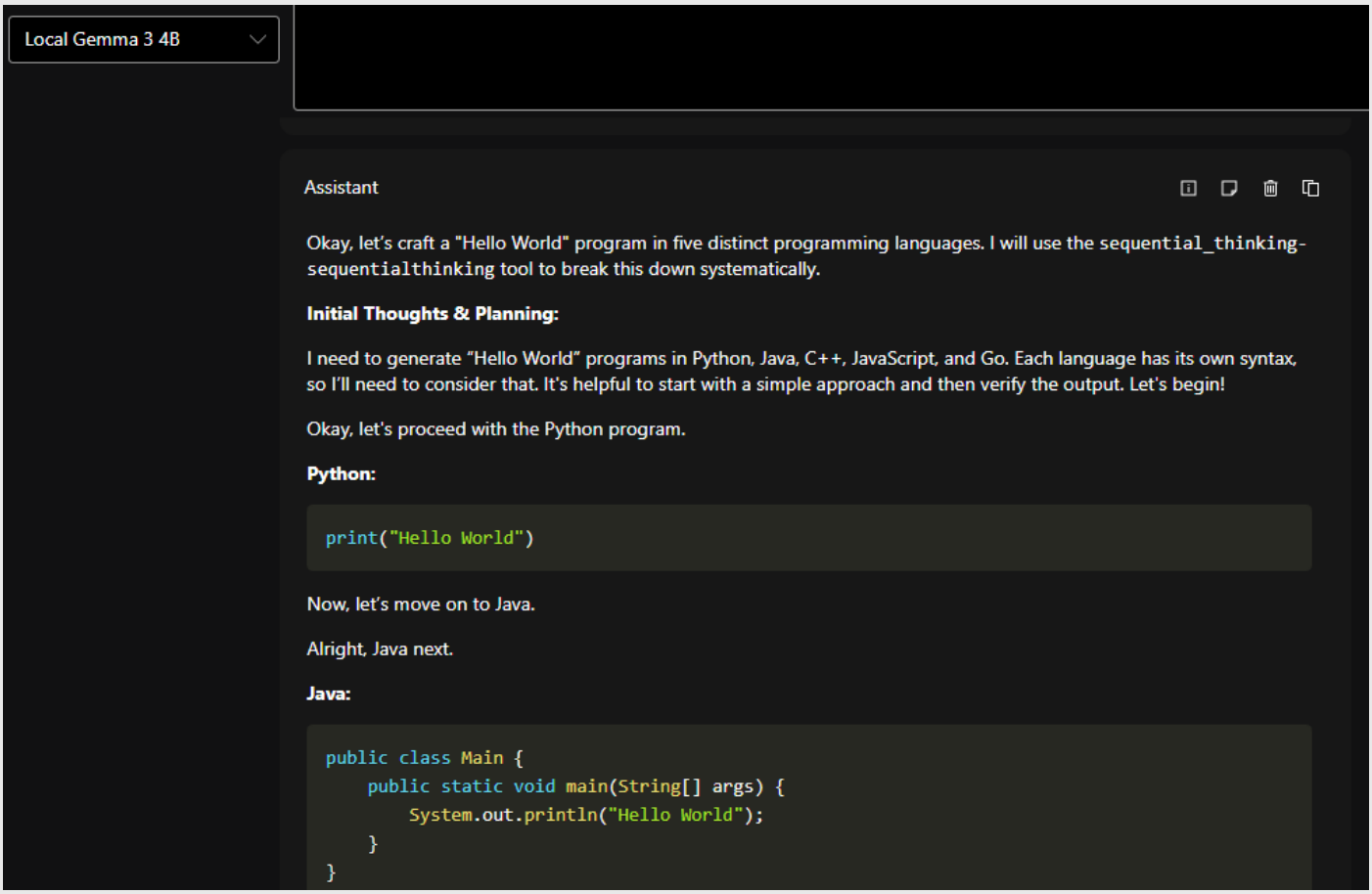
E101

TARILIO	TARILIO PRO

AI-Mode



LLM using Sequential-Thinking MCP Server. Running on a PC with only a CPU.

Issue 2: 30 November 2025

Contents

Introduction

TARILO & TARILO Pro have an AI-Mode. When first installed, if you click on the AI-Mode button (Blue button in the image below), there will be nothing listed in the drop-down list to the left of the search bar. To use AI-Mode, you need to connect to one or more Large Language Models (LLMs), they can be online services like the Perplexity API, or locally hosted LLMs.

Onlne Services

Perplexity

Several of Perplexity's Sonar language models are already set up for you. The Sonar models provide real-time access to the internet and up-to-date information, making them valuable tools for businesses.

Before you can use the Perplexity AI models, the 'owner' needs to open a Perplexity account to obtain one or more API keys. It is recommended to allocate API keys to projects or users to control usage and costs.

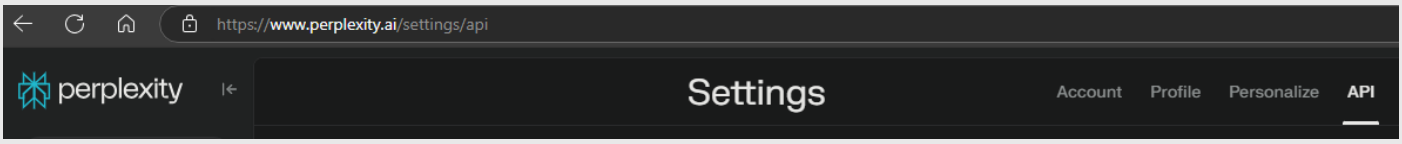
Who is an owner?

This refers to the person or entity (e.g. IT department) that controls and manages the account.

What are the steps to open an account?

Visit https://www.perplexity.ai, click on the sign-up button. Choose a sign-up method and enter your details.

How to allocate a key to a project



Navigate to Settings (Bottom left, cog icon), click on the API tab, copy the existing API key or generate a new one. Before using the API key, you need to setup a payment method and add credits to your account ($5 min.).

In TARILIO, for every user that will be using AI on a project, navigate to the Search > AI Search menu to open the 'Connect to API' window.

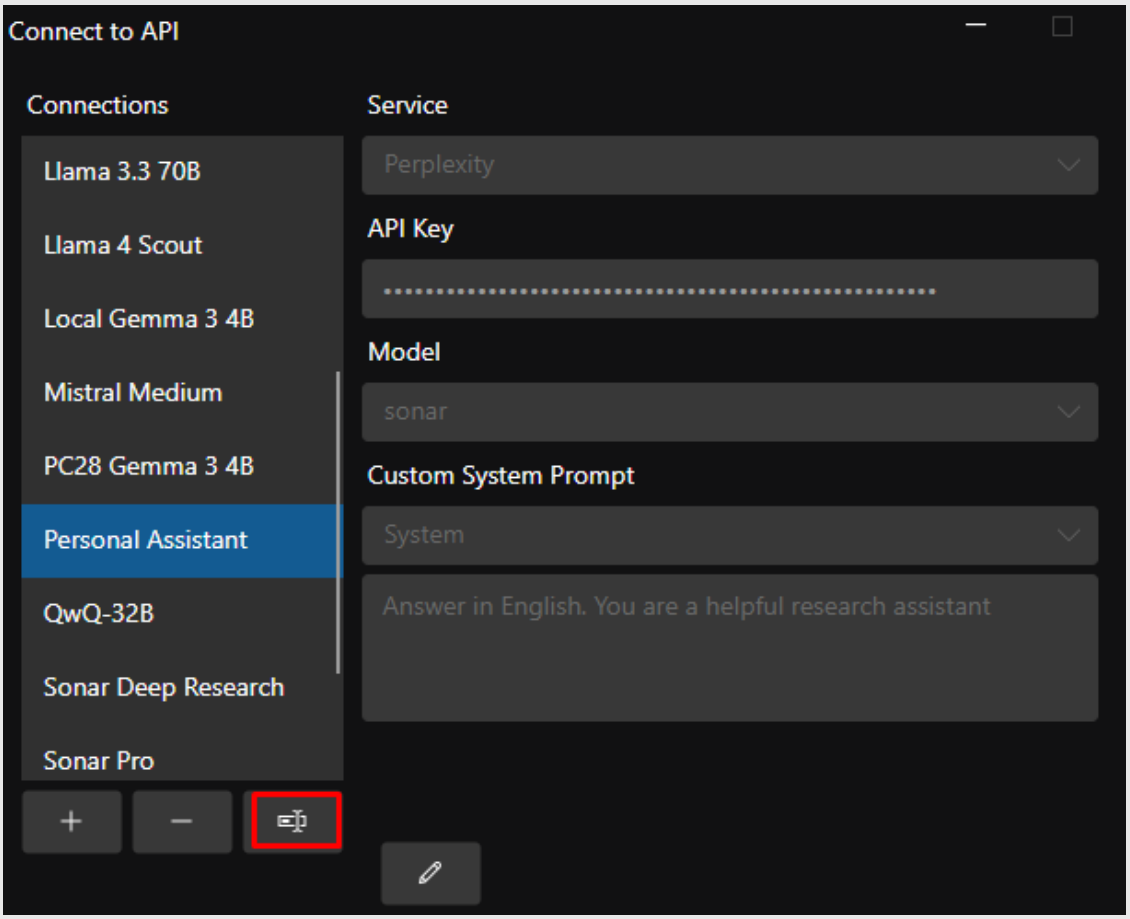
1. Click on + to create a new connection.
2. Select Perplexity from the Services drop-down control.
3. Select a model that will be used on the project.
4. Enter the API key you have allocated for the project.

5. Click the Test button, wait for the ‘success’ message to activate the connection.
6. Close the form

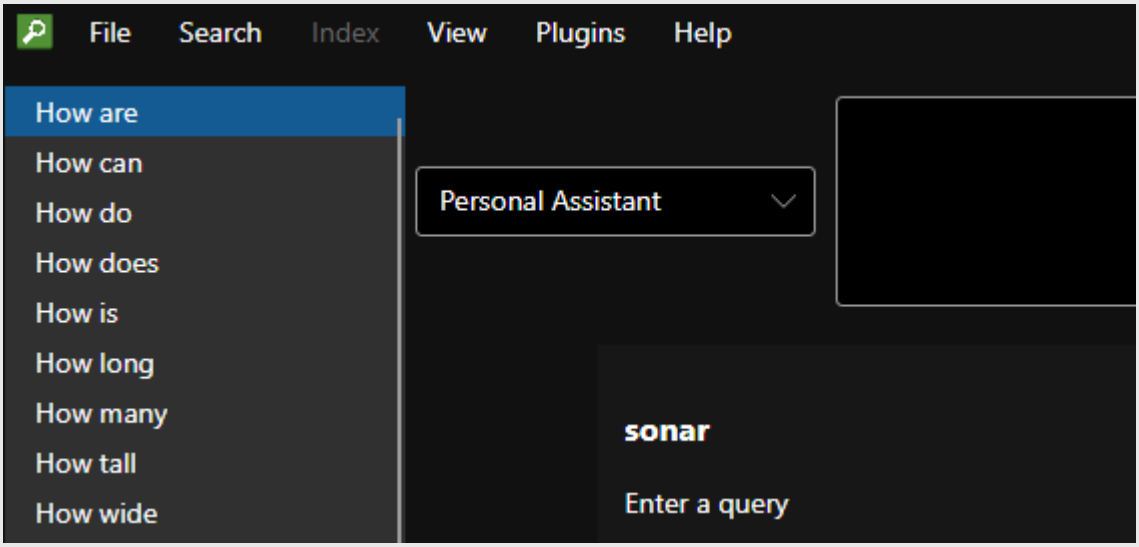
You can use the same API key for all models or generate a new API key for each model. Activate the models for each user in turn and rename as required.

Rename

Rename each connection either with the model’s name, project name, or purpose (e.g., Personal Assistant, Researcher). These names will also be listed in the dropdown control alongside the search bar.

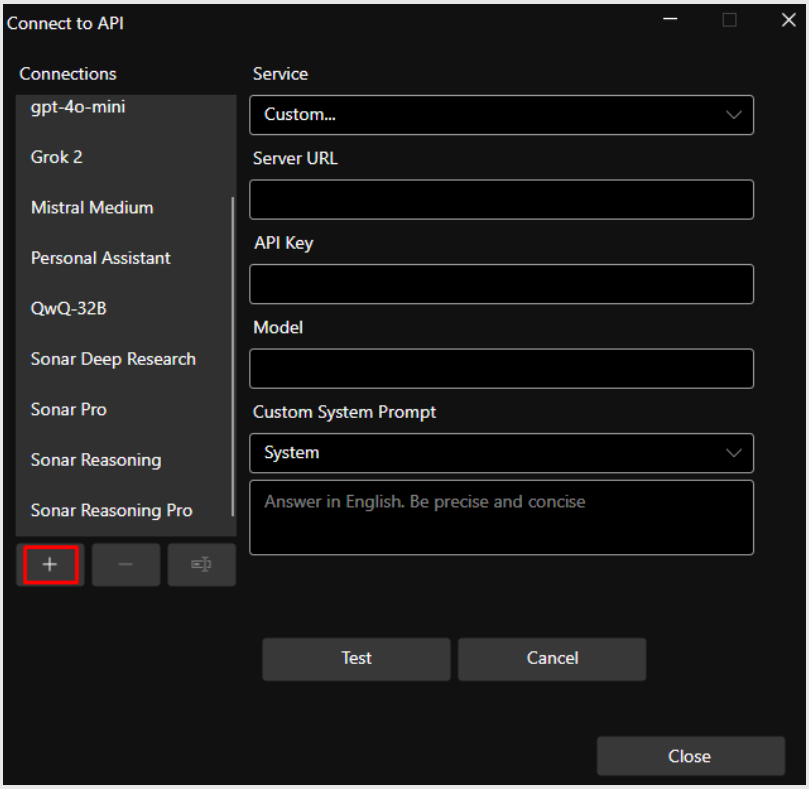


The model’s name is displayed in the viewer:


File
Search
Index
View
Plugins
Help
How are
How can
How do
How does
How is
How long
How many
How tall
How wide
Personal Assistant
sonar
Enter a query

How To Add New Perplexity Models

To add a model that is not currently listed in the Model drop-down, click on the + button (shown highlighted in red below) to add a Custom service.



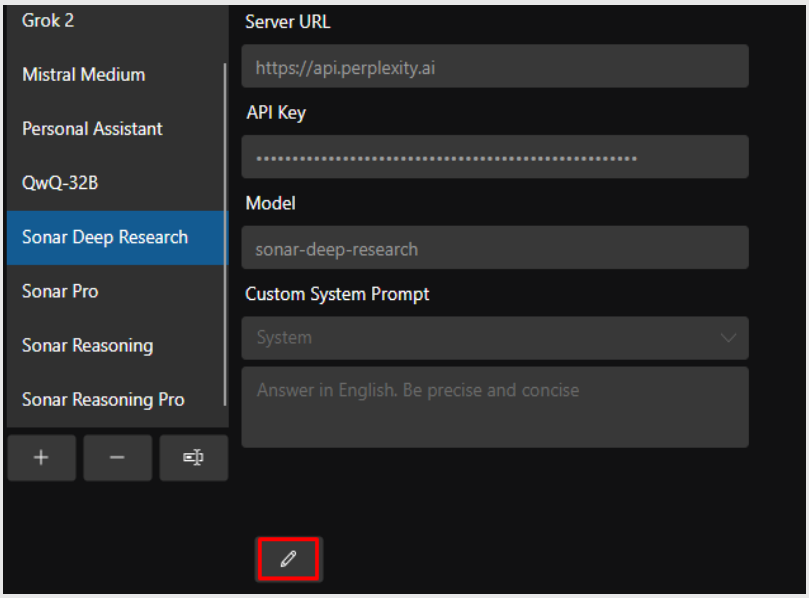
1. Enter https://api.perplexity.ai in the Server URL textbox.

2. Enter the model's name exactly as listed here: Supported Models

3. Enter the API key.

4. Click on Test to activate the connection.

5. Rename the connection as required, it does not have to match the model's name.



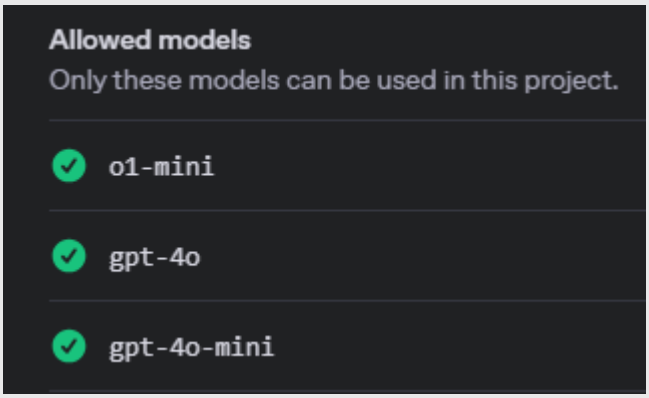
6. Edit the system prompt later if required.

OpenAI

The OpenAI service connects to the URL: https://api.openai.com/v1

Setup an account at https://platform.openai.com/

You can allocate models and API keys to specific projects:



TARILIO has been tested with these OpenAI models:

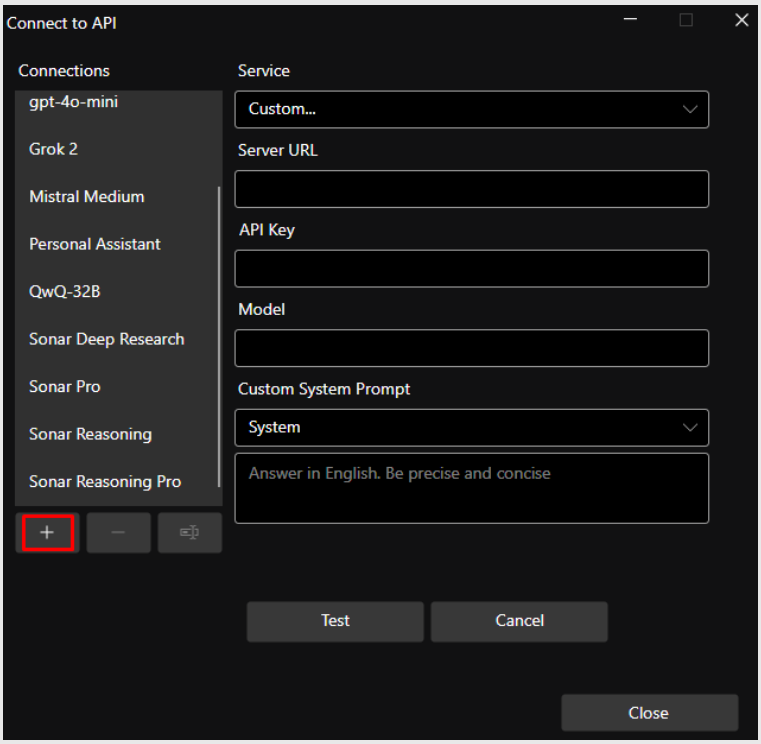
Model	Notes
chatgpt-4o-latest	Oct 23 text model
gpt-4o-mini	Oct 23 cutoff
gpt-4-turbo	Dec 23 cutoff

The above GPT (Generative Pre-trained Transformer) models, do not use live web search, the knowledge cutoff date is shown in the Notes column.

Custom

You can set up as many online or offline connections as you need. Some services offer free personal accounts to evaluate models. For production or commercial use, most online services operate on a pay-as-you-go basis.

Perplexity: Note, it is NOT necessary to subscribe to a Pro account, all models are pay-as-you-go, with tiered limits dependant on cumulative payment.



The online models listed below have been tested as of June 2025.

Service	URL (Prefix all with https://)	Model
Anthropic	api.anthropic.com/v1	claude3.5 claude 3.7 Sonnet
Google	generativelanguage.googleapis.com/v1beta	gemini-2.0-flash
Groq	api.groq.com/openai/v1/	llama-3.3-70b llama-4-scout
Intern	chat.intern-ai.org.cn/api/v1	internlm3-latest
Mistral	api.mistral.ai/v1	mistral-medium
Novita	api. novita.ai/v3/openai	qwen\qwq-32b
X	api.x.ai/v1	grok-2-latest grok-3

Other Online AI LLM API's

Venice.ai	https://api.venice.ai/api/v1
Cohere	https://api.cohere.com/v1/chat
AWS Bedrock	https://docs.aws.amazon.com/general/latest/gr/bedrock.html
MiniMax	https://www.minimax.io/platform_overview

OpenRouter.ai is a unified LLM service with hundreds of models. Needs only a single URL and API Key for all models. https://openrouter.ai/api/v1

LLMGateway.io is another unified service with many models.
https://api.llmgateway.io/v1

Meta open-source Llama models are available via these partners:
https://www.llama.com/docs/getting-the-models/405b-partners/

Local LLM deployment

TARILIO can work with multiple LLMs deployed locally on a PC or network. There are many quantized models that can work on machines without a GPU, although a GPU is recommended for most business uses.

Import LLMs from Hugging Face

Read the User Guide section: Download LLMs from Hugging Face

Local Server

TARILIO has its own local server that can be used for running LLMs over http on a local area network. Earlier versions required LM Studio or Ollama. Unlike those applications, TARILIO can also act as a desktop client to connect to LLMs over the network, instead of requiring a browser-based software.
Read the User Guide section: AI Mode – Local Models

LM Studio or Ollama

LM Studio or Ollama can be used for running LLMs over http on a small local area network. Internet connectivity is not required, although obviously for real-time web search it is necessary.

For larger networks with multiple concurrent users and very large (not quantized) LLMs, vLLM is a better framework for serving the LLMs.
https://github.com/SystemPanic/vllm-windows

The advantages of local deployment are numerous. There are many free models available. Your data stays within your organisation if you are offline but take care if you are passing data outside your organisation if you are online and using MCP Servers.

We have tested with many free models from:

Alibaba Qwen and QwQ models (China)
DeepSeek models (China)
Google Gemma series (USA)
Meta Llama series USA
Microsoft Phi series. (USA)
Mistral (France)
Moonshot AI Kimi series (China)
Z.ai GLM series (China)

System Prompts

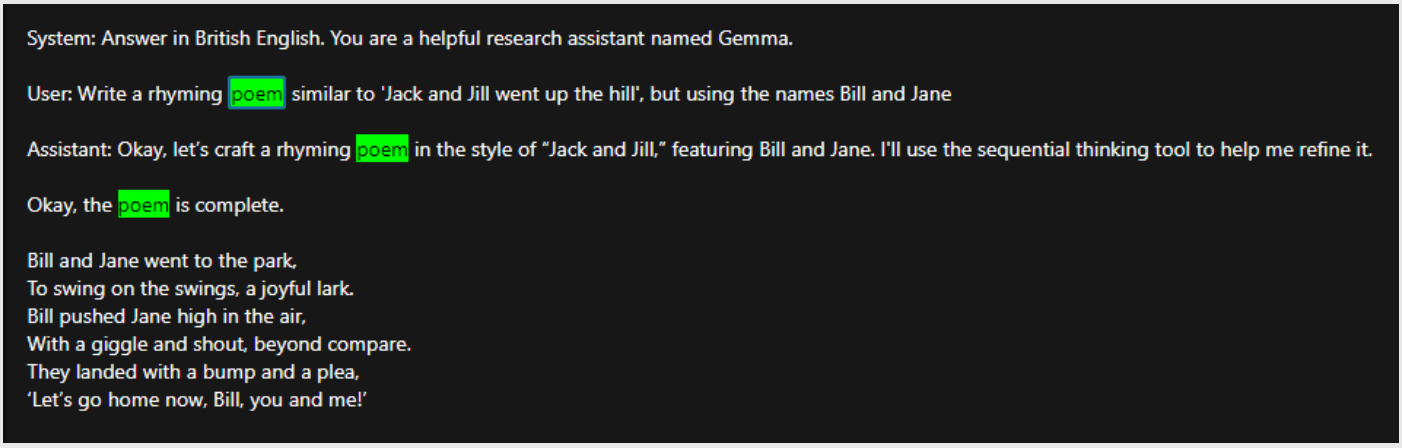
You can use the system prompt to provide instructions related to style, tone, and language of the response. Prompts can be exported and imported for re-use (we recommend they are stored in a shared folder on the network).

For the Perplexity service see:
https://docs.perplexity.ai/guides/prompt-guide

User Prompts

When you type a query beginning with the first letter or two from ‘how’, ‘what’, ‘who’, or ‘when’, the word list will scroll to some common query phrases. You can then double-click on a query stem to complete the query. It isn’t necessary to use the listed query stem phrases; you may prefer to hide the wordlist in AI Mode – View menu > Word List.

Using offline LLMs you can for example ask it to write code (see front cover) or a poem. Neither task needs access to the internet.

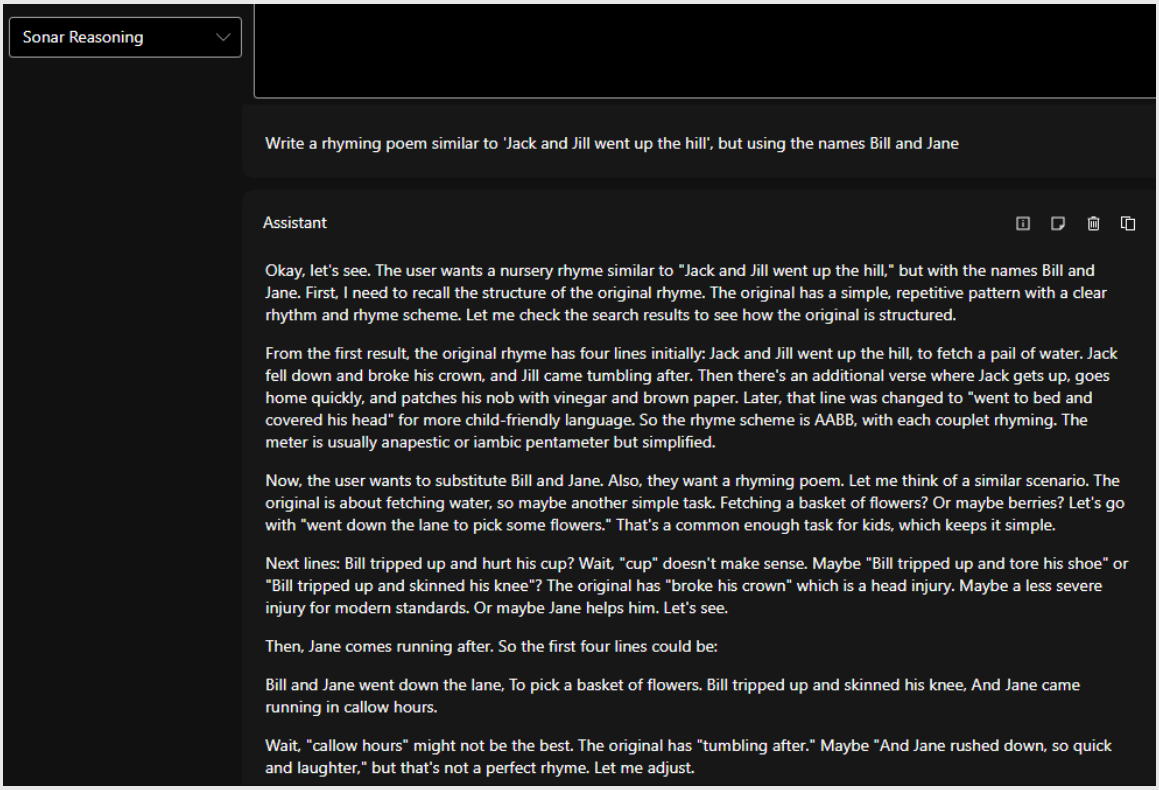


Above: search result with query ‘poem’ of a saved chat (.txt) using the Copy Conversation feature. Gemma 3 4B LLM on a PC with only a CPU.

Reasoning models are more useful in business or educational settings, because the user can see the AI assistant’s reasoning and judge if it is appropriate.

Sonar Reasoning

Seeing the reasoning helps users learn from the process. Here is an example of the same poem task:



Sonar Deep Research

User prompts can include specific website URLs. The length of the query is not restricted by the search bar; it has word wrap and multiline scroll.

TARILIO connects using streaming, so you don’t have to wait for the complete answer. It starts providing feedback as soon as it begins analysing the query.

Large Language Models (LLMs).

AI is developing at a rapid pace. Early non-reasoning models provide fast answers. They are being superseded for some tasks by reasoning models and deep research models, which have a slower response but are more suited to some business uses. When combined with real-time web search, models can overcome the knowledge cutoff date and provide links to sources, allowing users to verify information or conduct further research.

The future

Our aim is to incorporate the latest advances in LLMs to assist with information retrieval. Currently, Perplexity and certain other models enhance LLM knowledge with real-time web search.

In non-AI Mode, we index local files with a plug-in architecture that allows connection to external data sources like databases and third-party storage such as Dropbox, OneDrive, or Google Drive. On 17 April 2025, we introduced a free plugin for connection to AWS S3 Buckets.

Retrieval Augmented Generation (RAG) combines LLM knowledge with data from your private local or externally hosted sources. This, combined with reasoning or deep research models, will provide a powerful tool for retrieving and analysing a wealth of information that was not possible with older technology.

With these tools, you can make smarter data driven decisions and quickly enhance your knowledge in unfamiliar topics.

AI agents take this a step further; they are designed to autonomously perform tasks, make decisions, and interact with their environment.

AI agents integrate various AI forms and remain goal-oriented, taking targeted actions to achieve specific objectives. Think of it as having a smart virtual assistant or office system that can, for example, locate alternative suppliers, get quotes, and perhaps even produce an order ready for human approval, or book travel and accommodation for a business trip given your availability.

We will be pleased to quote for custom plugins to connect to your secure data sources, including automating tasks using MCP (Model Context Protocol) and A2A (Agent2Agent Protocol).

If you have specific information retrieval requirements, now or in the future, please contact our support at: support@electronart.co.uk, we may have it in the pipeline, or we'll put it on our roadmap!

-oo0oo-