

APIs and directly connected services: inference and hosting

Large Language Models and Agents

Hosted models – Trusted building blocks

- Research agents connect to hosted language models through an **inference API**
- Agents connect to external capabilities through **tools** and **MCP** servers
- Available tools include:
 - Web search, including **PubMed** and similar resources
 - Mathematical computation and expressions
 - Web interface development (Tailwind, shadcn)
- e-INFRA CZ provides the inference and hosted services that help researchers build and run agents

Inference Service

<https://llm.ai.e-infra.cz/v1>



- **API interface** to hosted LLM by CERIT-SC/e-INFRA CZ and Elixir CZ
 - Access to large open-weight models, including Kimi and GLM
 - Long-context models available, up to **1M tokens** where supported
- Compatible **OpenAI** and **Anthropic API**
 - Use with agentic frameworks such as Claude Code, OpenCode and Hermes
- Get your API key: <https://chat.ai.e-infra.cz>
- Setup guides: <https://docs.cerit.io>



Inference Service

<https://llm.ai.e-infra.cz/v1>

- Access with a **MetaCentrum account**
 - Cross-institution and international collaboration supported
- Fair-use policy
 - Up to **4 parallel requests**
 - Currently no fixed token or request quota
- Service status: <https://llm.ai.e-infra.cz/status>
- Monitor your usage: <https://llm.ai.e-infra.cz/usage>

Inference Service for Sensitive Data

<https://llm.cloud.trusted.e-infra.cz/v1>

- Dedicated cloud environment for **sensitive-data** processing
- Security controls aligned with the applicable **ISO 27000**-series standard
- Selected hosted models available
- Access restricted through a dedicated **VPN**
- Suitable for approved workflows involving sensitive data

Jupyter Notebook – Fire Starter

<https://hub.cloud.e-infra.cz>

- Want to try agents without installing them on your computer?
- No complicated setup required
- Start with the **DataScience AI** notebook
- Preconfigured environment with agents ready to explore
- Isolated notebook environment for experimentation
- Start experimenting directly in your **Web browser**

File Edit View Run Kernel Git Tabs Settings Help

Notebook Intelligence

User

15:08:17

create simple panda example into the current notebook

Claude Code

15:09:17

5 tool calls

Adding markdown cell

Adding code cell

Adding code cell

Adding code cell

Saving notebook

Let me run the cells so you can see the output:

Running cell

Calling tool 'mcp_nbi_run-cell'...

Parameters

{"cell_index": 1}

Running cell

Calling tool 'mcp_nbi_run-cell'...

Parameters

{"cell_index": 2}

Running cell

Calling tool 'mcp_nbi_run-cell'...

Parameters

{"cell_index": 3}

Reading cell output

Ask Notebook Intelligence...

Untitled.ipynb

🔍

🔒 ⚡ ▶

1

🔌 ⚡

Chatbook [conda env:base] * | Idle

CPU: 5.00 % | Mem: 1.68 / 16.00 GB

File Edit View Run Kernel Git Tabs Settings Help

Untitled.ipynb

× +

📄 + ✂️ 📄 📄 ▶ ⏮️ ⏭️ Code ⏏️ ↶️ ↷️ ⌚ git

~ Open in... ⚙️ Chatbook [conda env:base] ⦿

Simple Pandas Example

A quick demo of creating a DataFrame and doing basic operations.

[2]: import pandas as pd

Create a simple DataFrame from a dictionary
data = {
 "name": ["Alice", "Bob", "Charlie", "Diana"],
 "age": [25, 30, 35, 28],
 "city": ["Berlin", "Munich", "Hamburg", "Cologne"]
}
df = pd.DataFrame(data)
df

[2]:

	name	age	city
0	Alice	25	Berlin
1	Bob	30	Munich
2	Charlie	35	Hamburg
3	Diana	28	Cologne

[3]:
Basic operations
print("Average age:", df["age"].mean())

Add a new column
df["age_in_months"] = df["age"] * 12

Filter rows
adults_over_28 = df[df["age"] > 28]
adults_over_28

[3]:

	name	age	city	age_in_months
1	Bob	30	Munich	360

File Edit View Run Kernel Git Tabs Settings Help

OC | Greeting

×

+

@

hello

+ Thought: 149ms

Hello! How can I help you today?

■ Build · agentic · 8.0s

Build · agentic LiteLLM

/home/jovyan

8.0K ctrl+p commands

File Edit View Run Kernel Git Tabs Settings Help

Claude Code

×

+

@

Claude Code v2.1.283

agentic · API Usage Billing

/home/jovyan

✓ Using flicker-free rendering · if you want to go back, use /tui default
· Click to move your cursor in the text input
· Click to expand collapsed tool results
· By default, text auto-copies when you select it (/config to change)
· Hold Shift while selecting to use your terminal's native copy instead

hello

Thought for 3s

Hello! What would you like to work on?

I'm in plan mode, so if you've got a task in mind I'll explore the codebase and put together an approach for your approval before touching anything.

* Cogitated for 4s · done 1:34 PM

File Edit View Run Kernel Git Tabs Settings Help

Terminal 3

×

+

@



clarify: clarify
code_execution: execute_code
connections: manage_connections
file: patch, read_file, search_files, write_file
memory: memory
other: tool_call, tool_describe, tool_search
skills: skill_manage, skill_view, skills_list
terminal: terminal
(and 1 more toolsets...)

Available Skills
autonomous-ai-agents: claude-code, +4 more
creative: architecture-diagram, +9 more
devops: sdhc-review
email: email-inbox-triage, himalaya
media: gif-search, songsee, +1 more
note-taking: obsidian
productivity: airtable, box, +12 more
research: arxiv, +3 more
social-media: xurl
software-development: codebase-inspection, +11 more
web: blocked-page-recovery

16 tools · 54 skills · /help for commands
⬆️ 21 commits behind – run hermes update to update

Welcome to Hermes Agent! Type your message or /help for commands.
+ Tip: Ctrl+C interrupts the agent. Double-press within 2 seconds to force exit.

● hello

Initializing agent...

▼ Hermes

Hello – what can I help you with?

▼ kimi

10.1K/262.1K

[

4%

9.1s

↑...

Simple friendly greeting

›

Type / to browse commands, or Ctrl+P for the palette

›

accept edits on (shift+tab to cycle) · ← for agents

3

🔌 0

CPU: 0.00 % | Mem: 2.50 / 16.00 GB

Terminal 3 0 🔔

7

The Takeaway

Trusted building blocks for research agents

- Hosted inference APIs for agentic systems
- Compatible interfaces for existing agent frameworks
- A dedicated environment for approved sensitive-data workflows
- A ready-to-use JupyterHub sandbox for experimentation
- **Community support:** <https://element.cloud.e-infra.cz>

Thank you for your attention

Don't be afraid of AI

k8s@cerit-sc.cz

<https://docs.cerit.io>

<https://blog.e-infra.cz>