Get Plugged In: How to Use Generative AI Tools in Obsidian

Two community-developed plug-ins highlight how the next generation of app developers can embrace large language models on RTX AI PCs.

November 6, 2024 by Jesse Clayton

Share

Reading Time: 4 mins

Editor's note: This post is part of the AI Decoded series, which demystifies AI by making the technology more accessible, and showcases new hardware, software, tools and accelerations for GeForce RTX PC and NVIDIA RTX workstation users.

As generative AI evolves and accelerates industry, a community of AI enthusiasts is experimenting with ways to integrate the powerful technology into common productivity workflows.

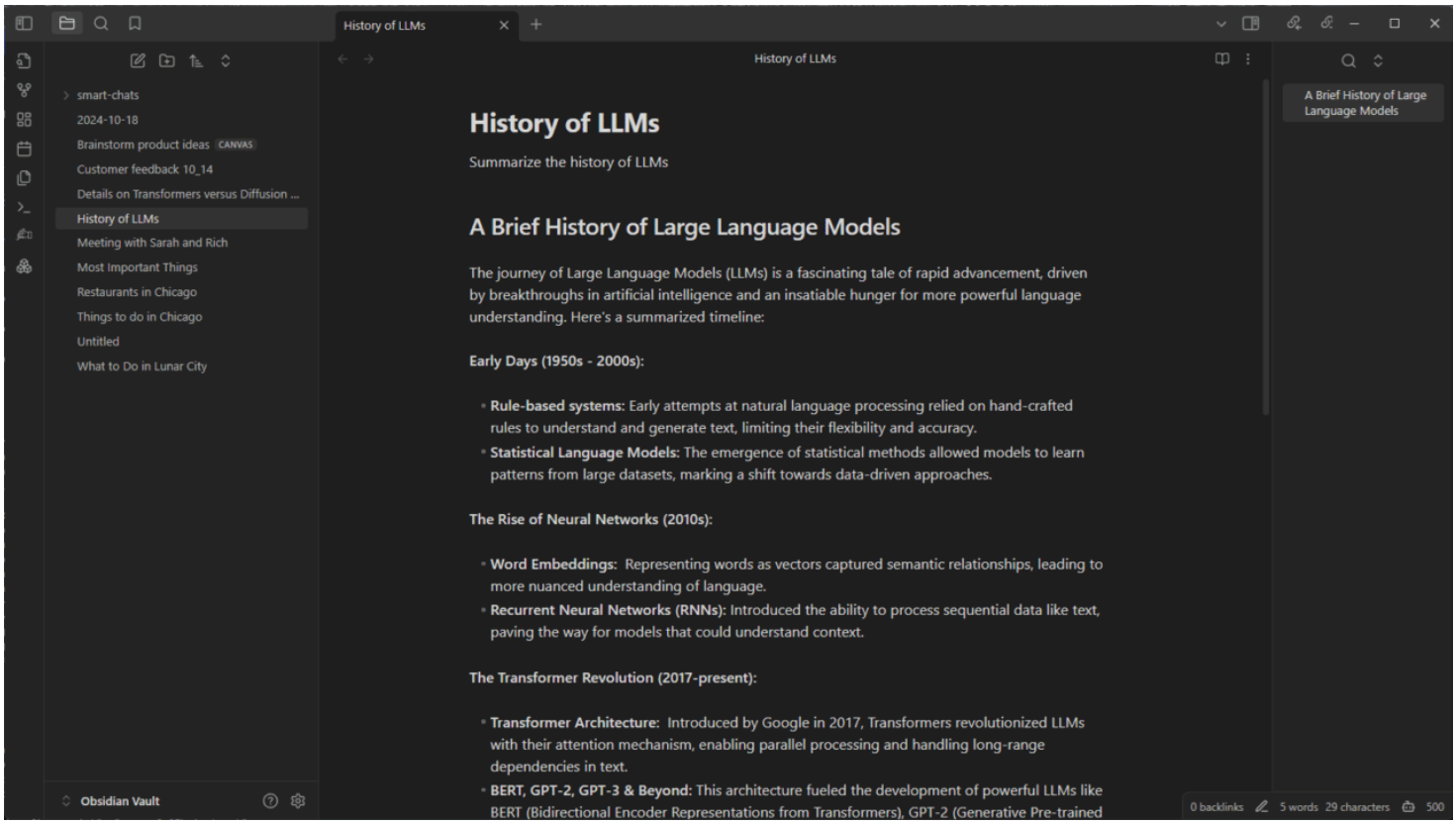
Applications that support community plug-ins give users the power to explore how large language models (LLMs) can enhance a variety of workflows. By using local inference servers powered by the

NVIDIA and our partners use cookies and other tools to collect information you provide as well as your interaction with our websites for performance improvement, analytics, and to assist in our marketing efforts. We also share this information with our social media, advertising, and analytics partners. You can manage your cookie settings by clicking on "Manage Settings". Please see our Cookie Policy for more information.

Manage Settings Agree

NVIDIA RTX-accelerated llama.cpp software library, users on RTX AI PCs can integrate local LLMs with ease.

Previously, we looked at how users can take advantage of Leo AI in the Brave web browser to optimize the web browsing experience. Today, we look at Obsidian, a popular writing and note-taking application, based on the Markdown markup language, that's useful for keeping complex and linked records for multiple projects. The app supports community-developed plug-ins that bring additional functionality, including several that enable users to connect Obsidian to a local inferencing server like Ollama or LM Studio.



Using Obsidian and LM Studio to generate notes with a 27B-parameter LLM accelerated by RTX.

Connecting Obsidian to LM Studio only requires enabling the local server functionality in LM Studio by clicking on the "Developer" icon on the left panel, loading any downloaded model, enabling the CORS toggle and clicking "Start." Take note of the chat completion URL from the "Developer" log console ("http://localhost:1234/v1/chat/completions" by default), as the plug-ins will need this information to connect.

Next, launch Obsidian and open the "Settings" panel. Click "Community plug-ins" and then "Browse." There are several community plug-ins related to LLMs, but two popular options are Text Generator and Smart Connections.

- Text Generator is helpful for generating content in an Obsidian vault, like notes and summaries on a research topic.

NVIDIA and our partners use cookies and other tools to collect information you provide as well as your interaction with our websites for performance improvement, analytics, and to assist in our marketing efforts. We also share this information with our social media, advertising, and analytics partners. You can manage your cookie settings by clicking on "Manage Settings". Please see our Cookie Policy for more information.

- Smart Connections is useful for asking questions about the contents of an Obsidian vault, such as the answer to an obscure trivia question previously saved years ago.

Each plug-in has its own way of entering the LM Server URL.

For Text Generator, open the settings and select “Custom” for “Provider profile” and paste the whole URL into the “Endpoint” field. For Smart Connections, configure the settings after starting the plug-in. In the settings panel on the right side of the interface, select “Custom Local (OpenAI Format)” for the model platform. Then, enter the URL and the model name (e.g., “gemma-2-27b-instruct”) into their respective fields as they appear in LM Studio.

Once the fields are filled in, the plug-ins will function. The LM Studio user interface will also show logged activity if users are curious about what’s happening on the local server side.

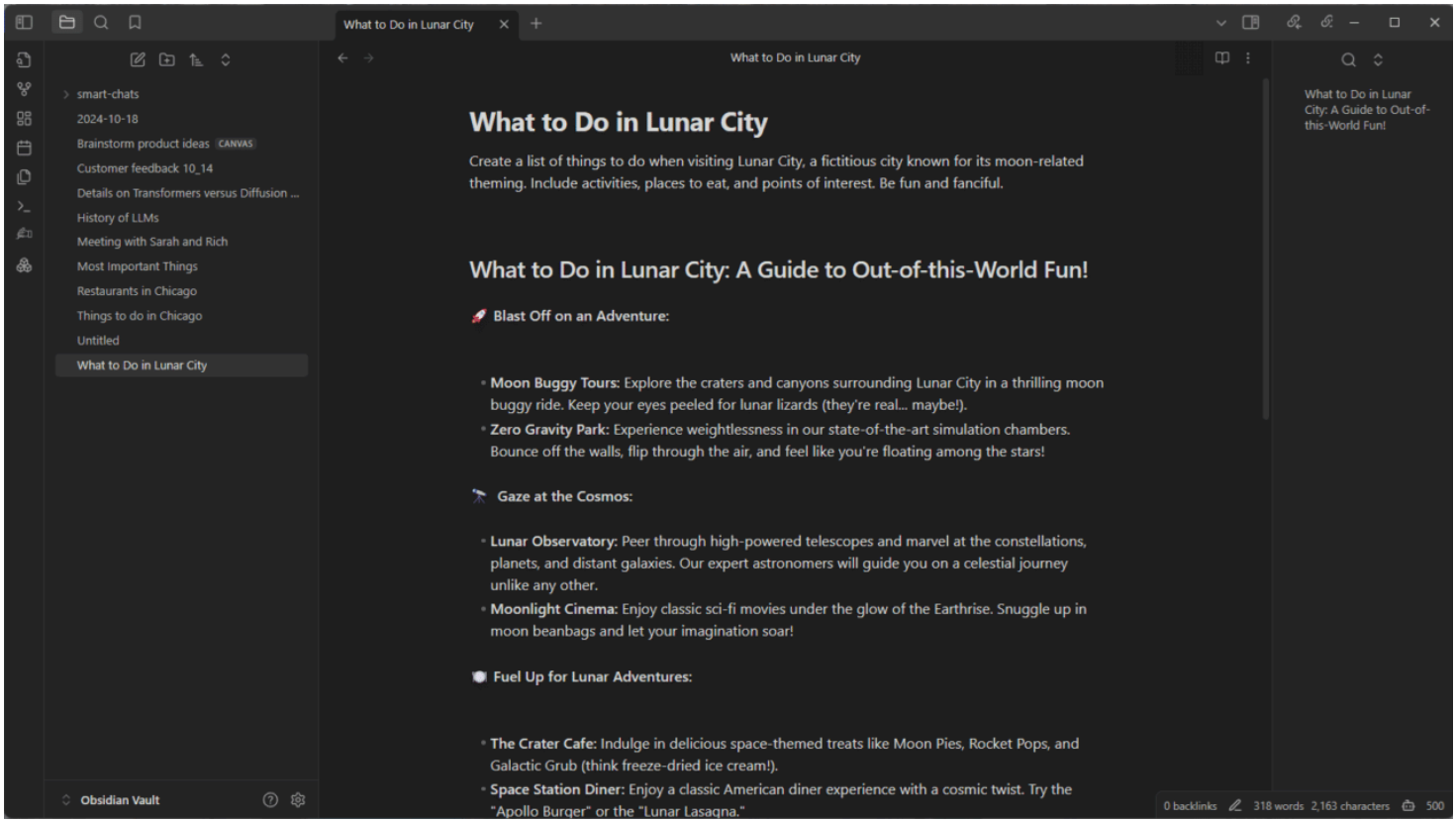
Transforming Workflows With Obsidian AI Plug-Ins

Both the Text Generator and Smart Connections plug-ins use generative AI in compelling ways.

For example, imagine a user wants to plan a vacation to the fictitious destination of Lunar City and brainstorm ideas for what to do there. The user would start a new note, titled “What to Do in Lunar City.” Since Lunar City is not a real place, the query sent to the LLM will need to include a few extra instructions to guide the responses. Click the Text Generator plug-in icon, and the model will generate a list of activities to do during the trip.

Obsidian, via the Text Generator plug-in, will request LM Studio to generate a response, and in turn LM Studio will run the Gemma 2 27B model. With RTX GPU acceleration in the user’s computer, the model can quickly generate a list of things to do.

NVIDIA and our partners use cookies and other tools to collect information you provide as well as your interaction with our websites for performance improvement, analytics, and to assist in our marketing efforts. We also share this information with our social media, advertising, and analytics partners. You can manage your cookie settings by clicking on "Manage Settings". Please see our Cookie Policy for more information.



The Text Generator community plug-in in Obsidian enables users to connect to an LLM in LM Studio and generate notes for an imaginary vacation. The Text Generator community plug-in in Obsidian allows users to access an LLM through LM Studio to generate notes for a fictional vacation.

Or, suppose many years later the user's friend is going to Lunar City and wants to know where to eat. The user may not remember the names of the places where they ate, but they can check the notes in their vault (Obsidian's term for a collection of notes) in case they'd written something down.

Rather than looking through all of the notes manually, a user can use the Smart Connections plug-in to ask questions about their vault of notes and other content. The plug-in uses the same LM Studio server to respond to the request, and provides relevant information it finds from the user's notes to assist the process. The plug-in does this using a technique called retrieval-augmented generation.

NVIDIA and our partners use cookies and other tools to collect information you provide as well as your interaction with our websites for performance improvement, analytics, and to assist in our marketing efforts. We also share this information with our social media, advertising, and analytics partners. You can manage your cookie settings by clicking on "Manage Settings". Please see our Cookie Policy for more information.

The Smart Connections community plug-in in Obsidian uses retrieval-augmented generation and a connection to LM Studio to enable users to query their notes.

These are fun examples, but after spending some time with these capabilities, users can see the real benefits and improvements for everyday productivity. Obsidian plug-ins are just two ways in which community developers and AI enthusiasts are embracing AI to supercharge their PC experiences. .

NVIDIA GeForce RTX technology for Windows PCs can run thousands of open-source models for developers to integrate into their Windows apps.

Learn more about the power of LLMs, Text Generation and Smart Connections by integrating Obsidian into your workflow and play with the accelerated experience available on RTX AI PCs

Generative AI is transforming gaming, videoconferencing and interactive experiences of all kinds. Make sense of what's new and what's next by subscribing to the AI Decoded newsletter.

NVIDIA and our partners use cookies and other tools to collect information you provide as well as your interaction with our websites for performance improvement, analytics, and to assist in our marketing efforts. We also share this information with our social media, advertising, and analytics partners. You can manage your cookie settings by clicking on "Manage Settings". Please see our Cookie Policy for more information.

Categories: Generative AI

Tags: AI Decoded | Artificial Intelligence | GeForce | NVIDIA RTX

All NVIDIA News

NVIDIA and our partners use cookies and other tools to collect information you provide as well as your interaction with our websites for performance improvement, analytics, and to assist in our marketing efforts. We also share this information with our social media, advertising, and analytics partners. You can manage your cookie settings by clicking on "Manage Settings". Please see our Cookie Policy for more information.

Austin Calling: As Texas Absorbs Influx of Residents, Rekor Taps NVIDIA Technology for Roadway Safety, Traffic Relief

Give AI Eyes to Any Industry Can Now Search and Summarize Vast Volumes of Visual Data

The NVIDIA Engineer Who Mastered a BA, MBA, PhD and QA

Startup Helps Surgeons Target Breast Cancers With AI-Powered 3D Visualizations

Scale New Heights With ‘Dragon Age: The Veilguard’ in the Cloud on GeForce NOW

NVIDIA and our partners use cookies and other tools to collect information you provide as well as your interaction with our websites for performance improvement, analytics, and to assist in our marketing efforts. We also share this information with our social media, advertising, and analytics partners. You can manage your cookie settings by clicking on "Manage Settings". Please see our Cookie Policy for more information.