AnythingLLM LM Studio

APRENDE A CREAR TU IA PRIVADA EN LOCAL CON TUS DATOS

miguelangelnavarro

Índice

La creciente importancia de las herramientas locales de gestión y utilización de Modelos de Lenguaje Extensos (LLMs) se ha vuelto cada vez más evidente en el panorama actual de la inteligencia artificial. Esta tendencia responde a la necesidad de los usuarios de ejercer un mayor control sobre sus datos y sus interacciones con la IA, buscando alternativas que ofrezcan privacidad, acceso sin conexión y la posibilidad de personalización.

En este documento detallamos el proceso para utilizar conjuntamente **Anything LLM** y **LM Studio** en la creación de agentes de inteligencia artificial capaces de trabajar con información propia. La combinación de estas dos herramientas permite a los usuarios construir sistemas de IA personalizados y privados, aprovechando la flexibilidad de Anything LLM para la gestión de agentes y documentos, junto con la capacidad de LM Studio para ejecutar modelos de lenguaje grandes (LLM) localmente.

1. Introducción a Anything LLM y LM Studio

Anything LLM se presenta como una aplicación de inteligencia artificial integral, diseñada para interactuar con una amplia variedad de modelos de lenguaje, documentos y agentes, todo ello manteniendo la privacidad del usuario.[1] Su principal objetivo es facilitar el uso de LLMs a usuarios sin necesidad de conocimientos de programación, ofreciendo una interfaz intuitiva y numerosas funcionalidades integradas.[1] La aplicación soporta tanto modelos de lenguaje que se ejecutan localmente como aquellos ofrecidos por proveedores empresariales, además de ser compatible con diversos formatos de documentos.[1] Una de sus características destacadas es la capacidad de crear agentes con funcionalidades personalizadas, como la búsqueda de información en documentos y la interacción con bases de datos vectoriales.[1] Anything LLM actúa como la plataforma central para la construcción de estos agentes, simplificando la gestión de datos y la comunicación con los modelos de lenguaje y las bases de datos vectoriales.

Por otro lado, **LM Studio** se describe como una interfaz de usuario, una API y un motor para la ejecución local de LLMs.[5] Esta herramienta permite a los usuarios descargar y ejecutar modelos GGUF desde **Hugging Face** directamente en sus propios equipos, utilizando tanto la CPU como la GPU para el procesamiento.[5] LM Studio también funciona como un SDK local de IA, facilitando la creación de aplicaciones sin la complejidad de gestionar dependencias subyacentes.[6] Además de su capacidad para ejecutar modelos, LM Studio ofrece una interfaz de chat amigable y funcionalidades para buscar y descargar modelos de manera sencilla.[6]

Una característica clave para su integración con **Anything LLM** es su capacidad para operar como un servidor local compatible con los endpoints de la API de OpenAI.[7] LM Studio proporciona la potencia computacional necesaria para ejecutar los modelos de lenguaje de forma local, siendo su compatibilidad con modelos GGUF de Hugging Face y su capacidad para emular la API de OpenAI aspectos fundamentales para su uso con Anything

LLM.

La ventaja de utilizar ambas herramientas en conjunto reside en su complementariedad.[9] Mientras que **Anything LLM proporciona la lógica** y las funcionalidades para la creación de agentes, la gestión de documentos y la interacción con el usuario [3], **LM Studio ofrece la capacidad de ejecutar los modelos de lenguaje de manera local**, lo que garantiza la privacidad y el control sobre los recursos computacionales.[6]

Esta combinación permite a los usuarios crear agentes personalizados que pueden acceder y procesar su propia información sin necesidad de depender de servicios en la nube, manteniendo así la confidencialidad de sus datos.[1] La sinergia entre ambas aplicaciones radica en que Anything LLM gestiona el marco del agente y los datos, mientras que LM Studio proporciona la infraestructura para ejecutar el LLM localmente, ofreciendo un equilibrio entre facilidad de uso y control para el usuario.

Es posible que algunos usuarios se pregunten por qué es necesario utilizar LM Studio si Anything LLM ya incluye soporte para la ejecución de LLMs locales.[11] Si bien Anything LLM puede simplificar el uso de modelos locales (por ejemplo, integrando Ollama en su versión de escritorio [13]), LM Studio ofrece un control más detallado sobre la selección y configuración de estos modelos, además de permitir el uso de una gama más amplia de modelos compatibles.[6] La integración con LM Studio proporciona una mayor flexibilidad y más opciones para aquellos usuarios que desean un control más avanzado sobre sus modelos de lenguaje locales.

2. Instalación de LM Studio

Para comenzar a utilizar LM Studio, es necesario asegurarse de que el sistema operativo cumpla con los requisitos mínimos. LM Studio es compatible con macOS (tanto en arquitecturas M1/M2/M3/M4 como Intel), Windows (en versiones de 64 bits x86 o ARM) y Linux (en arquitecturas x86 con procesadores que soporten AVX2).[6] En cuanto a las recomendaciones de hardware, se sugiere contar con al menos 8 GB de RAM, aunque se recomienda tener 16 GB o más para un rendimiento óptimo, especialmente al utilizar modelos de lenguaje grandes. También se requiere un espacio libre en disco de al menos 20 GB para la instalación y descarga de modelos.[15] Para usuarios de Windows y Linux, se aconseja disponer de una tarjeta gráfica (GPU) con al menos 4 GB de VRAM para acelerar el procesamiento.[16] Estos requisitos aseguran que la aplicación y los modelos de lenguaje puedan ejecutarse de manera fluida.

El primer paso para instalar LM Studio es dirigirse a su sitio web oficial.[6] En la página de descargas, se deben identificar y seleccionar la versión correspondiente al sistema operativo que se esté utilizando (Windows, macOS o Linux).[6] Es posible que el sitio web ofrezca enlaces directos para facilitar la descarga, como el enlace para Windows que se encuentra en.[6] Se recomienda revisar la sección de descarga y las opciones disponibles, tal como se muestra en la captura de pantalla del sitio web en.[15] Al iniciar la descarga, algunos navegadores podrían mostrar una advertencia indicando que se está descargando un archivo ejecutable. En este caso, se debe confirmar la descarga para continuar.[15] El sitio web oficial es la fuente principal para obtener la aplicación de LM Studio.

Una vez descargado el archivo de instalación, el proceso varía ligeramente según el sistema operativo:

- **Windows:** Se debe ejecutar el archivo .exe descargado.[13] El instalador guiará al usuario a través de los pasos necesarios para completar la instalación, como se puede observar en las capturas de pantalla del proceso de instalación en [15] y.[15] Durante la instalación, es posible que Windows Defender u otro software antivirus muestren una advertencia indicando que la aplicación no es confiable, ya que actualmente no

cuenta con una firma digital. Para continuar con la instalación, se debe hacer clic en "Más detalles" y luego en el botón "Ejecutar de todos modos".[13] Una vez finalizada la instalación, la aplicación estará disponible en el menú de inicio o en el escritorio.[15]

- **macOS:** Se debe hacer doble clic en el archivo .dmg descargado.[17] A continuación, se abrirá una ventana en la que se mostrará el icono de LM Studio. Para instalar la aplicación, se debe arrastrar este icono a la carpeta "Aplicaciones".[17] Una vez copiada, la aplicación se puede ejecutar desde la carpeta "Aplicaciones" o utilizando la función de búsqueda Spotlight (pulsando cmd + espaciobar y escribiendo "LM Studio").[17]
- **Linux:** El proceso de instalación en Linux puede variar dependiendo de la distribución. Algunas fuentes sugieren descargar el archivo desde el sitio web oficial.[14] Otra alternativa, mencionada en [18], es la instalación a través de Docker. Este método implica clonar el repositorio de LM Studio desde GitHub utilizando el comando git clone https://github.com/lm-studio/lm-studio.git y luego utilizar Docker Compose para ejecutar la aplicación. Este último método requiere tener Docker instalado en el sistema.

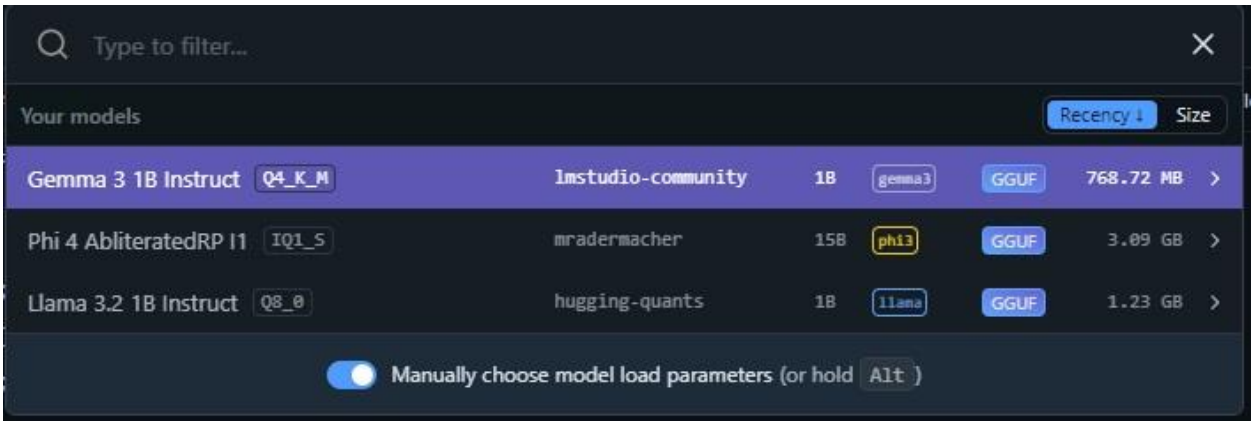
Es recomendable incluir capturas de pantalla del instalador de LM Studio en los diferentes sistemas operativos, si es posible, para ofrecer una guía visual más clara al usuario. Las capturas de pantalla del sitio web de descarga [15] y del proceso de instalación en Windows [15] son especialmente útiles.

Tabla: Requisitos del Sistema para LM Studio

Sistema Operativo	Requisitos Mínimos	Requisitos Recomendados
macOS	macOS 10.15 (Catalina) o posterior, Intel o Apple Silicon	macOS 11 o posterior, Apple Silicon (M1/M2/M3/M4)
Windows	Windows 10/11 (64-bit x86 o ARM), Procesador con AVX2	Windows 10/11 (64-bit x86 o ARM), Procesador con AVX2, 16 GB RAM, GPU con 4 GB VRAM
Linux	Distribución Linux compatible (x86), Procesador con AVX2	Distribución Linux compatible (x86), Procesador con AVX2, 16 GB RAM, GPU con 4 GB VRAM

3. Descarga de Modelos de Lenguaje (LM) para LM Studio

Los modelos de lenguaje son la base de la inteligencia artificial generativa, responsables de comprender y generar texto.[19] LM Studio permite ejecutar estos modelos directamente en el ordenador del usuario, ofreciendo privacidad y la posibilidad de trabajar sin conexión a internet.[20] Estos modelos varían en tamaño y capacidad, lo que influye en su rendimiento y los recursos computacionales que requieren.[11] La elección del modelo adecuado dependerá de las necesidades específicas del usuario y de las capacidades de su hardware.



Para buscar y descargar modelos compatibles con LM Studio, se debe abrir la aplicación y dirigirse a la pestaña **"Discover"** o "Descubrir" (el nombre puede variar según la versión).[14] En la interfaz principal, se encontrará una barra de búsqueda que permite encontrar modelos por nombre, palabras clave (como "llama", "mistral" o incluso términos en español) o pegando la URL del modelo desde Hugging Face.[7] Los resultados de la búsqueda mostrarán una lista de modelos disponibles en el repositorio de Hugging Face.[7] Al seleccionar un modelo de la lista, se podrán ver más detalles sobre sus características y requisitos.[14] Para iniciar la descarga del modelo deseado, se debe hacer clic en el botón **"Download"** o "Descargar".[14] Durante el proceso de descarga, se mostrará una barra de progreso indicando el estado.[15] Una vez completada, el modelo se guardará localmente en el ordenador del usuario.[22] La interfaz de descarga de modelos en LM Studio, tal como se muestra en capturas de pantalla como [16] y [23], facilita este proceso.

LM Studio utiliza principalmente modelos en formato GGUF, que son compatibles con la biblioteca llama.cpp, y también soporta el formato MLX para usuarios de macOS con chips de Apple.[6] Al buscar modelos, a menudo se encontrarán diferentes versiones con nombres como Q3_K_S, Q4, Q5, etc. Estas etiquetas hacen referencia a la cuantización del modelo, una técnica que comprime el tamaño del archivo del modelo a costa de una ligera pérdida de calidad.[10] Se recomienda elegir una opción de cuantización de 4 bits o superior si el sistema del usuario tiene suficiente capacidad para ejecutarlo.[21]

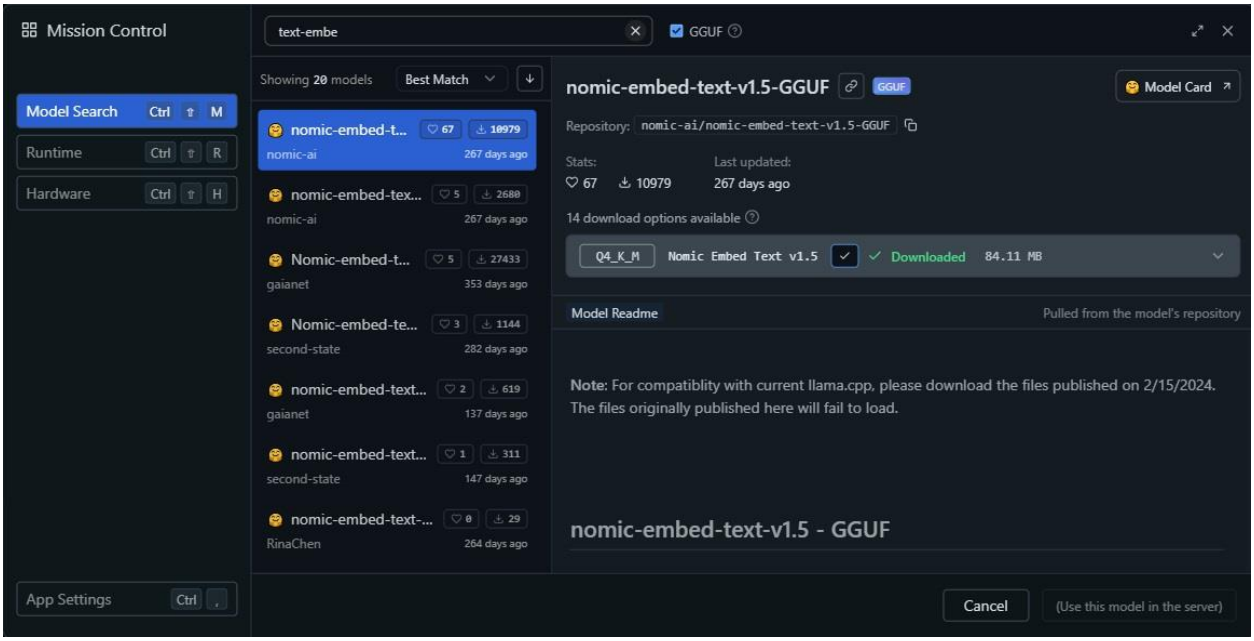
Comprender estos formatos y la cuantización ayuda a los usuarios a seleccionar los modelos más adecuados para su hardware y sus necesidades específicas. La cantidad de memoria VRAM disponible en la tarjeta gráfica del usuario influye directamente en el tamaño y el nivel de cuantización de los modelos que se pueden ejecutar de manera eficiente en LM Studio.[11]

GGUF (GPT-Generated Unified Format) es un formato de archivo diseñado para empaquetar modelos de lenguaje grandes (LLMs). Su propósito fundamental es consolidar todos los componentes esenciales de un modelo —incluyendo sus parámetros (pesos), la definición de su arquitectura, los datos del tokenizador y metadatos relevantes como plantillas de prompt— en un único fichero autocontenido. Esta estructura unificada simplifica significativamente la distribución, carga y ejecución de LLMs, particularmente en entornos locales (CPU/GPU), al eliminar dependencias externas y asegurar que toda la información necesaria viaje junta. GGUF se ha establecido como un estándar en herramientas como llama.cpp, sucediendo a formatos anteriores y ofreciendo mayor extensibilidad.

4. Descarga de Modelos de "Embedding" para Anything LLM

Los modelos de **"embedding"** son fundamentales para que Anything LLM pueda comprender el significado del texto en los documentos cargados. Estos modelos crean representaciones numéricas, conocidas como vectores, que capturan la semántica del texto.[4] Anything LLM utiliza estos "embeddings" para analizar el contenido de los documentos y realizar búsquedas semánticas relevantes.[4] Sin estos modelos, la aplicación no podría entender el significado subyacente de la información proporcionada.

Un embedding es una representación numérica de elementos discretos, como palabras, frases, documentos completos o incluso conceptos, transformándolos en vectores (listas de números) en un espacio matemático multidimensional y continuo. La propiedad fundamental de un embedding es que elementos con significados o contextos similares se ubican espacialmente cerca dentro de este espacio vectorial. Esta proximidad espacial codifica relaciones semánticas, permitiendo a los sistemas computacionales comprender, comparar y procesar el lenguaje u otros datos de manera que refleje su significado intrínseco y sus relaciones contextuales, en lugar de tratarlos como símbolos aislados.



Existen varias opciones para obtener modelos de **"embedding"** compatibles con Anything LLM, tanto de forma local como a través de servicios en la nube.

- **Opciones Locales:**
  - **LM Studio:** LM Studio puede utilizarse como proveedor de modelos de "embedding" locales.[24] Sin embargo, es importante tener en cuenta que LM Studio tiene una limitación: no puede funcionar simultáneamente como el modelo de lenguaje principal (LLM) y como el modelo de

"embedding".[24] Para utilizar LM Studio como "embedder", primero se debe descargar un modelo de "embedding" compatible desde la pestaña "Discover" de LM Studio. Este proceso es similar a la descarga de modelos de lenguaje, pero se deben buscar modelos utilizando términos como "embedding" o "instructor-large".[6] Una vez descargado el modelo de "embedding", se debe cargar explícitamente en LM Studio antes de iniciar el servidor de inferencia.[24]

- **Otras opciones:** Anything LLM también ofrece otras opciones para modelos de "embedding" locales, como "AnythingLLM Default", "Local AI" y "Ollama".[24] Si bien esta guía se centra en el uso de LM Studio, estas alternativas pueden ser consideradas por usuarios con necesidades específicas.

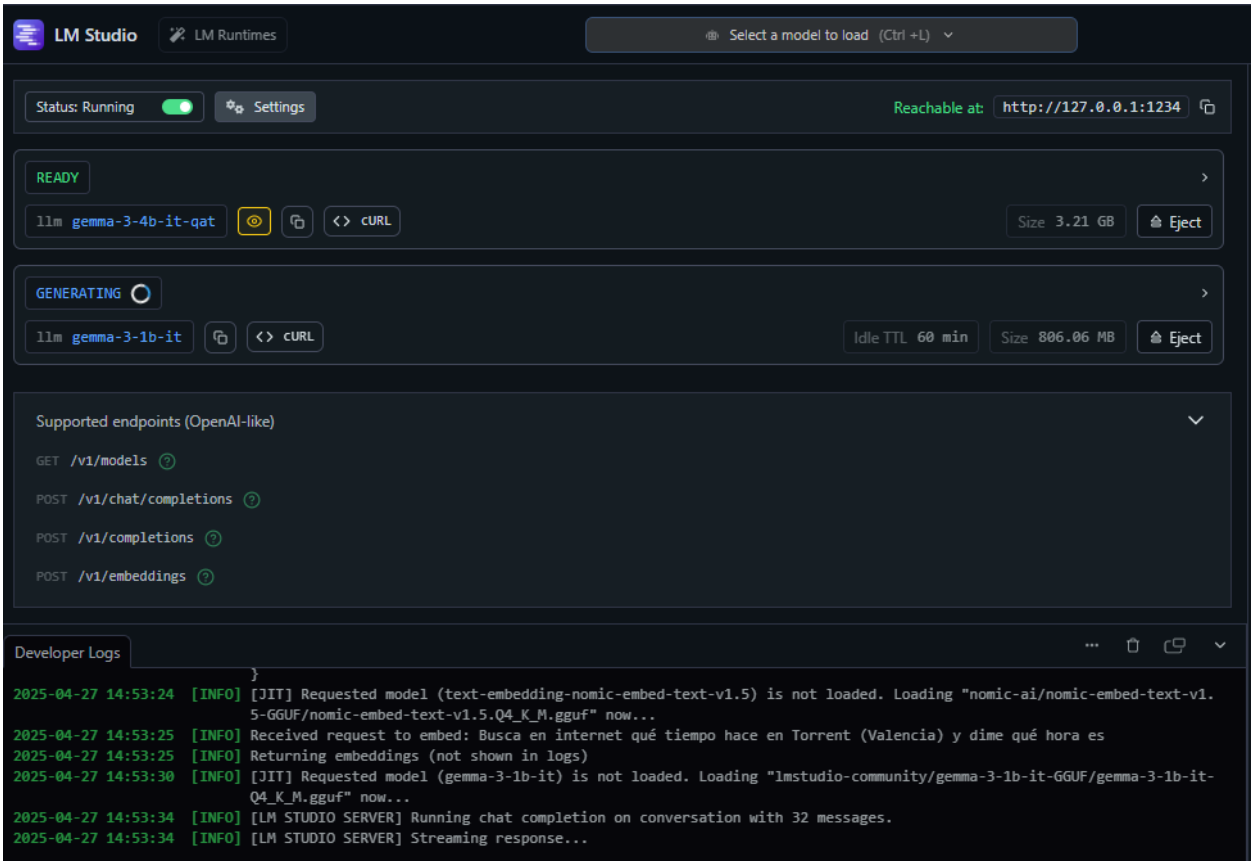
- **Opciones en la Nube:**
  - Para aquellos que prefieren utilizar servicios en la nube, Anything LLM soporta proveedores populares como OpenAI, Azure OpenAI y Cohere para modelos de "embedding".[24] El uso de estas opciones generalmente requiere una cuenta y una clave API del proveedor correspondiente.
- **Repositorios como Hugging Face:**
  - La mayoría de los modelos de "embedding" compatibles, especialmente para las opciones locales como LM Studio, Local AI u Ollama, se pueden encontrar en la plataforma Hugging Face.[24] Para buscar modelos de "embedding" en Hugging Face, se pueden utilizar términos como "embedding", "sentence-transformers" o "instructor". Es importante tener en cuenta que, para utilizar un modelo descargado de Hugging Face con LM Studio, generalmente se buscará la versión en formato GGUF del modelo.[11]

La plataforma Hugging Face se convierte así en una fuente principal para encontrar modelos de "embedding". Al navegar por su interfaz, los usuarios pueden utilizar los términos de búsqueda mencionados anteriormente para

localizar los modelos deseados.

5. Configuración de LM Studio como Servidor Local

Para que **Anything LLM** pueda comunicarse con el modelo de lenguaje ejecutado por LM Studio, es necesario configurar LM Studio para que funcione como un servidor local. El primer paso es abrir la aplicación LM Studio y dirigirse a la sección **"Local Inference Server"** (o un nombre similar, dependiendo de la versión) que generalmente se encuentra en la barra lateral izquierda de la interfaz.[26] En esta sección, se pueden configurar varios parámetros del servidor.



Uno de los parámetros más importantes es el **"Server Port"** o puerto del servidor, que especifica el número de puerto a través del cual el servidor

escuchará las conexiones entrantes. El puerto predeterminado suele ser el 1234.[26]

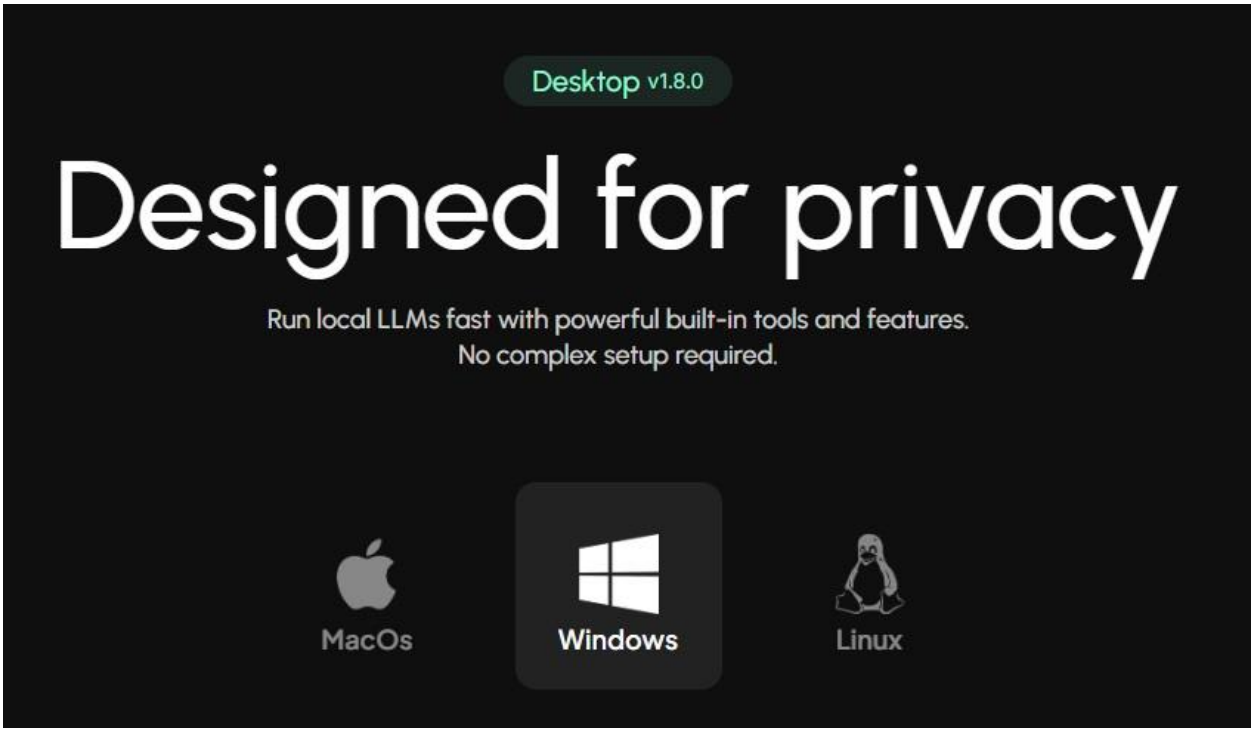
Una vez configurados los parámetros deseados, se debe hacer clic en el botón **"Start Server"**[26] Al hacerlo, LM Studio iniciará el servidor de inferencia local y mostrará un mensaje de éxito indicando la dirección en la que está funcionando, que normalmente será http://localhost:1234 si se utiliza el puerto predeterminado.[27] Es importante mantener este servidor en funcionamiento mientras se utiliza Anything LLM para que la comunicación entre ambas aplicaciones sea posible. La captura de pantalla del panel de configuración del servidor con el botón **"Start Server"** resaltado, como se muestra en [26], puede ser de gran ayuda para los usuarios. Iniciar el servidor de inferencia en LM Studio es un paso fundamental para que Anything LLM pueda acceder y utilizar los modelos de lenguaje gestionados por LM Studio.

El **"Server Port"** es crucial para la comunicación entre aplicaciones, ya que actúa como un punto de encuentro específico en la red local. Habilitar CORS podría ser necesario en escenarios donde Anything LLM y LM Studio se ejecutan en orígenes diferentes (aunque en este caso, al ejecutarse localmente, no siempre es imprescindible, pero puede serlo en ciertas configuraciones de red). Los registros detallados del servidor ("Verbose Server Logs") pueden ser útiles para la depuración en caso de problemas de conexión o rendimiento, ya que proporcionan información sobre la actividad del modelo y el servidor.

6. Configuración de Anything LLM para Conectarse a LM Studio

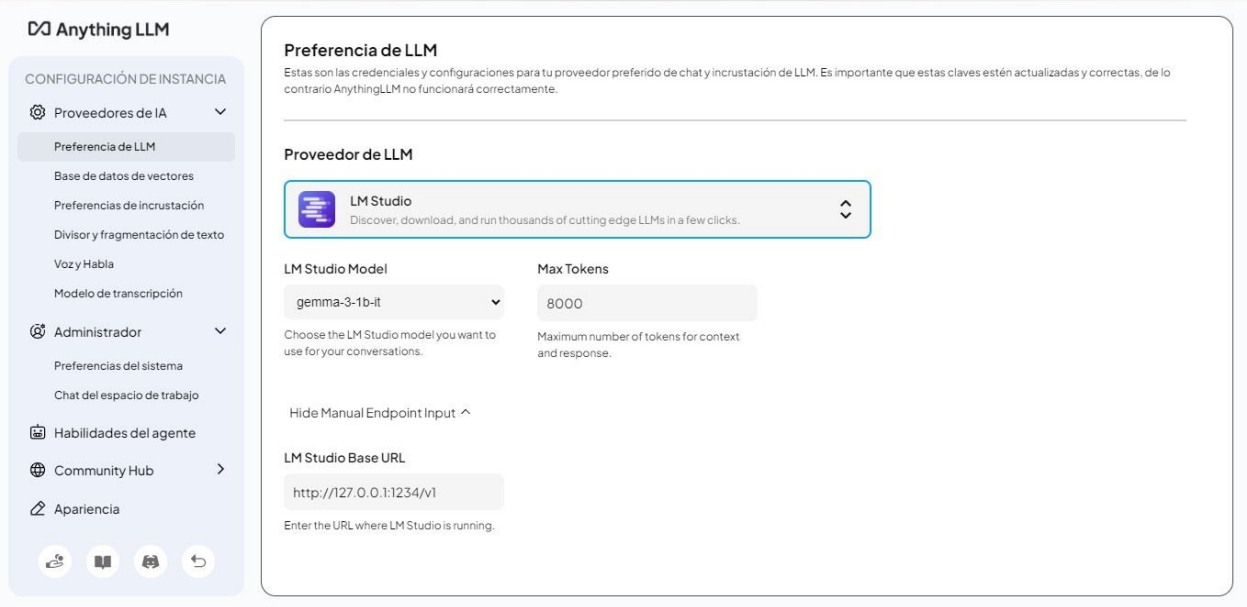
El siguiente paso es instalar la aplicación Anything LLM. Para ello, se debe acceder al sitio web oficial de Anything LLM.[1] En la página principal, se encontrará la opción para descargar la versión de escritorio (Desktop) de la aplicación.[1] Es importante mencionar que también existen opciones para la

instalación auto-hospedada (Self-hosted) y en la nube (Cloud) [1], pero esta guía se centrará en la versión de escritorio. Se debe seleccionar el instalador adecuado para el sistema operativo que se esté utilizando (Windows, macOS o Linux).[1] El sitio web puede ofrecer enlaces directos a los instaladores, como los que se encuentran en [17] para macOS[13] para Windows y [28] para Linux. Se recomienda revisar la sección de descarga y las opciones disponibles, tal como se muestra en la captura de pantalla del sitio web en.[29]



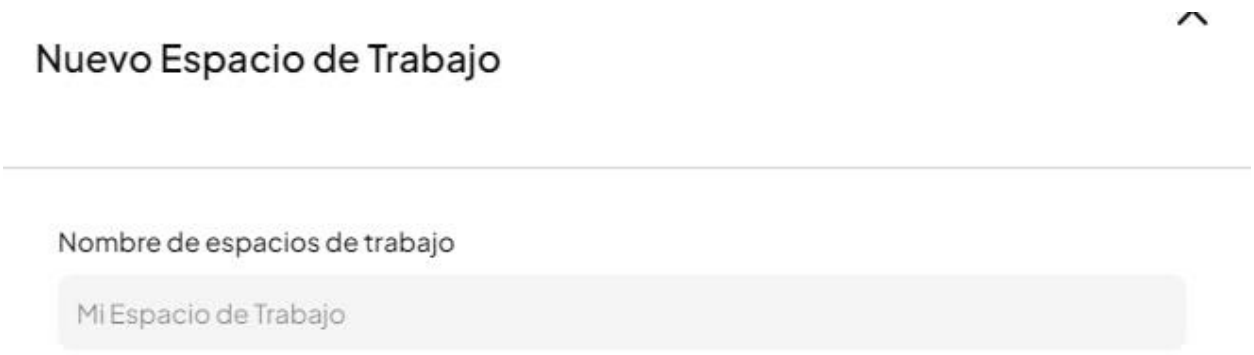
Una vez descargado el instalador, se debe ejecutar y seguir las instrucciones que aparecen en pantalla.[29] En el caso de Windows, es posible que se muestre una advertencia del antivirus indicando que la aplicación no es confiable. Si esto ocurre, se debe hacer clic en "Más detalles" y luego en "Ejecutar de todos modos" para continuar con la instalación.[13] Para usuarios de Linux, la instalación se realiza mediante un script proporcionado en el sitio web.[28] La versión de escritorio de Anything LLM ofrece una instalación simplificada con un solo clic.[1]

Una vez instalada la aplicación Anything LLM, se debe abrir para proceder con la configuración de la conexión al servidor LM Studio local. Dentro de la aplicación, se debe buscar la sección de "Settings" o "Configuración".[5] Esta sección generalmente se encuentra accesible a través de un icono o una opción en el menú principal de la aplicación, como se muestra en las capturas de pantalla.[29] Dentro de la configuración, se debe buscar la sección relacionada con la configuración del modelo de lenguaje (LLM), que puede tener nombres como "LLM", "Chat Settings" o similar.[30] En esta sección, se debe seleccionar "LM Studio" como el proveedor de LLM.[5] Esto generalmente se realiza a través de un menú desplegable o una lista de proveedores, como se ilustra en las capturas de pantalla.[30] Finalmente, se debe configurar la dirección del servidor LM Studio. Por lo general, esta dirección será http://localhost seguida del puerto que se configuró en LM Studio (el puerto predeterminado es 1234).[30] Se debe indicar en la interfaz de Anything LLM dónde ingresar esta información. Una vez ingresada la dirección y el puerto, se deben guardar los cambios en la configuración. Este proceso de conexión es fundamental para que Anything LLM pueda utilizar los modelos de lenguaje que se están ejecutando en el servidor de LM Studio.



7. Creación y Configuración de Espacios de Trabajo en Anything LLM

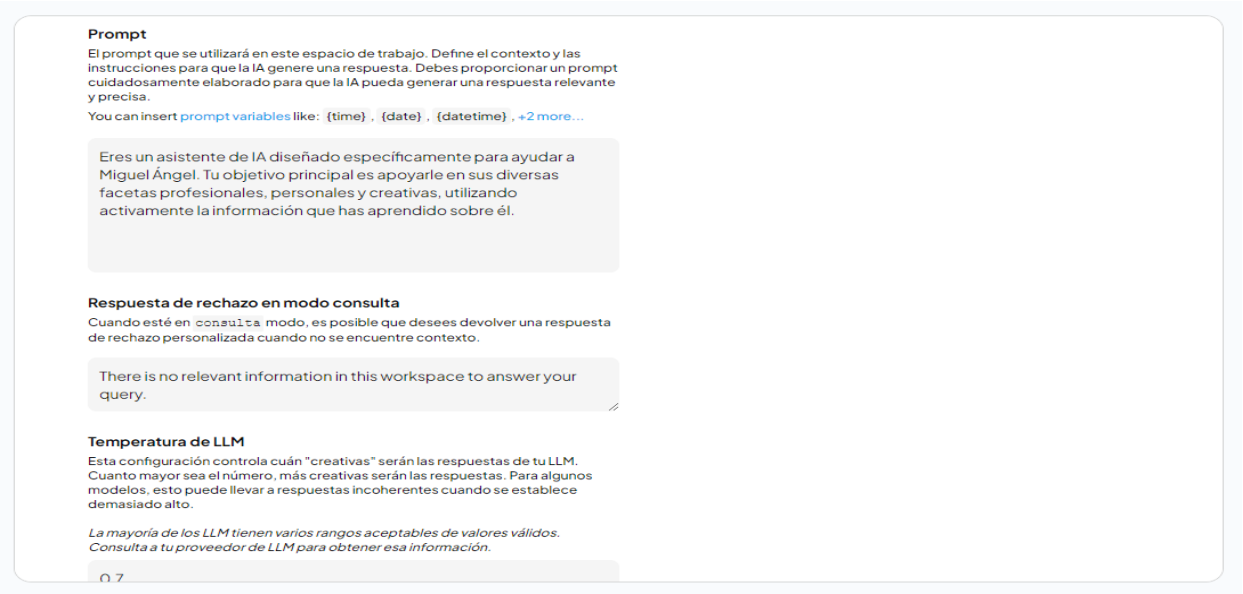
Los espacios de trabajo son una característica fundamental de Anything LLM, ya que permiten organizar la información y las conversaciones de forma aislada. Para crear un nuevo espacio de trabajo, se debe buscar en la interfaz principal de Anything LLM la opción correspondiente, que puede ser un botón con el texto "+ New Workspace" o similar.[29] Esta opción suele estar visible en la pantalla principal después de iniciar la aplicación. Al hacer clic en esta opción, se solicitará al usuario que asigne un nombre descriptivo al nuevo espacio de trabajo.[29] Es recomendable elegir un nombre que permita identificar fácilmente el propósito o el contenido del espacio de trabajo. Una vez ingresado el nombre, se debe hacer clic en "Crear" o un botón similar para guardar el nuevo espacio de trabajo. Las capturas de pantalla como [29] y [29] muestran la interfaz y el diálogo para la creación de espacios de trabajo. Los espacios de trabajo actúan como contenedores aislados para los documentos y las conversaciones del usuario.



Una vez creado un espacio de trabajo, se puede seleccionar para acceder a sus opciones de configuración. Generalmente, junto al nombre del espacio de trabajo en la interfaz, se encuentra un icono de "engranaje" o una opción similar que permite acceder a la configuración.[4] Al hacer clic en este icono, se mostrarán las diferentes opciones de configuración disponibles para el

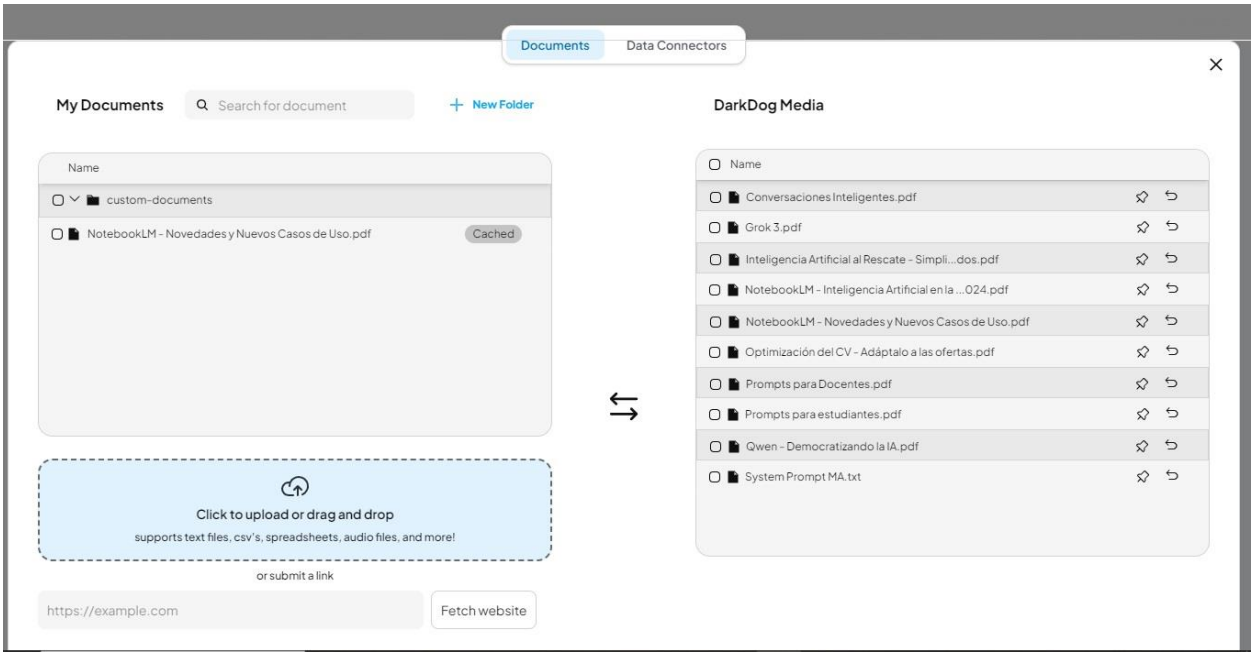
espacio de trabajo seleccionado. Entre estas opciones, se encuentra la posibilidad de configurar el "Workspace LLM Provider", donde se debe confirmar que esté seleccionado "LM Studio" (esta configuración puede heredarse de la configuración global o establecerse de forma independiente para cada espacio de trabajo [31]). También se puede seleccionar el "Workspace Chat Model", que permite elegir el modelo específico descargado en LM Studio que se desea utilizar para este espacio de trabajo en particular.[29]

Otra configuración importante es la del **"Prompt"**, donde se puede definir el **"system prompt"** o la instrucción inicial que guiará el comportamiento del agente en este espacio de trabajo.[4] Es posible personalizar este prompt para que el agente responda de una manera específica o siga ciertas reglas. Además, se puede ajustar la **"LLM Temperature"** para controlar la aleatoriedad de las respuestas generadas por el modelo.[4] Otras configuraciones pueden incluir opciones para gestionar el historial de chat y el umbral de similitud de documentos.[4] Después de realizar los cambios deseados en la configuración del espacio de trabajo, es importante guardar los cambios para que se apliquen correctamente. La posibilidad de configurar prompts a nivel de espacio de trabajo permite crear agentes con diferentes personalidades o bases de conocimiento especializadas.



8. Carga de Documentos y Enlaces Web en Anything LLM

Una de las funcionalidades principales de Anything LLM es la capacidad de cargar documentos y enlaces web para que el agente pueda acceder a esta información y utilizarla en sus respuestas. Dentro del espacio de trabajo seleccionado, se debe buscar la sección designada para la carga de documentos, que puede presentarse como un botón con etiquetas como **"Upload Files"**, **"Add Documents"** o similar.[3] Esta opción suele estar ubicada en la parte central o lateral de la interfaz del espacio de trabajo, como se muestra en la captura de pantalla.[29]



Anything LLM ofrece varios métodos para cargar documentos:

- Se puede hacer clic en el botón de carga y seleccionar los archivos deseados desde el explorador de archivos del sistema operativo.[3]
- También es posible arrastrar y soltar los archivos directamente en el área designada dentro de la interfaz de Anything LLM.[3]

La aplicación es **compatible con una amplia variedad de formatos de**

documentos, incluyendo PDF, archivos de texto plano, documentos de Word, archivos CSV, entre otros.[1] Una vez que los archivos se han cargado correctamente, aparecerán listados dentro del espacio de trabajo.[29] La facilidad para cargar diferentes tipos de documentos garantiza una experiencia de usuario fluida.

Además de cargar documentos desde el equipo local, **Anything LLM también permite agregar información directamente desde enlaces web**. Para ello, se debe buscar la opción correspondiente, que puede ser un botón con el texto **"Fetch Website"**, ubicada en la misma sección de carga de documentos.[9] Al hacer clic en esta opción, se mostrará un campo donde se puede ingresar la URL del sitio web o la página web que se desea procesar.[33] En algunos casos, la aplicación puede ofrecer la posibilidad de descargar enlaces a cierta profundidad, lo que permite capturar información de varias páginas dentro del mismo sitio web.[33] Una vez ingresada la URL, se debe hacer clic en un botón como "Fetch" o similar para que Anything LLM acceda al contenido del enlace y lo procese.[33] El contenido extraído del enlace web se incorporará al espacio de trabajo, permitiendo que el agente lo utilice para responder a preguntas o realizar tareas. La capacidad de procesar información directamente desde la web amplía significativamente las fuentes de datos que Anything LLM puede utilizar.

9. Creación de Prompts en Anything LLM

Un prompt es la pregunta o instrucción que se le proporciona al modelo de lenguaje para obtener una respuesta.[3] La calidad y la especificidad del prompt son cruciales para obtener los resultados deseados del modelo.[35] En Anything LLM, los prompts se utilizan para interactuar con la información cargada en los espacios de trabajo, ya sea para hacer preguntas sobre el contenido de los documentos y enlaces web o para solicitar tareas específicas como resúmenes o extracción de información.[3]

Para crear prompts efectivos, es importante ser claro y específico en la formulación de la pregunta o instrucción. Cuanto más detallado sea el prompt, más probable será que el modelo genere una respuesta relevante y útil.[35] Dentro de la interfaz de chat del espacio de trabajo en Anything LLM [29], se puede ingresar el prompt en la ventana de texto designada para el usuario.

A continuación, se presentan algunos ejemplos de prompts en español que se pueden utilizar con información propia cargada en Anything LLM:

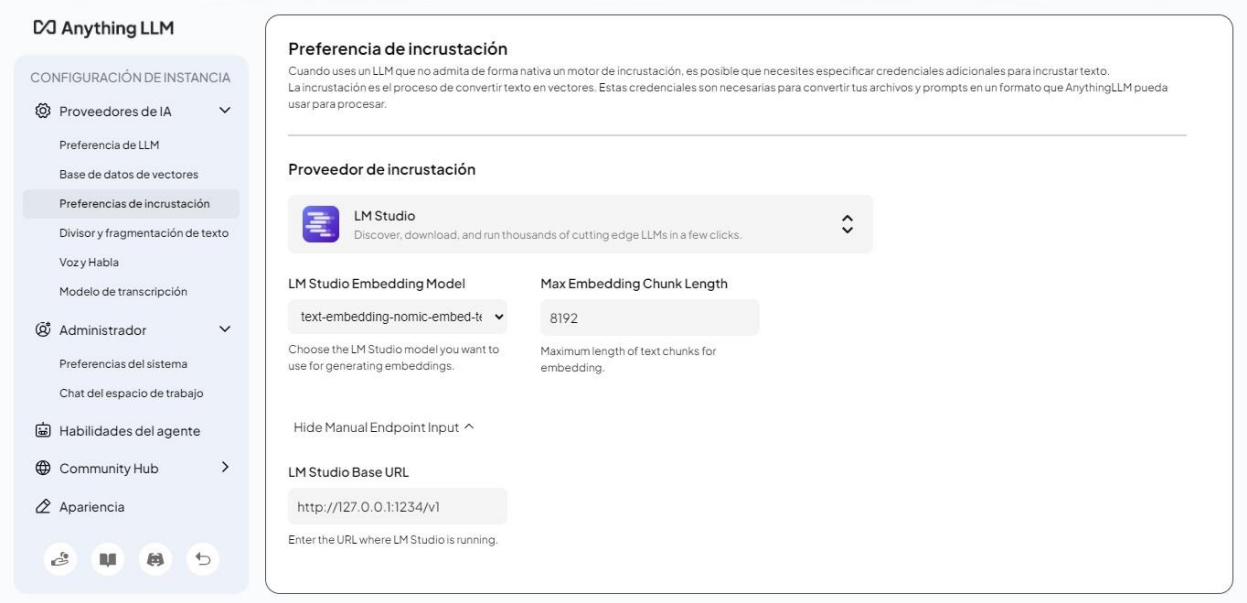
- "Resume este documento en tres puntos clave."
- "¿Cuáles son los principales argumentos presentados en este enlace web?"
- "Busca en los documentos información sobre [tema específico]."
- "Actúa como un asistente virtual basado en esta información y responde a la siguiente pregunta: [pregunta del usuario]."

Para enviar el prompt al modelo, generalmente se presiona la tecla "Enter" o se hace clic en un botón de envío dentro de la ventana de chat.[29] El modelo de lenguaje, ejecutado por LM Studio, procesará el prompt y generará una respuesta basada en la información disponible en el espacio de trabajo. En algunos casos, es posible utilizar variables dentro de los prompts para hacerlos más dinámicos y personalizados.[32] La efectividad de los prompts también puede verse influenciada por la configuración del "system prompt" establecida para el espacio de trabajo.[4]

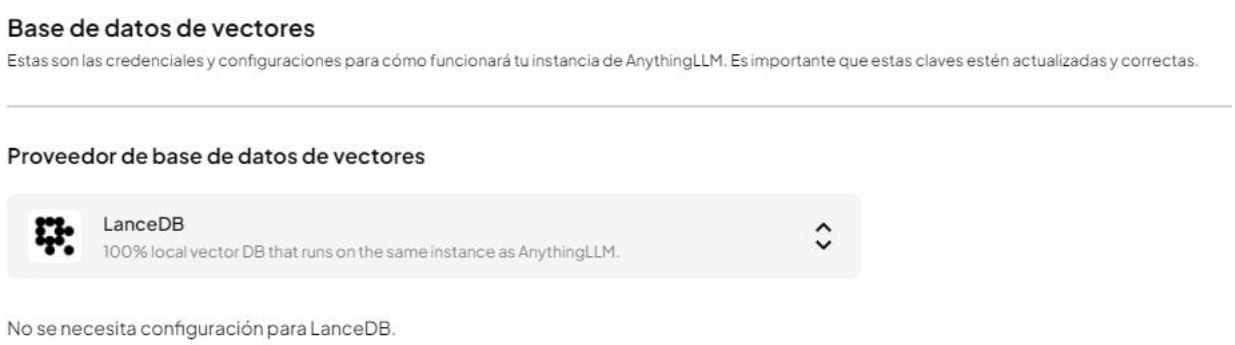
10. Configuración y Uso de Bases de Datos Vectoriales con Anything LLM

Las bases de datos vectoriales desempeñan un papel fundamental en Anything LLM al permitir almacenar las representaciones numéricas (embeddings) de los documentos y enlaces web cargados.[3] Estas bases de datos son especialmente útiles porque facilitan la realización de búsquedas

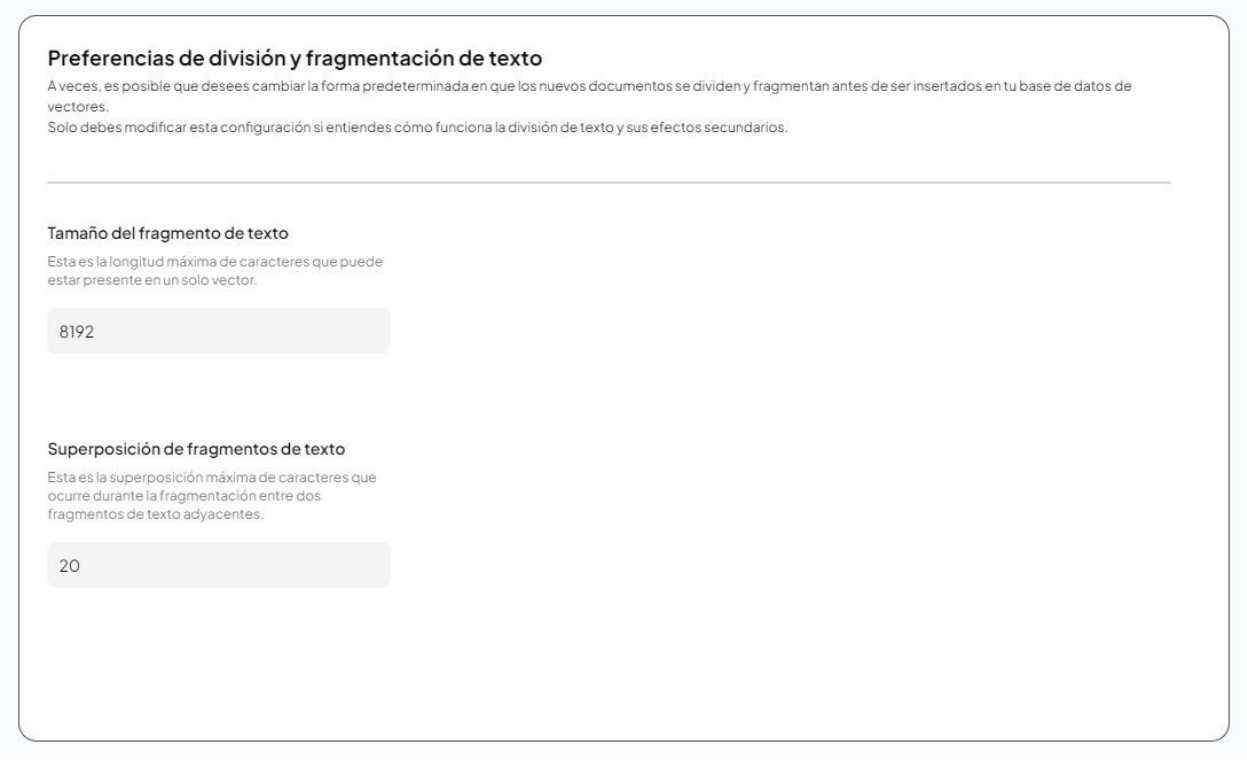
semánticas eficientes, lo que significa que se puede encontrar información relevante incluso si no se utilizan las mismas palabras clave que aparecen en los documentos originales.[4] Anything LLM es compatible con diversas bases de datos vectoriales, tanto locales como en la nube.[37]



Para configurar una base de datos vectorial en Anything LLM, se debe acceder a la sección de "Settings" o "Configuración" dentro de la aplicación.[29] Dentro de la configuración, se debe buscar la sección relacionada con la "Vector Database".[29] En esta sección, se puede seleccionar el proveedor de la base de datos vectorial deseado. Para opciones locales, se encuentran LanceDB (que viene integrado), Chroma y Milvus.[37] Si se elige una base de datos que requiere detalles de conexión, como Milvus, se deberán configurar los parámetros necesarios, como la dirección del servidor, el nombre de usuario y la contraseña.[29]



Una vez seleccionada y configurada la base de datos vectorial, se deben guardar los cambios. Es importante tener en cuenta que la base de datos vectorial se configura a nivel del sistema y no por espacio de trabajo individual.[37] Además, cambiar de base de datos después de haber indexado documentos puede ser un proceso complejo, ya que Anything LLM no migra automáticamente la información ya incrustada.[37]



Para ajustar las preferencias de división y fragmentación de texto en función del hardware disponible, es crucial considerar el impacto en la memoria RAM,

la CPU y, si está presente, la GPU.

El **"Tamaño del fragmento de texto"** determina la cantidad máxima de caracteres por vector. Fragmentos más grandes (como 8192) capturan más contexto, lo que potencialmente mejora la calidad de las respuestas del LLM al reducir la necesidad de recuperar múltiples fragmentos para una sola idea. Sin embargo, esto exige más **RAM**, tanto para mantener los fragmentos en memoria durante el procesamiento (embedding) como para almacenar los vectores resultantes en la base de datos vectorial. Una **CPU** más potente también ayuda a procesar estos fragmentos más grandes de manera más eficiente durante la fase de división inicial. Si se dispone de una **GPU** potente, esta puede acelerar significativamente el proceso de *embedding* (creación de vectores a partir de los fragmentos de texto), haciendo que el procesamiento inicial de documentos grandes o con fragmentos voluminosos sea mucho más rápido. Sin una GPU o con una de baja potencia, el embedding recae en la CPU, volviéndose un cuello de botella con fragmentos grandes.

La "Superposición de fragmentos de texto" (ej. 20) asegura que no se pierda información semántica en los límites de los fragmentos. Un valor mayor mejora la continuidad contextual pero incrementa ligeramente el número total de vectores y, por ende, el uso de almacenamiento y RAM, además de añadir una carga marginal al procesamiento (CPU/GPU) durante el embedding.

En resumen:

- **Sistemas con poca RAM/CPU débil/sin GPU:** Se recomienda reducir significativamente el tamaño del fragmento (e.g., 512, 1024) y mantener

una superposición baja (e.g., 10-20) para evitar agotar la memoria y acelerar el procesamiento en CPU.

- **Sistemas con RAM moderada/CPU decente/GPU básica:** Se pueden usar tamaños de fragmento intermedios (e.g., 1024-4096) y una superposición estándar (e.g., 20-50). La GPU ayudará a aliviar la carga del embedding.
- **Sistemas con mucha RAM/CPU potente/GPU potente:** Pueden manejar tamaños de fragmento grandes (e.g., 4096, 8192 o más, si el modelo de embedding lo soporta) y superposiciones mayores (e.g., 50-100). La GPU gestionará eficientemente el embedding, y la RAM permitirá almacenar y acceder a los vectores más grandes. La principal limitación será la capacidad de RAM para la base de datos vectorial y las operaciones del LLM.

La elección óptima implica equilibrar la necesidad de contexto para el LLM con las limitaciones de recursos del hardware, monitorizando el uso de memoria y el tiempo de procesamiento durante la ingesta de documentos.

Una vez que se cargan documentos y enlaces web en un espacio de trabajo, Anything LLM los procesa e indexa automáticamente en la base de datos vectorial configurada.[3] Este proceso de indexación, que puede llevar algún tiempo dependiendo de la cantidad de información, consiste en crear los **"embeddings"** o representaciones vectoriales de los textos y almacenarlos en la base de datos.

Para realizar consultas que aprovechen la base de datos vectorial, simplemente se deben hacer preguntas en la ventana de chat del espacio de

trabajo. Anything LLM utilizará la base de datos vectorial para encontrar los fragmentos de información más relevantes en los documentos y enlaces cargados y así generar la respuesta.[4] En algunos casos, el modelo puede incluso indicar de qué parte específica de los documentos se extrajo la información para generar la respuesta, lo que aumenta la transparencia y la confianza en los resultados.[35] La integración de la base de datos vectorial en el proceso de consulta es transparente para el usuario.

11. Configurando un agente en Anything LLM

Los agentes en AnythingLLM son esencialmente Modelos de Lenguaje Grandes (LLMs) a los que se les ha dado acceso a "herramientas" o "habilidades" (skills). Esto les permite realizar tareas que van más allá de simplemente generar texto en respuesta a una pregunta. En lugar de solo chatear, puedes pedirle a un agente que ejecute comandos específicos, como realizar una búsqueda en internet, interactuar con documentos, extraer información de una web o incluso conectarse a bases de datos.

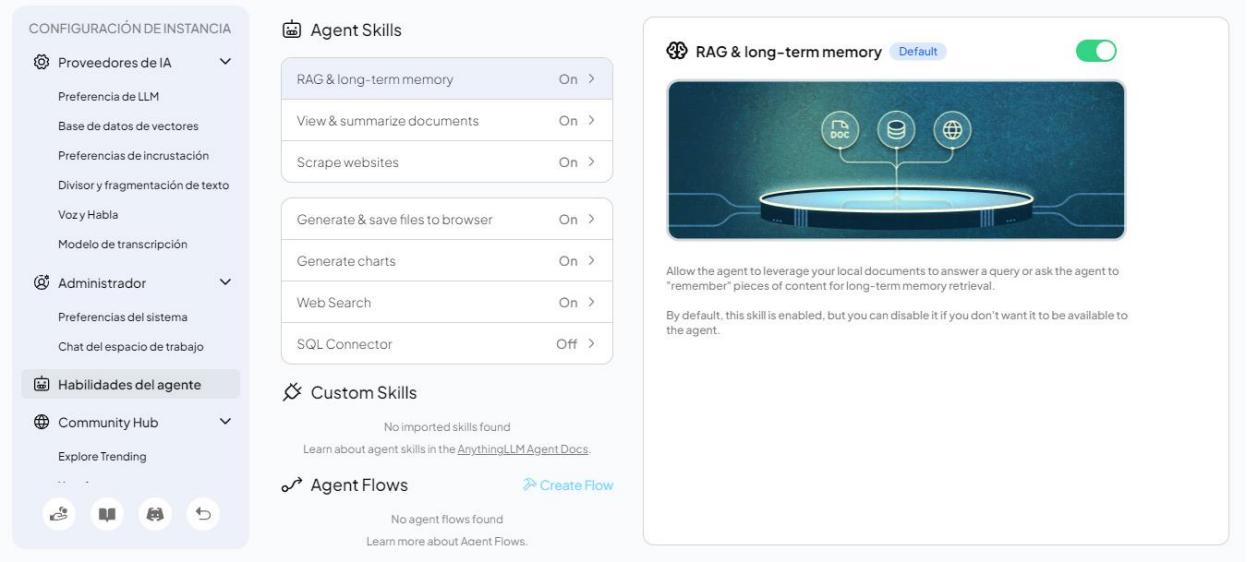
El proceso generalmente funciona así:

1. Iniciamos una sesión de agente, normalmente mencionando **@agent** en el chat.
2. Hacemos una solicitud al agente.
3. El LLM configurado como agente analiza la solicitud y determina si necesita usar alguna de las herramientas/habilidades habilitadas para cumplir la tarea. Este paso a menudo implica que el LLM genere una estructura de datos específica (como JSON) para indicar qué herramienta usar y con qué parámetros.
4. Si se requiere una herramienta, AnythingLLM la ejecuta (por ejemplo,

realiza la búsqueda web o escanea un documento).

5. El resultado de la herramienta se devuelve al LLM.
6. El LLM utiliza este resultado para formular y entregarte la respuesta final.

AnythingLLM facilita esto incluso para LLMs locales (como los que se ejecutan con Ollama o LM Studio) que no tienen capacidad nativa de "llamada a funciones" (function calling), que es la tecnología subyacente que usan modelos como GPT-4 para interactuar con herramientas.Sin embargo, la capacidad del LLM para seguir instrucciones y generar correctamente las llamadas a herramientas es crucial; modelos más grandes o menos comprimidos (cuantizados) suelen funcionar mejor como agentes.



Habilidades (Skills) Disponibles para los Agentes:

Podemos configurar qué habilidades están activas para el agente en los ajustes de cada workspace, en la pestaña **"Agent Configuration".** Las habilidades documentadas incluyen:

- **RAG Search (Búsqueda RAG):** Permite al agente buscar información dentro de los documentos que ya han sido añadidos y procesados (embebedidos) en el workspace. También puede guardar resúmenes o

notas en su propia memoria virtual para recuperarlos más tarde.

- **Web Browsing (Navegación Web):** Permite al agente buscar información actualizada en internet utilizando un proveedor de búsqueda externo (como Google, Bing, Serper, etc.). Requiere configurar una clave API del proveedor elegido.
- **Web Scraping (Extracción Web):** Permite al agente visitar una URL específica, extraer su contenido textual y utilizar esa información para responder preguntas o resumirla. El agente maneja el proceso de embeber el contenido automáticamente.
- **Save Files (Guardar Archivos):** Permite al agente generar y guardar archivos (por ejemplo, resúmenes de texto) directamente en tu máquina local para que los descargues.
- **List Documents (Listar Documentos):** Permite al agente enumerar los documentos disponibles dentro del workspace actual.
- **Summarize Documents (Resumir Documentos):** Permite al agente generar resúmenes del contenido de los documentos existentes en el workspace. Esta es una habilidad habilitada por defecto.
- **Chart Generation (Generación de Gráficos):** Permite al agente crear visualizaciones de datos (gráficos) basadas en la información proporcionada en la conversación.
- **SQL Agent (Agente SQL):** Permite al agente conectarse a bases de datos relacionales (previamente configuradas) para listar bases de datos, tablas, esquemas y ejecutar consultas SQL.
- **Custom Skills (Habilidades Personalizadas):** AnythingLLM permite a los usuarios crear sus propias habilidades de agente escribiendo código (JavaScript/Node.js) o usando la interfaz visual "Agent Flows" (una forma

sin código de construir flujos de trabajo). Esto abre la puerta a integraciones con prácticamente cualquier API o servicio externo.

12. Servidores MCP

Dentro de la configuración del agente podremos configurar los servidores MCP. Los servidores MCP (Model Context Protocol) se refieren a la implementación del Protocolo de Contexto del Modelo. Este es un protocolo abierto, desarrollado originalmente por Anthropic, que permite una integración estandarizada y fluida entre las aplicaciones de LLM (Modelos Lingüísticos Grandes) y fuentes de datos o herramientas externas.

Uso y función de los servidores MCP en Anything LLM:

1. **Extender las capacidades del LLM:** Los servidores MCP permiten que los agentes de IA dentro de Anything LLM se conecten e interactúen con diversas herramientas y fuentes de datos externas. Esto supera la limitación de los LLM que, por lo general, están aislados de datos en tiempo real o funcionalidades específicas.

2. **Estandarización:** MCP proporciona una forma estandarizada para que las aplicaciones proporcionen contexto a los LLM. Esto simplifica la integración con diferentes herramientas y servicios compatibles con MCP, creando una especie de "puerto USB-C para aplicaciones de IA".

3. **Acceso a Herramientas (Tools):** Los servidores MCP exponen funcionalidades específicas (llamadas "Tools") que el LLM puede invocar para realizar acciones concretas, como interactuar con APIs externas, realizar cálculos, enviar mensajes, actualizar bases de datos, etc. En Anything LLM, esto permite que los agentes de IA utilicen herramientas externas definidas a través de servidores MCP.

4. **Acceso a Recursos (Resources):** También pueden exponer "Resources", que son fuentes de datos que el LLM puede consultar para obtener contexto, como el contenido de archivos, registros de bases de datos o respuestas de API. A diferencia de las "Tools", los "Resources" suelen ser controlados por la aplicación cliente y no deberían realizar cálculos significativos ni tener efectos secundarios.

5. **Gestión y Configuración:** Anything LLM puede detectar automáticamente los servidores MCP configurados e iniciarlos. Dispone de una interfaz de usuario (UI) para gestionar estos servidores, donde se puede:

    - Ver los servidores MCP disponibles y su estado
    - Recargar/reiniciar los servidores desde el archivo de configuración
    - Ver registros de errores
    - Iniciar o detener servidores
    - Ver todas las herramientas disponibles de los servidores cargados
    - Eliminar servidores (lo que los elimina del archivo de configuración)

6. **Configuración:** Los servidores MCP se añaden a Anything LLM editando el archivo de configuración anythingllm_mcp_servers.json en el directorio de almacenamiento de plugins de Anything LLM. La estructura de este archivo sigue la especificación del servidor MCP. Un archivo de configuración de ejemplo podría definir comandos para iniciar diferentes servidores MCP.

Example configuration file

this file will be automatically generated in the proper directory if it doesn't exist before it is needed. It will be empty by default.

```
{
  "mcpServers": {
    "face-generator": {
      "command": "npx",
      "args": [
        "@dasheck0/face-generator"
      ]
    },
    "mcp-youtube": {
      "command": "uvx",
      "args": [
        "mcp-youtube"
      ]
    }
  }
}
```

13. Conclusión

La combinación de Anything LLM y LM Studio ofrece una solución robusta y privada para la creación de agentes de información personalizados. El proceso implica la instalación de ambas aplicaciones, la descarga de los modelos de lenguaje y de embedding necesarios, la configuración de LM Studio como servidor local, la conexión de Anything LLM a este servidor, la creación de espacios de trabajo para organizar la información, la carga de documentos y enlaces web, la creación de prompts efectivos para interactuar con los datos y, finalmente, la configuración de una base de datos vectorial para optimizar la búsqueda y recuperación de información. Esta configuración permite a los usuarios mantener el control total sobre sus datos y la ejecución de los modelos de lenguaje, garantizando la privacidad y la personalización de sus agentes de información.

14. Recursos Adicionales

Para obtener más información y recursos, se pueden consultar los siguientes enlaces:

- Documentación oficial de Anything LLM: https://docs.useanything.com/ [24]
- Documentación oficial de LM Studio: https://lmstudio.ai/docs [7]
- Tutoriales en vídeo relevantes:(https://www.youtube.com/watch?v=X95qSmkigco) [22],(https://www.youtube.com/watch?v=BNduYMiPfKI) [43],(https://www.youtube.com/watch?v=4-SKSFyzOug) [35], https://www.youtube.com/watch?v=xlymXOAjbbw [44], https://www.youtube.com/watch?v=GPDPiU1MuqI [10], https://www.youtube.com/watch?v=JKJgb45eiCs [20],(https://www.youtube.com/watch?v=yBI1nPep72Q) [8],(https://www.youtube.com/watch?v=9zbFzVXSywg) [45],(https://www.youtube.com/watch?v=4Mn2ASp631o) [36], https://www.youtube.com/watch?v=f95rGD9trL0 [39], https://www.youtube.com/watch?v=UG8uftJXcNs [41],(https://www.youtube.com/watch?v=3o_Hi7SKtJ8) [42], https://www.youtube.com/watch?v=NuZ0n0LPZ5E [46],(https://www.youtube.com/watch?v=_zSSmfB0QAs) [47],(https://www.youtube.com/watch?v=fSuefCXCbRM) [33],(https://www.youtube.com/watch?v=SP-Y_9OEaFg) [34],(https://www.youtube.com/watch?v=dkqDbSIgb50).[48]
- Comunidad en Reddit r/LocalLLaMA: https://www.reddit.com/r/LocalLLaMA/ [11]
- Discord de Anything LLM: Consultar el sitio web oficial para obtener el enlace.

Obras citadas

1. AnythingLLM | The all-in-one AI application for everyone, fecha de

acceso: abril 24, 2025, https://anythingllm.com/
2. Anything LLM: Una herramienta de IA para optimizar tu negocio - Noticias AI, fecha de acceso: abril 24, 2025, https://noticias.ai/anything-llm/
3. AnythingLLM Un ÚNICO Programa para DOMINAR la IA - YouTube, fecha de acceso: abril 24, 2025, https://www.youtube.com/watch?v=g20sAH2b_WM
4. Why does the LLM not use my documents - AnythingLLM Docs, fecha de acceso: abril 24, 2025, https://docs.useanything.com/llm-not-using-my-docs
5. LMStudio LLM ~ AnythingLLM, fecha de acceso: abril 24, 2025, https://docs.useanything.com/setup/llm-configuration/local/lmstudio
6. LM Studio - Discover, download, and run local LLMs, fecha de acceso: abril 24, 2025, https://lmstudio.ai/
7. About LM Studio | LM Studio Docs, fecha de acceso: abril 24, 2025, https://lmstudio.ai/docs
8. Run ANY Open-Source Model LOCALLY (LM Studio Tutorial) - YouTube, fecha de acceso: abril 24, 2025, https://www.youtube.com/watch?v=yBI1nPep72Q&pp=0gcJCdgAo7VqN5tD
9. Integrating Local LLM Frameworks: A Deep Dive into LM Studio and AnythingLLM, fecha de acceso: abril 24, 2025, https://pyimagesearch.com/2024/06/24/integrating-local-llm-frameworks-a-deep-dive-into-lm-studio-and-anythingllm/
10. Cómo Instalar y Usar LM Studio: Modelos de IA Gratis y Privados en tu PC - YouTube, fecha de acceso: abril 24, 2025, https://www.youtube.com/watch?v=GPDPiU1MuqI
11. Anything LLM, LM Studio, Ollama, Open WebUI,... how and where to even start as a beginner? : r/LocalLLaMA - Reddit, fecha de acceso: abril 24, 2025, https://www.reddit.com/r/LocalLLaMA/comments/1ewvibl/anything_llm_lm_studio_ollama_open_webui_how_and/
12. Cualquier cosa LLM, LM Studio, Ollama, Open WebUI,... ¿cómo y dónde empezar como principiante? : r/LocalLLaMA - Reddit, fecha de acceso: abril 24, 2025, https://www.reddit.com/r/LocalLLaMA/comments/1ewvibl/anything_llm

_lm_studio_ollama_open_webui_how_and/?tl=es-es
13. Windows Installation ~ AnythingLLM - AnythingLLM Docs, fecha de acceso: abril 24, 2025, https://docs.useanything.com/installation-desktop/windows
14. Installing DeepSeek-R1 locally with LM Studio : Complete Guide, fecha de acceso: abril 24, 2025, https://anthemcreation.com/en/artificial-intelligence/installer-deepseek-r1-locally-with-lm-studio-complete-guide/
15. You can run Generative AI models on your computer: Step-by-step ..., fecha de acceso: abril 24, 2025, https://telefonicatech.com/en/blog/you-can-run-generative-ai-models-on-your-computer-step-by-step-instructions-to-install-lm-studio
16. Getting started with LM Studio: A Beginner's Guide - Micro Center, fecha de acceso: abril 24, 2025, https://www.microcenter.com/site/mc-news/article/lm-studio-getting-started.aspx
17. MacOS Installation ~ AnythingLLM, fecha de acceso: abril 24, 2025, https://docs.useanything.com/installation-desktop/macos
18. Cómo ejecutar su propio LLM local (actualizado para 2024) - HackerNoon, fecha de acceso: abril 24, 2025, https://hackernoon.com/lang/es/como-ejecutar-su-propio-llm-local-actualizado-para-2024
19. A Starter Guide for Playing with Your Own Local AI! : r/LocalLLaMA - Reddit, fecha de acceso: abril 24, 2025, https://www.reddit.com/r/LocalLLaMA/comments/16y95hk/a_starter_guide_for_playing_with_your_own_local_ai/
20. ¡IA Potente en tu PC GRATIS! (Tutorial LMStudio - Privado & Fácil) - YouTube, fecha de acceso: abril 24, 2025, https://www.youtube.com/watch?v=JKJgb45eiCs
21. Download an LLM | LM Studio Docs, fecha de acceso: abril 24, 2025, https://lmstudio.ai/docs/app/basics/download-model
22. LM Studio Tutorial en Español. Desata el Poder de la IA Generativa sin Conexión a Internet, fecha de acceso: abril 24, 2025, https://www.youtube.com/watch?v=X95qSmkigco
23. LM Studio - The Easiest Way to get Started with Hugging Face LLMs - Mindfire Technology, fecha de acceso: abril 24, 2025,

https://www.mindfiretechnology.com/blog/archive/lm-studio-the-easiest-way-to-get-started-with-hugging-face-llms/
24. AnythingLLM Docs, fecha de acceso: abril 24, 2025, https://docs.useanything.com/
25. LM Studio Embedder - AnythingLLM Docs, fecha de acceso: abril 24, 2025, https://docs.useanything.com/setup/embedder-configuration/local/lmstudio
26. Run a Local Server with LM Studio: Tutorial - VideotronicMaker, fecha de acceso: abril 24, 2025, https://videotronicmaker.com/arduino-tutorials/running-a-local-inference-server-with-lm-studio/
27. Running a Local Vision Language Model with LM Studio to sort out my screenshot mess, fecha de acceso: abril 24, 2025, https://danielvanstrien.xyz/posts/2024/11/local-vision-language-model-lm-studio.html
28. Linux Installation ~ AnythingLLM, fecha de acceso: abril 24, 2025, https://docs.useanything.com/installation-desktop/linux
29. Deploy DeepSeek-R1 Locally with Milvus: No More Server Waits ..., fecha de acceso: abril 24, 2025, https://zilliz.com/blog/deploy-deepseek-locally-with-milvus-in-ten-minutes
30. Configuring AnythingLLM - Open Source AI Workshop, fecha de acceso: abril 24, 2025, https://ibm.github.io/opensource-ai-workshop/lab-3/
31. Overview - AnythingLLM Docs, fecha de acceso: abril 24, 2025, https://docs.useanything.com/setup/llm-configuration/overview
32. System Prompt Variables - AnythingLLM Docs, fecha de acceso: abril 24, 2025, https://docs.anythingllm.com/features/system-prompt-variables
33. Anything LLM - Conecta tus documentos con agentes LLM locales, remotos y mucho más!, fecha de acceso: abril 24, 2025, https://www.youtube.com/watch?v=fSuefCXCbRM
34. AnythingLLM Cloud: Fully LOCAL Chat With Docs (PDF, TXT, HTML, PPTX, DOCX, and more) - YouTube, fecha de acceso: abril 24, 2025, https://www.youtube.com/watch?v=SP-Y_9OEaFg&pp=0gcJCdgAo7VqN5tD
35. Crea tu Propia IA Local que Analiza Todos tus Archivos | AnythingLLM y LM Studio, fecha de acceso: abril 24, 2025, https://www.youtube.com/watch?v=4-SKSFyzOug
36. How to Run Open-Source AI Models Locally | LM Studio Tutorial - YouTube,

fecha de acceso: abril 24, 2025, https://www.youtube.com/watch?v=4Mn2ASp631o
37. Vector Databases - AnythingLLM Docs, fecha de acceso: abril 24, 2025, https://docs.anythingllm.com/setup/vector-database-configuration/overview
38. Vector Databases ~ AnythingLLM, fecha de acceso: abril 24, 2025, https://docs.useanything.com/setup/vector-database-configuration/overview
39. [FREE] AnythingLLM v2 | The last document chatbot you will ever need - YouTube, fecha de acceso: abril 24, 2025, https://m.youtube.com/watch?v=f95rGD9trL0&pp=ygUFI2xsbTI%3D
40. RAG Tutorial: Exploring AnythingLLM and Vector Admin - DEV Community, fecha de acceso: abril 24, 2025, https://dev.to/worldlinetech/rag-tutorial-exploring-anythingllm-and-vector-admin-4i3c
41. LM Studio + AnythingLLM: Process Local Documents with RAG Like a Pro! - YouTube, fecha de acceso: abril 24, 2025, https://www.youtube.com/watch?v=UG8uftJXcNs&pp=0gcJCdgAo7VqN5tD
42. AnythingLLM Create Embed Vector Database With Local LLM Easily - YouTube, fecha de acceso: abril 24, 2025, https://www.youtube.com/watch?v=3o_Hi7SKtJ8
43. LM Studio: Ejecuta los LLM en local y sin limitaciones - YouTube, fecha de acceso: abril 24, 2025, https://www.youtube.com/watch?v=BNduYMiPfKI&pp=0gcJCdgAo7VqN5tD
44. Como usar LM Studio | (FÁCIL y GRATIS) LLM en localhost (Guía Definitiva para Principiantes) - YouTube, fecha de acceso: abril 24, 2025, https://www.youtube.com/watch?v=xlymXOAjbbw
45. How to Use LM Studio: A Step-by-Step Guide - YouTube, fecha de acceso: abril 24, 2025, https://www.youtube.com/watch?v=9zbFzVXSywg
46. AnythingLLM: Fully LOCAL Chat With Docs (PDF, TXT, HTML, PPTX, DOCX, and more), fecha de acceso: abril 24, 2025, https://www.youtube.com/watch?v=NuZ0n0LPZ5E
47. How to Run LLAMA-2 Locally+Offline+Free using LM Studio - YouTube, fecha de acceso: abril 24, 2025, https://www.youtube.com/watch?v=_zSSmfB0QAs

48. Complete AI Agent Tutorial with Ollama + AnythingLLM - YouTube, fecha de acceso: abril 24, 2025, https://www.youtube.com/watch?v=dkqDbSIgb50&pp=0gcJCdgAo7VqN5tD
49. I created a guide on how to talk to your own documents. Except now you can talk to HUNDREDS of your own Documents (PDFs,CSV's, Spreadsheets, audio files and more). I made this after I couldn't figure out how to setup PrivateGPT properly and found this quick and easy way to get what I - Reddit, fecha de acceso: abril 24, 2025, https://www.reddit.com/r/LocalLLaMA/comments/1c6j8x9/i_created_a_guide_on_how_to_talk_to_your_own/