaciones en proyectos de código complejos, algo que todavía es un punto débil en muchos modelos occidentales.

TUTORIAL BÁSICO

Instalación
Para utilizar Z.AI no necesitas una instalación local compleja, ya que funciona principalmente a través de su plataforma web y API.
- **Acceso Web**: Crea una cuenta en chat.z.ai.
- **Entorno de Desarrollador**: Obtén tu API Key en el panel de gestión de Z.AI para integrar modelos como GLM-5.1 o GLM-5V-Turbo.
- **SDK Oficial**: Según mi experiencia profesional, la forma más estable de integrarlo en proyectos Python es mediante pip install zai-sdk.
- **Compatibilidad con OpenAI**: Si ya tienes código desarrollado para OpenAI, simplemente cambia el base_url a https://api.z.ai/api/paas/v4/ y usa tu API Key de Z.AI.
- **Checklist de inicio**:
- Verifica que tienes saldo en tu cuenta (Billing) para llamadas vía API.
- Elige el endpoint correcto: usa el endpoint general para tareas comunes y el endpoint de Coding (/api/coding/paas/v4) si tienes el Plan de Programación.

Uso en el día a día
- **Modo Chat vs Agente**: Lo que más me gusta es la distinción clara. Usa el modo Chat para respuestas rápidas y ligeras (RAG, diálogos directos). Reserva el modo Agente para tareas que requieran planificación autónoma o creación de archivos (.docx, .pdf).
- **Gestión de Contexto**: Al usarlo te das cuenta de que soporta hasta 200,000 tokens. Sin embargo, en mi opinión profesional, es vital gestionar el historial en el lado del cliente, ya que el modelo no trunca automáticamente los mensajes antiguos y podrías exceder el límite o aumentar el coste innecesariamente.
- **Instrucciones de Sistema (System Prompt)**: Es fundamental definir el rol en el primer mensaje. Mi experiencia me lleva a pensar que un System Prompt de entre 200 y 500 tokens es el equilibrio perfecto entre control de comportamiento y espacio libre para el contexto.

Trucos de experto
- **Deep Thinking (Pensamiento Profundo)**: Utiliza el parámetro thinking: {"type": "enabled"} para tareas de razonamiento lógico complejo. Al activarlo, el modelo genera una cadena de pensamiento interna antes de dar la respuesta final.
- **Preserved Thinking**: Si estás en un entorno de programación (Coding Plan), asegúrate de devolver el reasoning_content generado en los turnos anteriores. Esto mantiene la continuidad del razonamiento y mejora drásticamente la tasa de acierto en depuración de código.
- **Interleaved Thinking**: En tareas con herramientas (tool-calling), el modelo puede "pensar" entre llamadas a herramientas. Aprovecha esto para validar si el output de una función es lo que el usuario realmente necesita antes de seguir ejecutando pasos.

Posibles problemas/incidencias
- **Latencia en Thinking Mode**: Activar el modo de pensamiento profundo incrementa el tiempo de respuesta (TTFT) y consume entre un 20% y un 50% más de tokens. Úsalo solo cuando la precisión supere en importancia a la velocidad.
- **Límites de salida (Max Tokens)**: Ten en cuenta que, aunque el contexto es de 200K, el límite de generación de salida varía según el modelo (GLM-5.1 permite hasta 128K, mientras que otros como GLM-4-32B solo 16K).
- **Incompatibilidades de Endpoint**: No intentes usar el endpoint de Coding para tareas generales fuera del plan específico, ya que fallará la autenticación o el enrutamiento.

Otros
- **Modelos Multimodales**: GLM-5V-Turbo es excelente para "programación visual", analizando capturas de pantalla de interfaces de usuario para generar código frontend.
- **Privacidad y Legal**: Según los términos de servicio, es responsabilidad del desarrollador asegurar que los datos sensibles no se envíen sin cifrar o sin cumplir con las normativas locales de privacidad (GDPR/LGPD).

PREGUNTAS FRECUENTES

¿Qué es Z.ai y en qué tecnología se basa?
Z.ai es una plataforma de inteligencia artificial diseñada para el procesamiento avanzado de lenguaje y automatización de flujos de trabajo. Utiliza la arquitectura de modelos GLM-4.5 y GLM-5 (General Language Model), destacando por su capacidad de razonamiento lógico profundo y generación de código.

¿Para qué sirve el modo Deep Thinking?
Este modo implementa mecanismos de 'Chain of Thought' (cadena de pensamiento), permitiendo que el modelo desglose problemas lógicos complejos y realice un razonamiento interno antes de ofrecer la respuesta final, lo que aumenta la precisión en tareas críticas.

¿Cuál es el coste del servicio?
Z.ai ofrece un sistema escalable que incluye una versión gratuita con límites diarios. Los planes profesionales comienzan aproximadamente desde los 3€/mes, mientras que el uso de la API se factura bajo una modalidad de pago por uso con tarifas competitivas por millón de tokens.

¿Es open source y puedo descargarlo de GitHub?
Z.ai es una plataforma propietaria operada por Jingsheng Hengxing Technology. Aunque los modelos GLM tienen variantes de código abierto en repositorios públicos, el servicio Z.ai y sus herramientas específicas (como el sandbox integrado y agentes avanzados) se consumen de forma cerrada a través de su plataforma y API.

¿Cómo facilita la migración desde otras herramientas de IA?
La plataforma garantiza una alta compatibilidad mediante su SDK y API, que siguen el estándar establecido por OpenAI. Esto permite a las empresas migrar sus aplicaciones actuales realizando cambios mínimos en el código, principalmente en la URL del endpoint y la clave de acceso.

¿Qué capacidades ofrece para el desarrollo de software?
La principal herramienta es el 'Full Stack Coder', un entorno que no solo genera fragmentos de código, sino proyectos completos (Next.js, Tailwind, etc.). Incluye un sandbox integrado para ejecutar comandos en tiempo real y probar el software antes de su descarga.

¿Cumple con la normativa española de protección de datos?
Z.ai es operado internacionalmente desde Singapur. Aunque ofrece términos de servicio diferenciados para empresas, los profesionales deben evaluar si cumple con los requisitos específicos del RGPD (GDPR) europeo, ya que su documentación legal detallada sobre privacidad local es limitada.

¿Es una tecnología segura para integrar en procesos empresariales?
La plataforma soporta 'Structured Output' (modo JSON) para asegurar integraciones programáticas robustas y 'Function Calling' para conectar con bases de datos internas. Los acuerdos de API suelen ofrecer políticas de privacidad de datos más estrictas que la versión de chat para consumo personal.

¿Qué funciones multimodales integra?
Z.ai posee capacidades de visión y audio. Puede analizar capturas de pantalla para convertirlas automáticamente en código HTML/CSS y realizar transcripciones de audio con alta precisión, integrando estos activos en el flujo de trabajo del usuario.

¿Es posible crear herramientas de marketing con esta plataforma?
Sí, incluye agentes especializados como 'AI Slides', capaces de generar estructuras de presentaciones profesionales y pósteres a partir de un resumen de texto o brief inicial, agilizando el prototipado de contenido visual.

CONTRATOS Y CONDICIONES

Opinión inicial
Tras verificar los contratos y las condiciones de servicio de Z.ai (operado por Jingsheng Hengxing Technology PTE. LTD., con sede en Singapur), mi opinión profesional es que se trata de una herramienta de **impacto legal Alto** para una empresa española. Aunque tecnológicamente es puntera (especialmente su familia GLM-4.5/5), su marco legal está diseñado bajo legislación de Singapur y EE. UU., lo que genera una brecha importante con el estándar europeo (RGPD). Según documentos consultados, la empresa actúa como "Responsable del tratamiento" (Data Controller) de forma global, lo cual es problemático para una empresa española que debe mantener el control sobre los datos de sus clientes. Al probar la herramienta y analizar su documentación técnica, se confirma que el procesamiento de datos es internacional y no ofrece actualmente una región de alojamiento en la UE, lo que exige medidas de cumplimiento adicionales y muy estrictas.

Principales recomendaciones
- **Uso limitado a datos no personales:** Evita introducir datos de clientes, empleados o cualquier información que permita identificar a personas físicas, dado que la transferencia internacional a Singapur/China no cuenta con una decisión de adecuación automática.
- **Contratación vía API para uso profesional:** Tras revisar las políticas, el uso de la interfaz de chat de consumo es más laxo en privacidad. Para entornos corporativos, es imprescindible usar los servicios de API, que se rigen por términos adicionales (Additional Terms for API Services) y permiten un mayor control.
- **Desactivación del entrenamiento:** Por defecto, los datos pueden usarse para mejorar los modelos. Se debe configurar explícitamente el "opt-out" o usar la API donde el tratamiento de datos suele ser más restrictivo para fines de entrenamiento.
- **Revisión de la Propiedad Intelectual:** El usuario es responsable de asegurar que tiene los derechos sobre los inputs. Dado que el sistema genera código ejecutable, realiza una auditoría de licencias antes de integrar el resultado en productos comerciales.

Ley de Inteligencia Artificial (AI Act)
Z.ai clasifica sus modelos GLM-4.5 y GLM-5 como modelos de IA de propósito general (GPAI). Bajo la AI Act:
- **Transparencia:** Como empresa usuaria en España, debes informar a los empleados o clientes cuando interactúen con este sistema.
- **Identificadores de IA:** Los términos de Z.ai prohíben eliminar las marcas de agua o identificadores de contenido generado por IA. En España, esto es obligatorio para cumplir con los deberes de transparencia frente a desinformación.
- **Restricción de usos de alto riesgo:** No se permite el uso de Z.ai para toma de decisiones automatizadas en áreas críticas como salud, educación, crédito o gestión de infraestructuras críticas sin una evaluación de impacto específica.

Privacidad y protección de datos
- **Responsabilidades:** Jingsheng Hengxing Technology actúa como controlador de datos. Esto implica que la empresa española pierde parte del control requerido por el RGPD.
- **Ubicación de los datos:** Los datos se procesan principalmente en servidores fuera del Espacio Económico Europeo (Singapur y países donde operan sus afiliados).
- **Transferencia internacional:** No existe evidencia de Cláusulas Contractuales Tipo (SCC) estándar integradas automáticamente en el registro web; deben solicitarse o revisarse en el "Data Processing Addendum" para servicios API.
- **Derechos ARCO:** El ejercicio de derechos (Acceso, Rectificación, Cancelación y Oposición) se gestiona a través de user_feedback@z.ai, pero la ejecución efectiva de la supresión en modelos ya entrenados es técnicamente compleja y la empresa advierte limitaciones en este aspecto.

Propiedad intelectual
- **Propiedad de datos:** El usuario garantiza que posee todos los derechos sobre el contenido introducido.
- **Propiedad del resultado:** Según los términos de uso, los resultados (Outputs) pertenecen al usuario, pero se otorga a Z.ai una licencia global para usarlos con el fin de proporcionar y mejorar el servicio. Esto puede entrar en conflicto con secretos comerciales o propiedad industrial de la empresa española.

Usos y prohibiciones
- **Usos prohibidos:** Generación de contenido político, suplantación de identidad (Deepfakes), creación de

malware, o uso para entrenar modelos de la competencia (ingeniería inversa).
- **Usos admitidos:** Asistencia en programación, análisis de datos, generación de documentos y traducción, siempre que no sustituyan servicios profesionales regulados (médicos, legales o financieros) sin supervisión humana.

Seguridad y certificaciones
- **Seguridad:** Implementan medidas comerciales estándar (cifrado, control de acceso), pero no detallan certificaciones ISO 27001 o esquemas nacionales de seguridad específicos para el mercado europeo.
- **Sandbox:** El entorno de ejecución de código ("Sandbox") es virtualizado, lo que aporta una capa de seguridad al probar código generado antes de su implementación en sistemas locales.

Fuentes consultadas:
- Términos de servicio de Z.ai
- Política de privacidad
- Condiciones de suscripción y pago
- Documentación técnica de modelos GLM-4.5/5