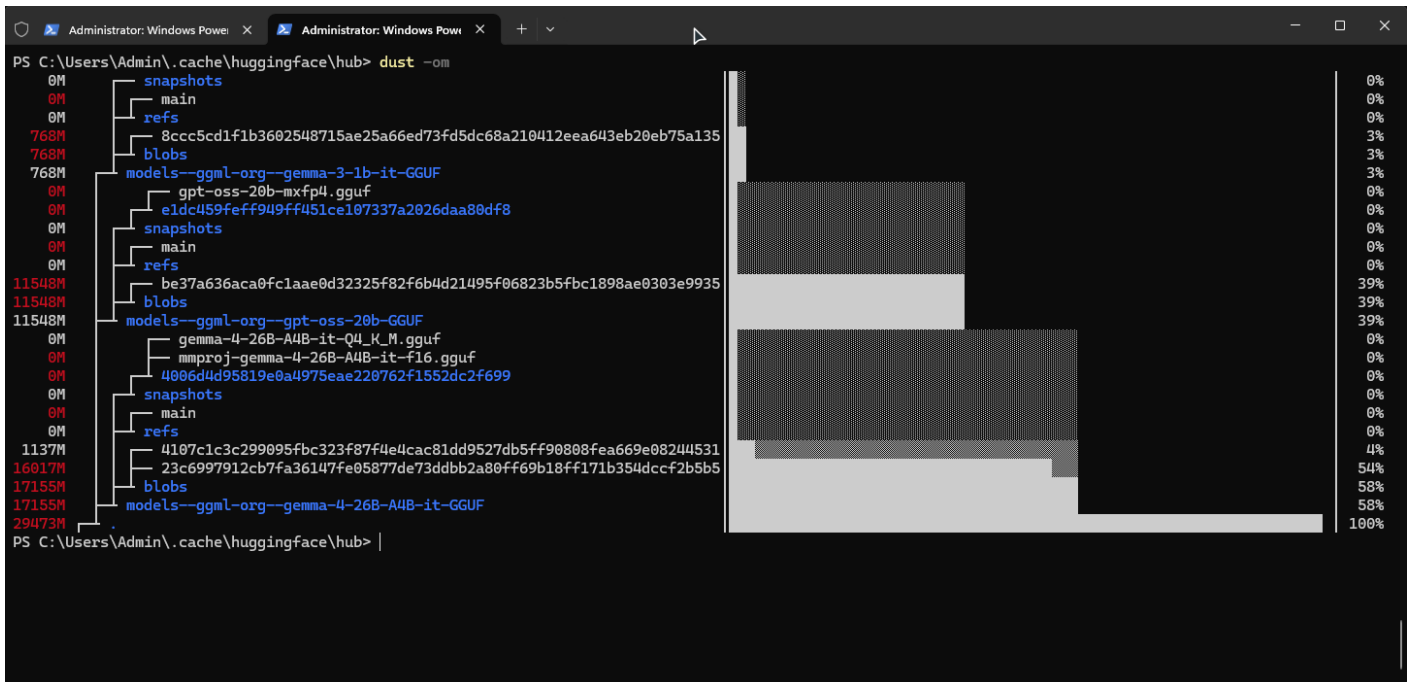


Some Output

Gemma 4 26B A4B - Server

```
llama-server.exe -hf ggml-org/gemma-4-26B-A4B-it-GGUF:Q4_K_M -ngl 99 --  
rpc 192.168.0.91:21000,192.168.0.92:21000 -c 4096
```



Input:

Hello Gemma. HOw are you doing? HOw much do you know?

Thinking

- "Hello Gemma. HOw are you doing? HOw much do you know?"
 - Greeting: "Hello Gemma."
 - Social Inquiry: "How are you doing?"
 - Capability Inquiry: "How much do you know?"

- *Greeting/Social*: I'm an AI, so I don't have feelings, but I'm functioning well and ready to help.
- *Capability/Knowledge*: I'm a large language model trained by Google. I have access to a vast amount of information (text, code, facts, languages, etc.) up to my training cutoff. I can process complex instructions, summarize, translate, code, and more.
- *Tone*: Friendly, professional, helpful, and transparent.
- *Part 1: Greeting/Status*
 - "Hello! I'm doing great, thank you for asking. I'm ready and eager to help you with whatever you need."
- *Part 2: Knowledge Scope*
 - *What am I?* A large language model trained by Google.
 - *What do I know?* A massive dataset of text, code, and information.
 - *Specifics*: History, science, math, literature, coding, languages, trivia, etc.
 - *Limitations*: I don't "know" things in the human sense (no consciousness/personal experience). I don't have real-time access to everything (though I can browse if enabled, but generally, I rely on training data). I can make mistakes (hallucinations).
- *Drafting the response*: "Hello! I'm doing very well, thank you for asking. How are you?"

As for how much I know, it's a bit hard to put a specific number on it, but I have been trained on a massive dataset of text and code. You can think of me as having access to a vast library of human knowledge.

Here is a breakdown of what I can do:

- **General Knowledge**: I can answer questions about history, science, geography, literature, and more.

- **Technical Skills:** I can write and debug code in many different programming languages, solve math problems, and explain complex technical concepts.
- **Language & Writing:** I can translate languages, summarize long articles, write stories/poems/emails, and check grammar.
- **Reasoning:** I can help with brainstorming, logic puzzles, and organizing information.

A small disclaimer: While I know a lot, I don't 'know' things the way a human does. I don't have personal experiences, feelings, or a physical presence. Also, I can occasionally make mistakes or present outdated information, so it's always a good idea to double-check important facts!"

- Ensure the tone is inviting.
- Make sure the "How much do you know?" part is comprehensive but not overwhelming.

Output:

```
Hello! I'm doing very well, thank you for asking! I'm powered up and ready to help you. How are you doing today?
```

```
As for how much I know, it is difficult to put an exact number on it, but you can think of me as having access to a massive, digital library containing a vast portion of human knowledge.
```

```
Here is a breakdown of the scope of my "knowledge":
```

```
### 1. What I can do well:
```

```
* General Knowledge: I can answer questions about history, science, geography, literature, pop culture, and more.
```

```
* Technical Skills: I can write, explain, and debug code in many programming languages (like Python, JavaScript, C++, etc.), solve mathematical equations, and explain complex scientific concepts.
```

```
* Language & Communication: I can translate dozens of languages, summarize long articles, write different styles of text (emails, essays, poems, stories), and help with grammar and editing.
```

```
* Reasoning & Logic: I can help with brainstorming, organizing ideas, solving logic puzzles, and breaking down complex problems into manageable steps.
```

```
### 2. My Limitations (What I *don't* know):
* **No Personal Experience:** I don't "know" what coffee tastes like
or how it feels to be happy. I only know the *descriptions* of those
things found in text.
* **No Sentience:** I don't have feelings, beliefs, or a physical
presence in the world.
* **The "Real-Time" Gap:** While I am constantly being updated, I
don't always have "live" access to events happening at this very second
unless I use specific tools to search the web.
* **Potential for Errors:** I can sometimes "hallucinate," which
means I might confidently state something that is factually incorrect.
It is always a good idea to double-check critical information.

**In short: I know a lot about the world as it has been recorded in
text, but I am a tool designed to assist you, not an infallible source
of truth.**

Is there something specific you'd like to test my knowledge on?
```

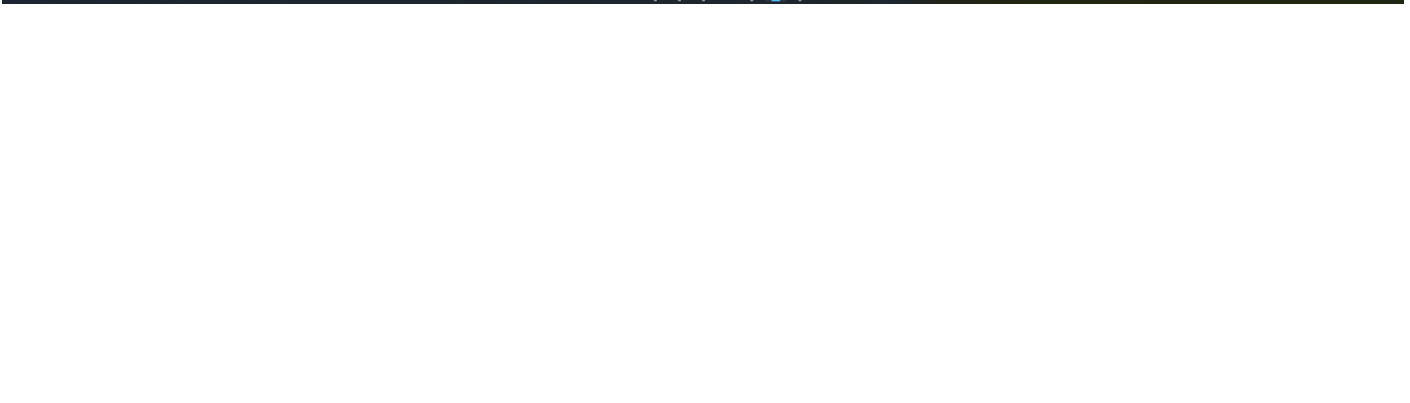
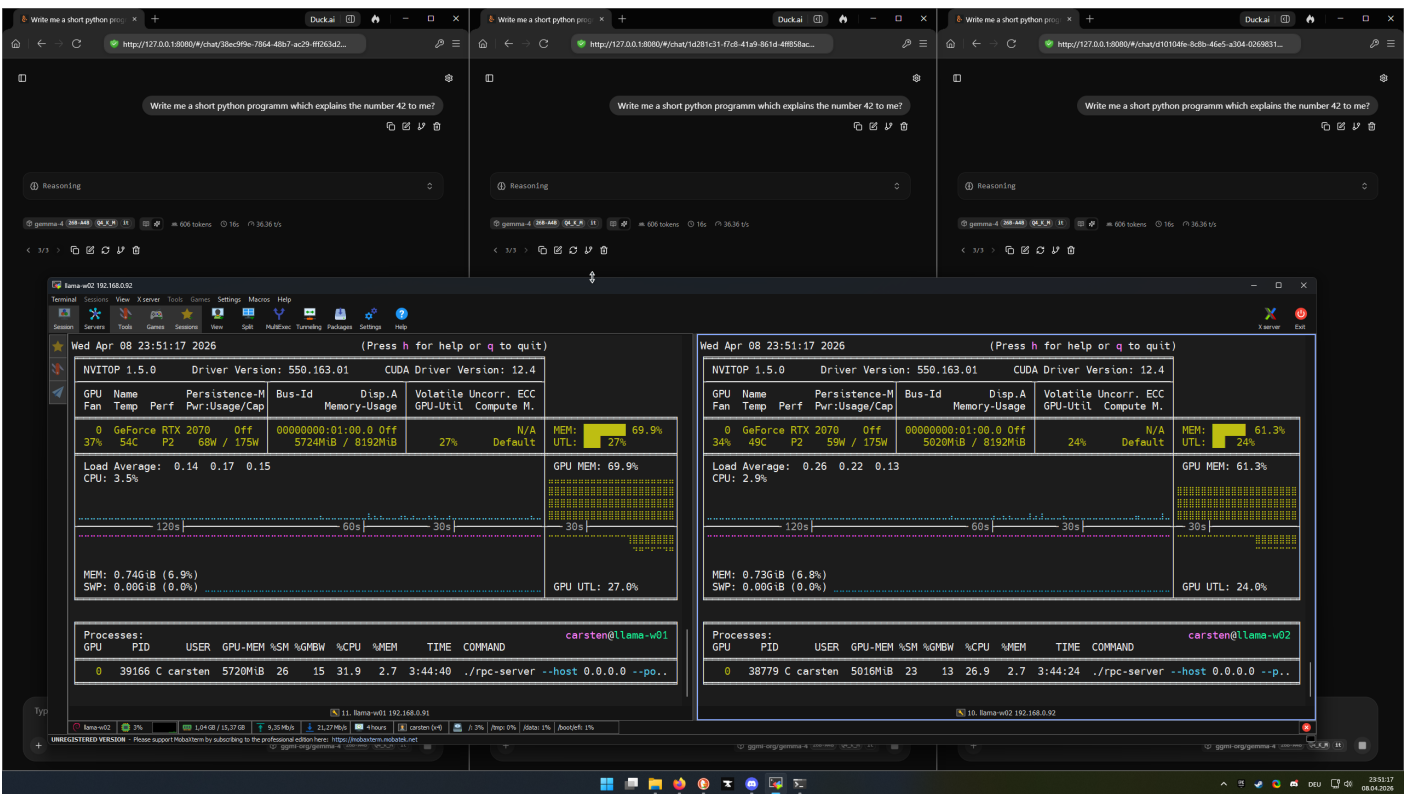
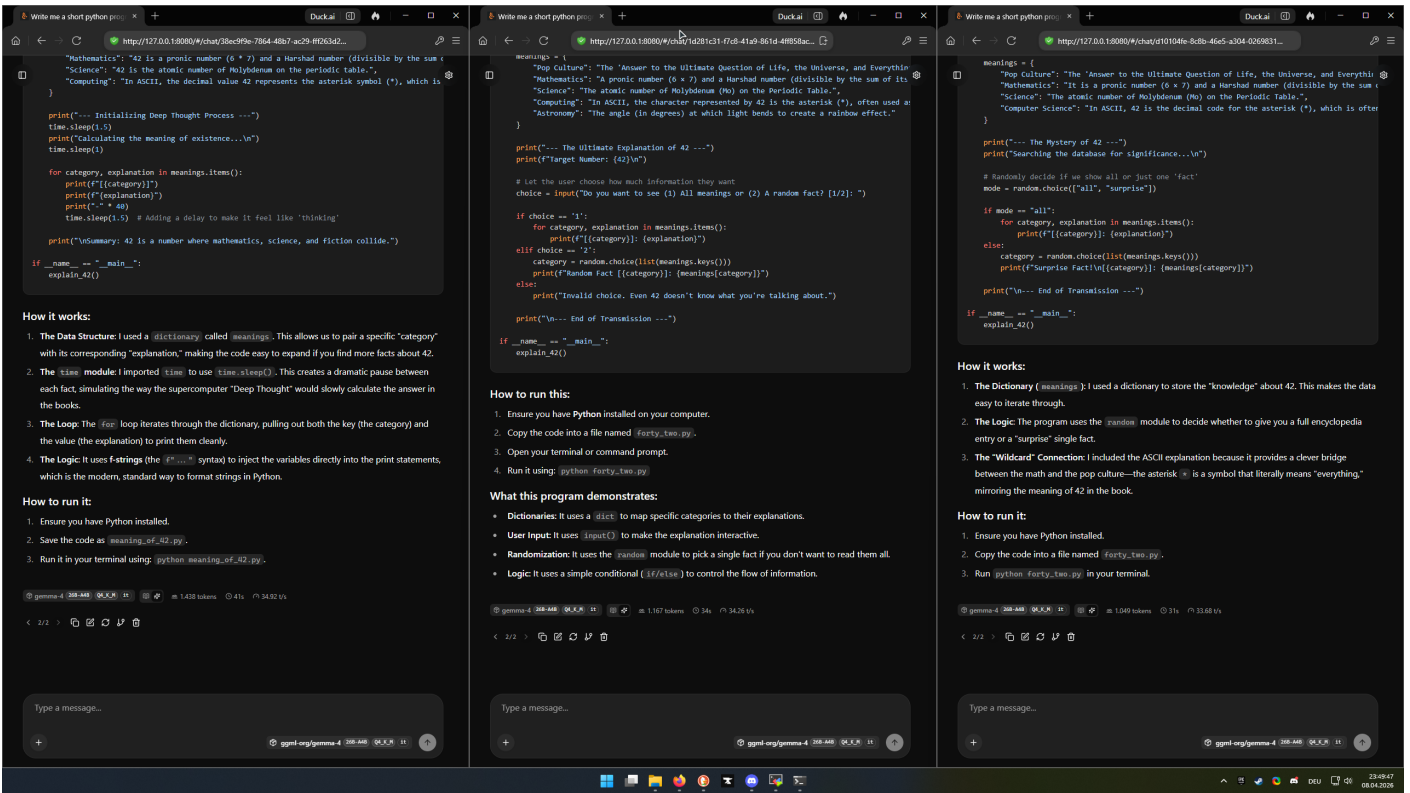
ggml-org/gemma-4-26B-A4B-it-GGUF:Q4_K_M -- 1.111 tokens -- 35s -- 31.14 t/s

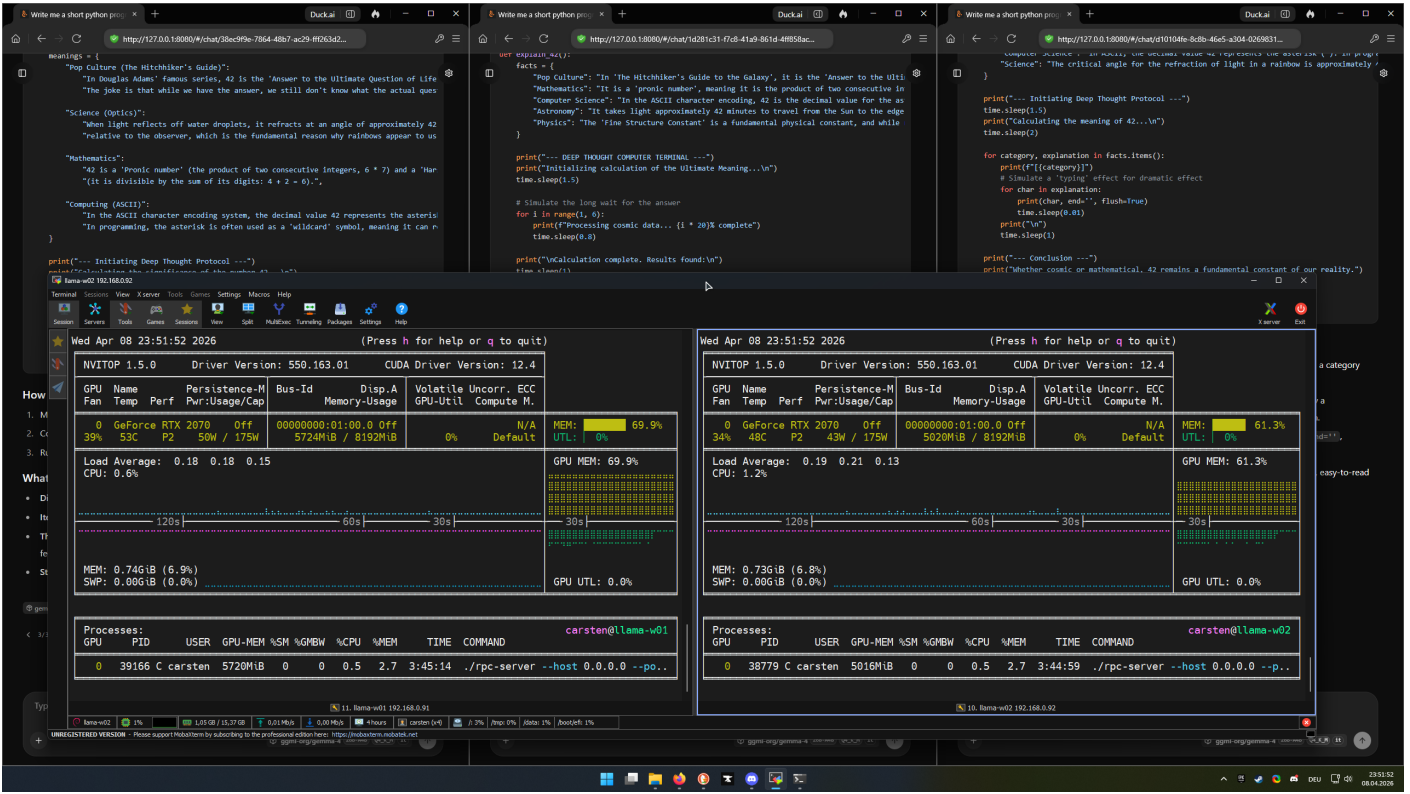
Only One Machine available only anymore

```
start: binding port with default address family
main: loading model
srv   load_model: loading model 'C:\Users\Admin\.cache\huggingface\hub\models--ggml-org--gemma-4-26B-A4B-it-GGUF\snapshots\4006d4d95819e0a4975eae220762f1552dc2f699\gemma-4-26B-A4B-it-Q4_K_M.gguf'
common_init_result: fitting params to device memory, for bugs during this step try to reproduce them with -fit off, or provide --verbose logs if the bug only
occurs with -fit on
llama_params_fit_impl: projected to use 17415 MiB of device memory vs. 11084 MiB of free device memory
llama_params_fit_impl: cannot meet free memory target of 1024 MiB, need to reduce device memory by 7354 MiB
llama_params_fit_impl: context size set by user to 4096 -> no change
llama_params_fit: failed to fit params to free device memory: n_gpu_layers already set by user to 99, abort
llama_params_fit: fitting params to free memory took 0.65 seconds
llama_model_load_from_file_impl: using device Vulkan0 (NVIDIA GeForce RTX 5070) (0000:01:00:0) - 11084 MiB free
llama_model_loader: loaded meta data with 47 key-value pairs and 658 tensors from C:\Users\Admin\.cache\huggingface\hub\models--ggml-org--gemma-4-26B-A4B-it-GGUF\snapshots\4006d4d95819e0a4975eae220762f1552dc2f699\gemma-4-26B-A4B-it-Q4_K_M.gguf (version GGUF V3 (latest))
llama_model_loader: Dumping metadata keys/values. Note: KV overrides do not apply in this output.
```

It is going to be a tough one. And it does not load, as not enough memory available.

Three chats at the same time





Revision #1

Created 2026-04-08 21:39:37 UTC by Carsten

Updated 2026-04-08 21:56:09 UTC by Carsten