nexplore
mission possible

03.07.2024

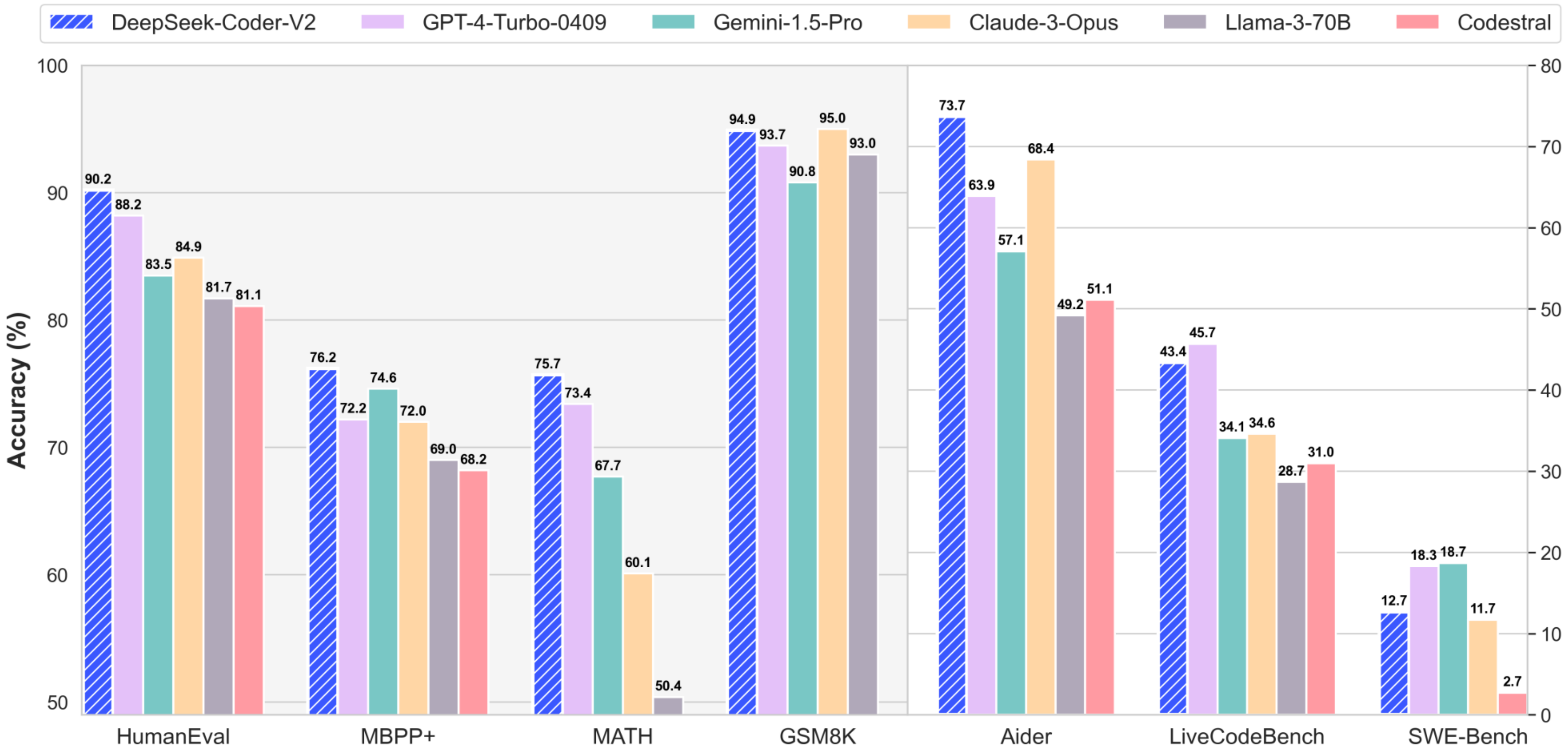
Jan Badertscher - Better than GPT-4 for Coding

AI Meetup Bern

Agenda

1. State of the Art Models
2. How to run Open Models for Coding?
3. Code Completion vs. Code Instructions
4. Demo
5. Outlook into the future

“DeepSeek Coder V2 beats GPT-4 at coding”



https://github.com/deepseek-ai/DeepSeek-Coder-V2

State of the Art

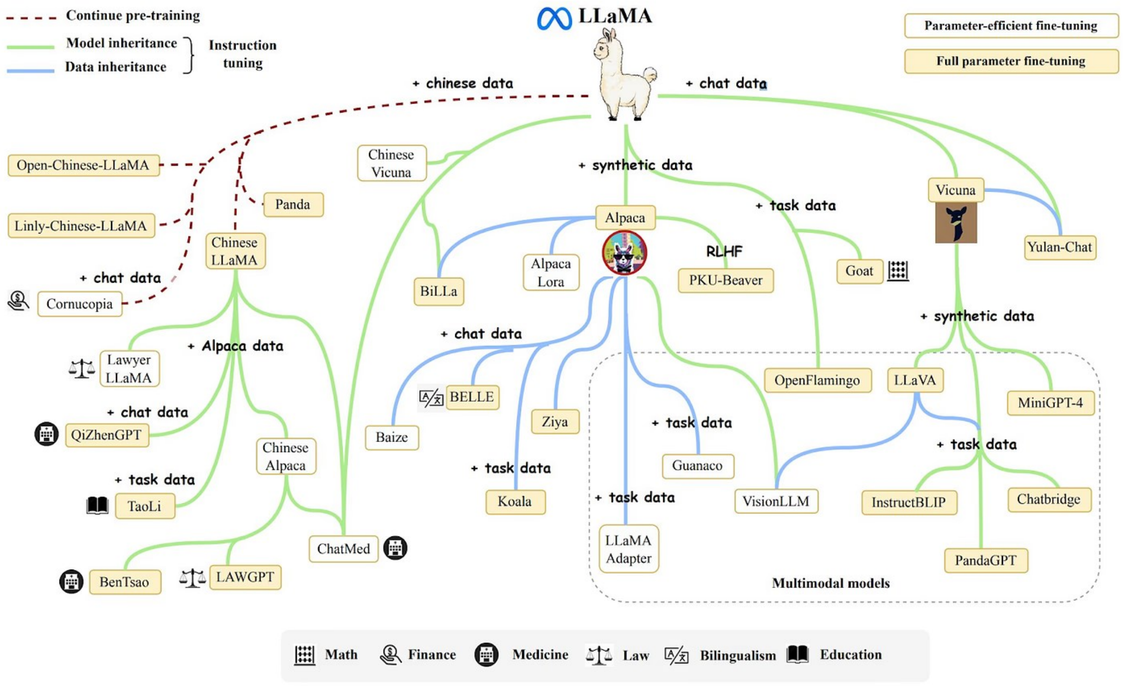
Coding Models and Leaderboards

The LLM Space

Many model types

- OpenAI's GPT is not the only LLM
- Meta's Llama
- Google's Gemma
- Microsoft's Phi
- …

Llama Model Tree



Source: https://blackbearlabs.ai/blog-detail/open-large-language-models-history-2023-report

Source: https://lifearchitect.ai/timeline/

Open vs. Closed

Things to consider

- Open Source != Open Weights
- Consider:
  - Training Data
  - Model Weights
  - Code
  - License
- Model License lists: https://github.com/eugeneyan/open-llms

License Types

LICENSE	PERMISSIVE OR COPYLEFT	PATENT GRANT	COMMERCIAL USE	REDISTRIBUTION	MODIFICATION
Apache 2.0	Permissive	Yes	Yes (with attribution)	Yes	Yes
MIT	Permissive	No	Yes (with attribution)	Yes	Yes
GPL-3.0	Copyleft	Yes (for GPL-3.0 licensed software only)	Yes (with source code)	Yes (with source code)	Yes
Proprietary	Varies	Varies	Varies	Varies	Varies

API vs Local models for coding

API

- Github Copilot (Codex / GPT-3.5)
- Claude 3.5 Sonnet
- GPT-4

Local (Open Weights)

- CodeLlama
- CodeQwen
- DeepSeek-Coder
- Phind
- Granit-Code
- Mixtral
- Command-R
- Codestral
- CodeGemma
- Starcoder2

Remember

- New models, architectures, paradigms
- Licencing
- Hosting vs API

Leaderboards

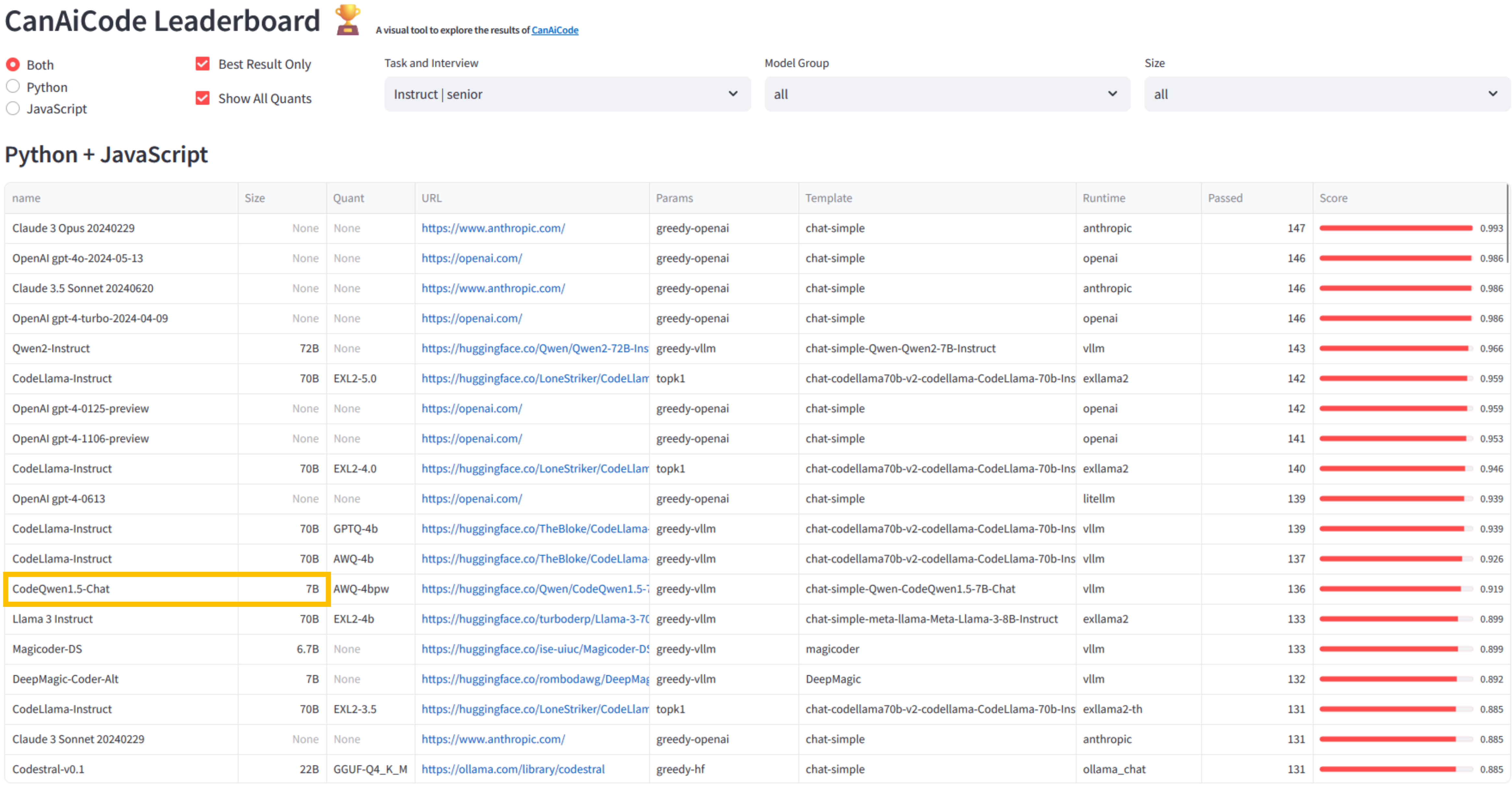
Evaluate Coding Models

Benchmarks

- HumanEval
- EvalPlus
- LLM as a Judge
- CanAiCode
- Aider
- SWE-bench

Leaderboards

- https://huggingface.co/spaces/bigcode/bigcode-models-leaderboard
- https://aider.chat/docs/leaderboards/
- https://prollm.toqan.ai/leaderboard/coding-assistant
- https://evalplus.github.io/leaderboard.html
- https://huggingface.co/spaces/mike-ravkine/can-ai-code-results
- https://www.swebench.com/



CanAiCode Leaderboard 🏆

A visual tool to explore the results of CanAiCode

● Both
○ Python
○ JavaScript

☑ Best Result Only

☑ Show All Quants

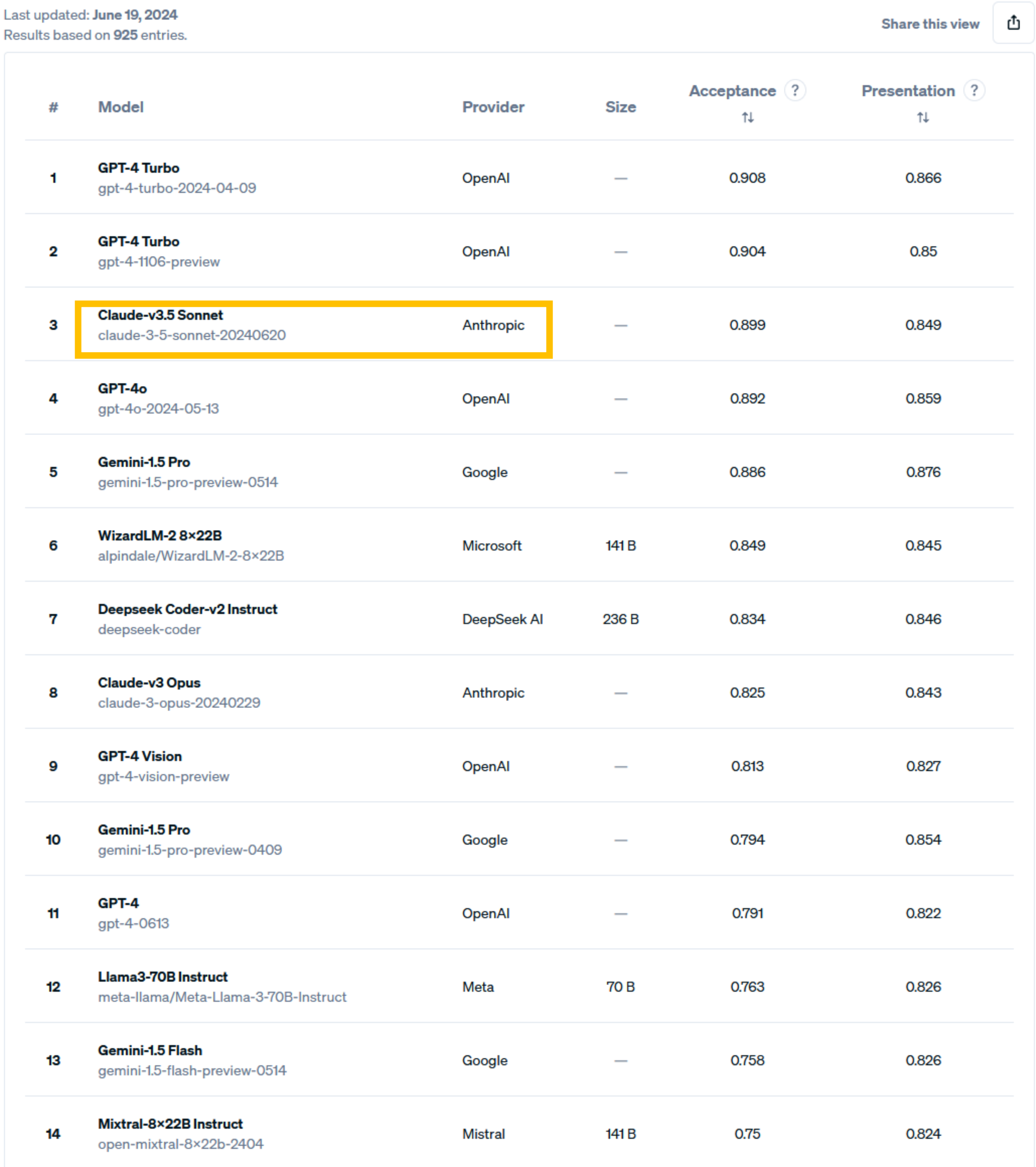
Task and Interview: Instruct | senior

Model Group: all

Size: all

Python + JavaScript

name	Size	Quant	URL	Params	Template	Runtime	Passed	Score
Claude 3 Opus 20240229	None	None	https://www.anthropic.com/	greedy-openai	chat-simple	anthropic	147	0.993
OpenAI gpt-4o-2024-05-13	None	None	https://openai.com/	greedy-openai	chat-simple	openai	146	0.986
Claude 3.5 Sonnet 20240620	None	None	https://www.anthropic.com/	greedy-openai	chat-simple	anthropic	146	0.986
OpenAI gpt-4-turbo-2024-04-09	None	None	https://openai.com/	greedy-openai	chat-simple	openai	146	0.986
Qwen2-Instruct	72B	None	https://huggingface.co/Qwen/Qwen2-72B-Ins	greedy-vllm	chat-simple-Qwen-Qwen2-7B-Instruct	vllm	143	0.966
CodeLlama-Instruct	70B	EXL2-5.0	https://huggingface.co/LoneStriker/CodeLlam	topk1	chat-codellama70b-v2-codellama-CodeLlama-70b-Ins	exllama2	142	0.959
OpenAI gpt-4-0125-preview	None	None	https://openai.com/	greedy-openai	chat-simple	openai	142	0.959
OpenAI gpt-4-1106-preview	None	None	https://openai.com/	greedy-openai	chat-simple	openai	141	0.953
CodeLlama-Instruct	70B	EXL2-4.0	https://huggingface.co/LoneStriker/CodeLlam	topk1	chat-codellama70b-v2-codellama-CodeLlama-70b-Ins	exllama2	140	0.946
OpenAI gpt-4-0613	None	None	https://openai.com/	greedy-openai	chat-simple	litellm	139	0.939
CodeLlama-Instruct	70B	GPTQ-4b	https://huggingface.co/TheBloke/CodeLlama-	greedy-vllm	chat-codellama70b-v2-codellama-CodeLlama-70b-Ins	vllm	139	0.939
CodeLlama-Instruct	70B	AWQ-4b	https://huggingface.co/TheBloke/CodeLlama-	greedy-vllm	chat-codellama70b-v2-codellama-CodeLlama-70b-Ins	vllm	137	0.926
CodeQwen1.5-Chat	7B	AWQ-4bpw	https://huggingface.co/Qwen/CodeQwen1.5-7	greedy-vllm	chat-simple-Qwen-CodeQwen1.5-7B-Chat	vllm	136	0.919
Llama 3 Instruct	70B	EXL2-4b	https://huggingface.co/turboderp/Llama-3-70	greedy-vllm	chat-simple-meta-llama-Meta-Llama-3-8B-Instruct	exllama2	133	0.899
Magicoder-DS	6.7B	None	https://huggingface.co/ise-uiuc/Magicoder-DS	greedy-vllm	magicoder	vllm	133	0.899
DeepMagic-Coder-Alt	7B	None	https://huggingface.co/rombodawg/DeepMag	greedy-vllm	DeepMagic	vllm	132	0.892
CodeLlama-Instruct	70B	EXL2-3.5	https://huggingface.co/LoneStriker/CodeLlam	topk1	chat-codellama70b-v2-codellama-CodeLlama-70b-Ins	exllama2-th	131	0.885
Claude 3 Sonnet 20240229	None	None	https://www.anthropic.com/	greedy-openai	chat-simple	anthropic	131	0.885
Codestral-v0.1	22B	GGUF-Q4_K_M	https://ollama.com/library/codestral	greedy-hf	chat-simple	ollama_chat	131	0.885



Last updated: **June 19, 2024**
Results based on **925** entries.

#	Model	Provider	Size	Acceptance ? ⇅	Presentation ? ⇅
1	**GPT-4 Turbo** gpt-4-turbo-2024-04-09	OpenAI	—	0.908	0.866
2	**GPT-4 Turbo** gpt-4-1106-preview	OpenAI	—	0.904	0.85
3	**Claude-v3.5 Sonnet** claude-3-5-sonnet-20240620	Anthropic	—	0.899	0.849
4	**GPT-4o** gpt-4o-2024-05-13	OpenAI	—	0.892	0.859
5	**Gemini-1.5 Pro** gemini-1.5-pro-preview-0514	Google	—	0.886	0.876
6	**WizardLM-2 8×22B** alpindale/WizardLM-2-8×22B	Microsoft	141 B	0.849	0.845
7	**Deepseek Coder-v2 Instruct** deepseek-coder	DeepSeek AI	236 B	0.834	0.846
8	**Claude-v3 Opus** claude-3-opus-20240229	Anthropic	—	0.825	0.843
9	**GPT-4 Vision** gpt-4-vision-preview	OpenAI	—	0.813	0.827
10	**Gemini-1.5 Pro** gemini-1.5-pro-preview-0409	Google	—	0.794	0.854
11	**GPT-4** gpt-4-0613	OpenAI	—	0.791	0.822
12	**Llama3-70B Instruct** meta-llama/Meta-Llama-3-70B-Instruct	Meta	70 B	0.763	0.826
13	**Gemini-1.5 Flash** gemini-1.5-flash-preview-0514	Google	—	0.758	0.826
14	**Mixtral-8×22B Instruct** open-mixtral-8×22b-2404	Mistral	141 B	0.75	0.824

https://prollm.toqan.ai/leaderboard/coding-assistant

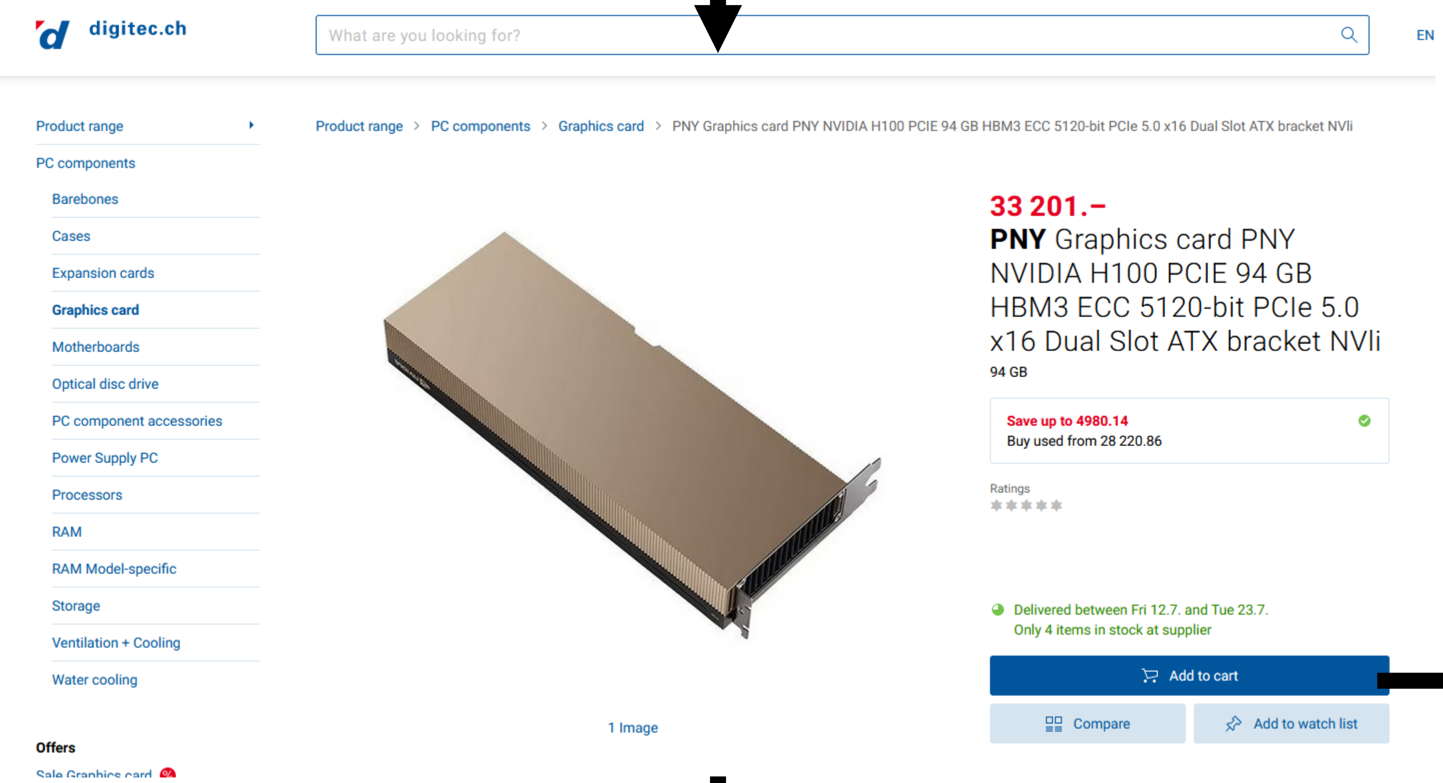
Take it with a grain of salt

- Figures up to date?
- Cheating in Benchmarks
- Measuring the right thing?
- Outdated Benchmarks Evaluation

How to run Open Models?

Hardware

Do you belong to the GPU Poor?



digitec.ch

What are you looking for?

EN

- Product range
- PC components
  - Barebones
  - Cases
  - Expansion cards
  - **Graphics card**
  - Motherboards
  - Optical disc drive
  - PC component accessories
  - Power Supply PC
  - Processors
  - RAM
  - RAM Model-specific
  - Storage
  - Ventilation + Cooling
  - Water cooling

Offers

Sale Graphics card

Product range > PC components > Graphics card > PNY Graphics card PNY NVIDIA H100 PCIE 94 GB HBM3 ECC 5120-bit PCIe 5.0 x16 Dual Slot ATX bracket NVli

1 Image

33 201.–

PNY Graphics card PNY NVIDIA H100 PCIE 94 GB HBM3 ECC 5120-bit PCIe 5.0 x16 Dual Slot ATX bracket NVli

94 GB

Save up to 4980.14
Buy used from 28 220.86

Ratings

Delivered between Fri 12.7. and Tue 23.7.
Only 4 items in stock at supplier

Add to cart

Compare

Add to watch list

Not GPU poor

All others

☹

API vs Host

API

- Groq
- Codestral API
- TogetherAI
- HuggingFace Inference
- Deepinfra
- Replicate
- AWS Bedrock
- Azure AI
- OpenRouter

Host

Large scale

- vLLM
- Nvidia NIM

Small scale

- Ollama
- LM Studio
- Open Webui
- Llama.cpp

IDE Copilots

Extensions

VS Code

- Continue.dev
  https://continue.dev
- Tabby
  https://github.com/Eugeny/tabby

Other projects for

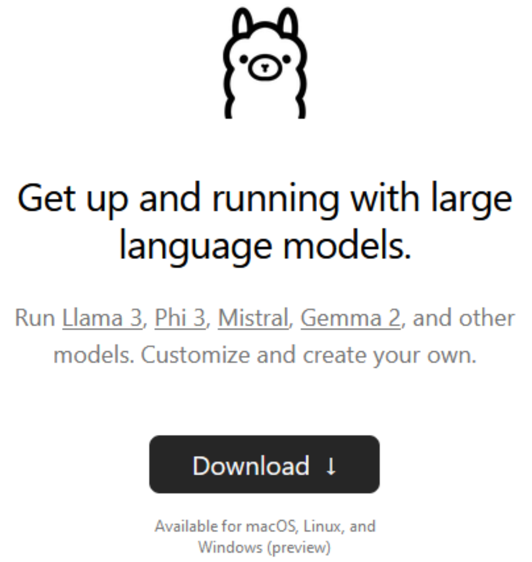
- JetBrains
- Jupyter Labs
- ...

Functionality

Tasks

- Tab Autocomplete
- Explain Code
- Debug Code
- Chat with your codebase
- Generate code

Easy start with Ollama

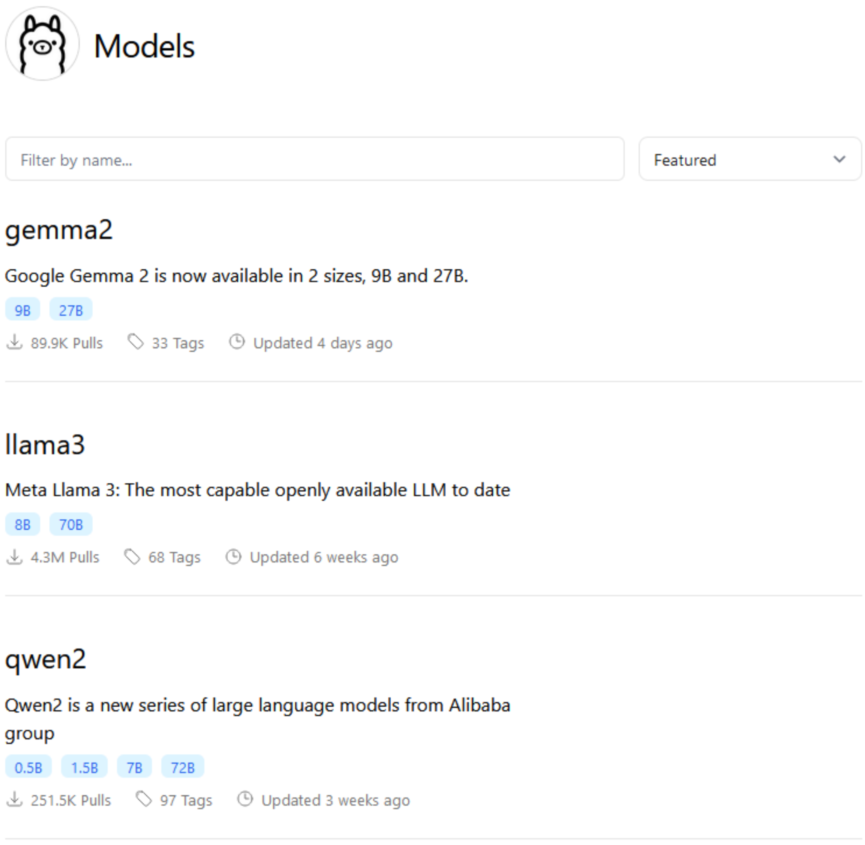


Get up and running with large language models.

Run Llama 3, Phi 3, Mistral, Gemma 2, and other models. Customize and create your own.

Available for macOS, Linux, and Windows (preview)

Install from Ollama.com



Models

gemma2

Google Gemma 2 is now available in 2 sizes, 9B and 27B.

9B 27B

89.9K Pulls 33 Tags Updated 4 days ago

llama3

Meta Llama 3: The most capable openly available LLM to date

8B 70B

4.3M Pulls 68 Tags Updated 6 weeks ago

qwen2

Qwen2 is a new series of large language models from Alibaba group

0.5B 1.5B 7B 72B

251.5K Pulls 97 Tags Updated 3 weeks ago

Select a model in the catalog

Code completion vs. Chat

Code completion

- Models that do Fill in the middle
- Base models
- Can't chat

- Small, fast and dumb

Chat models

- Can follow chat instructions
- Can generate complete code

- Big, slow and smart

Some good models

	Small	Medium	Large
Code Completion	Codegemma 3b code Starcoder2 3b	CodeQwen1.5 7b	DeepSeek-Coder-V2-Base 236B
Chat	---	CodeQwen1.5 7b chat	DeepSeek-Coder-V2-Base 236B Instruct

Running Models

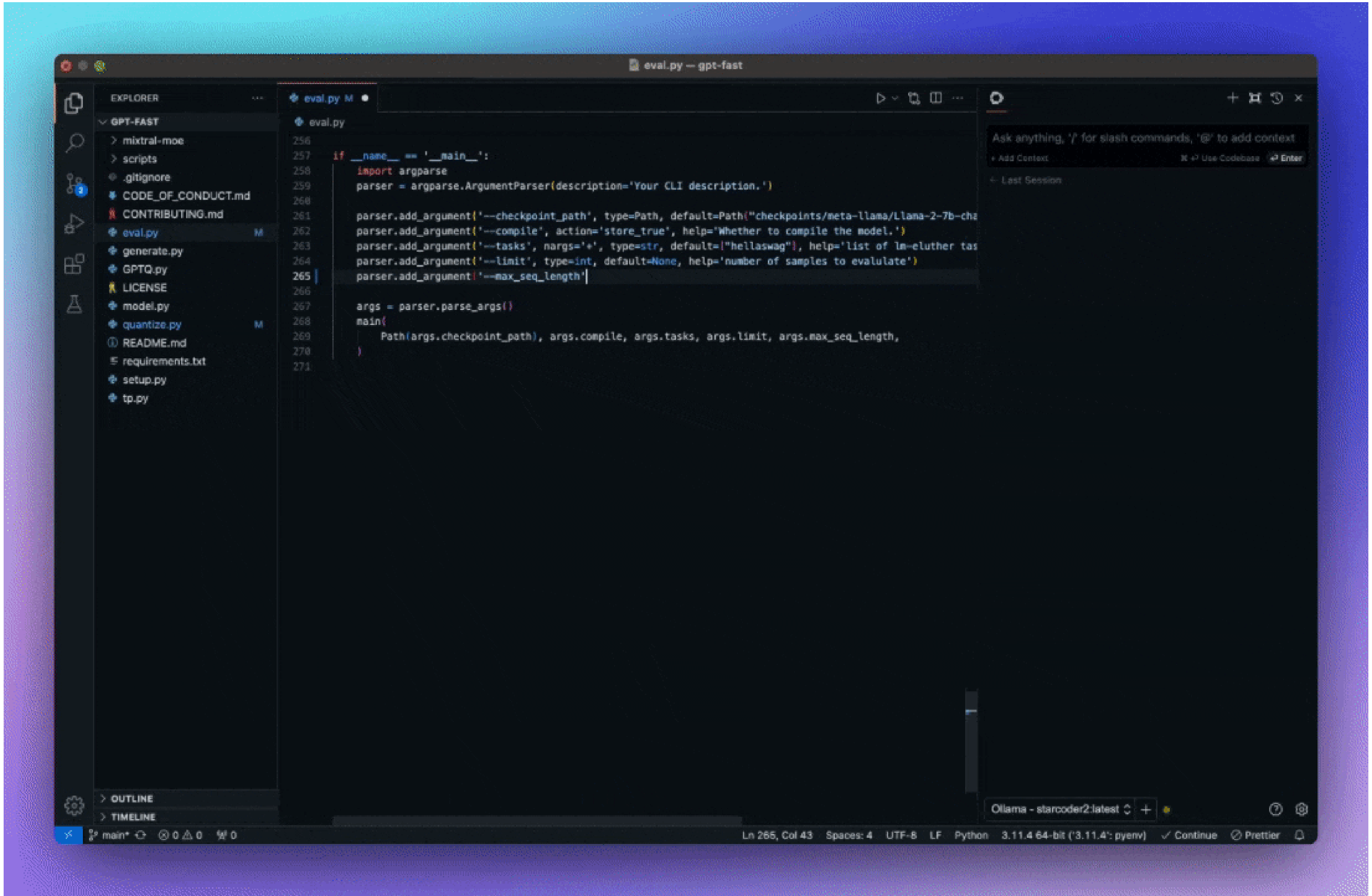
- Parameter size depends on vRAM
- Size != Quality

Use the IDE Extension

What’s the promise?

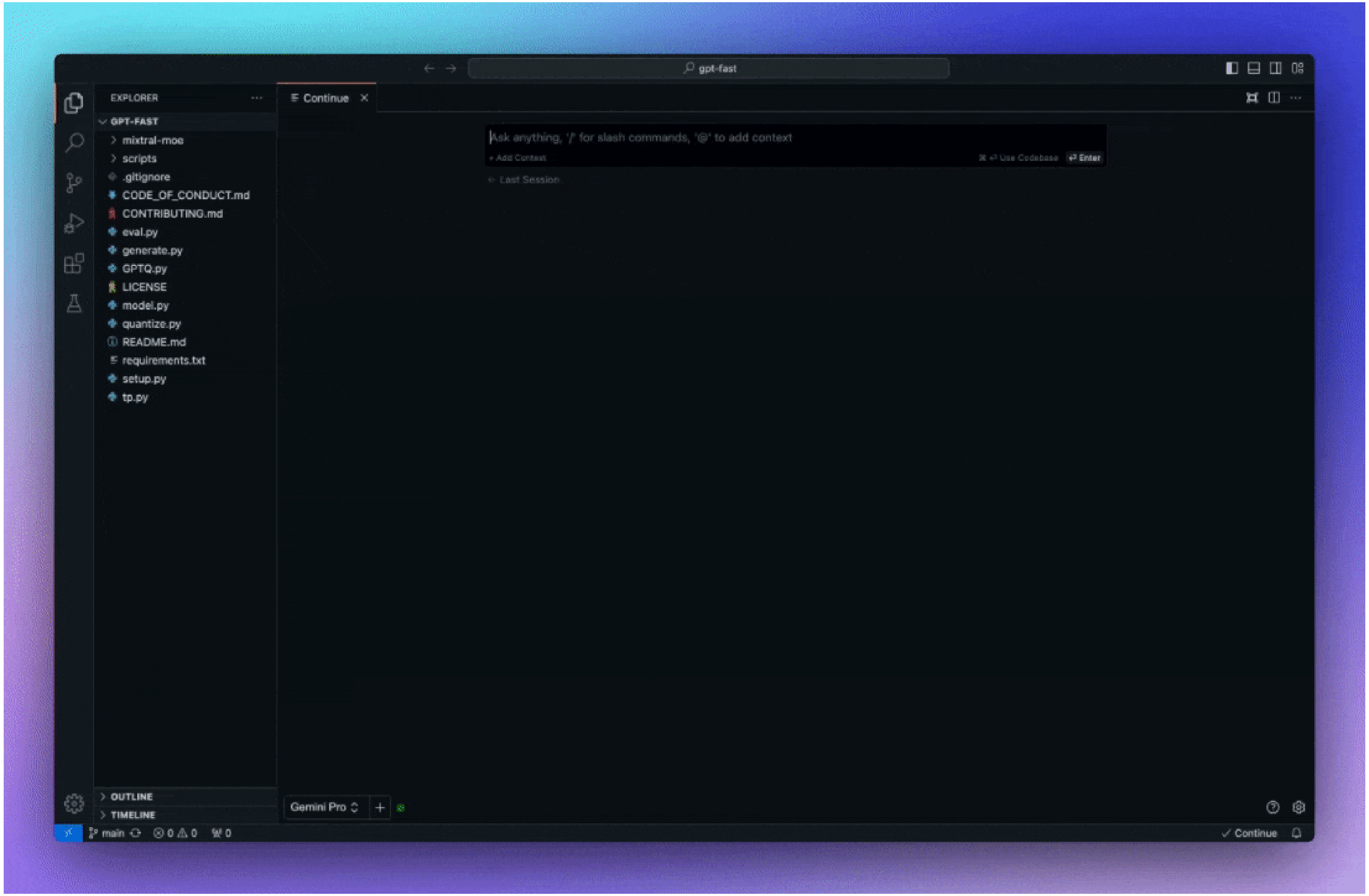
- Better suggestions
- Faster writing
- Smarter bug fixing

Smarter Auto Complete

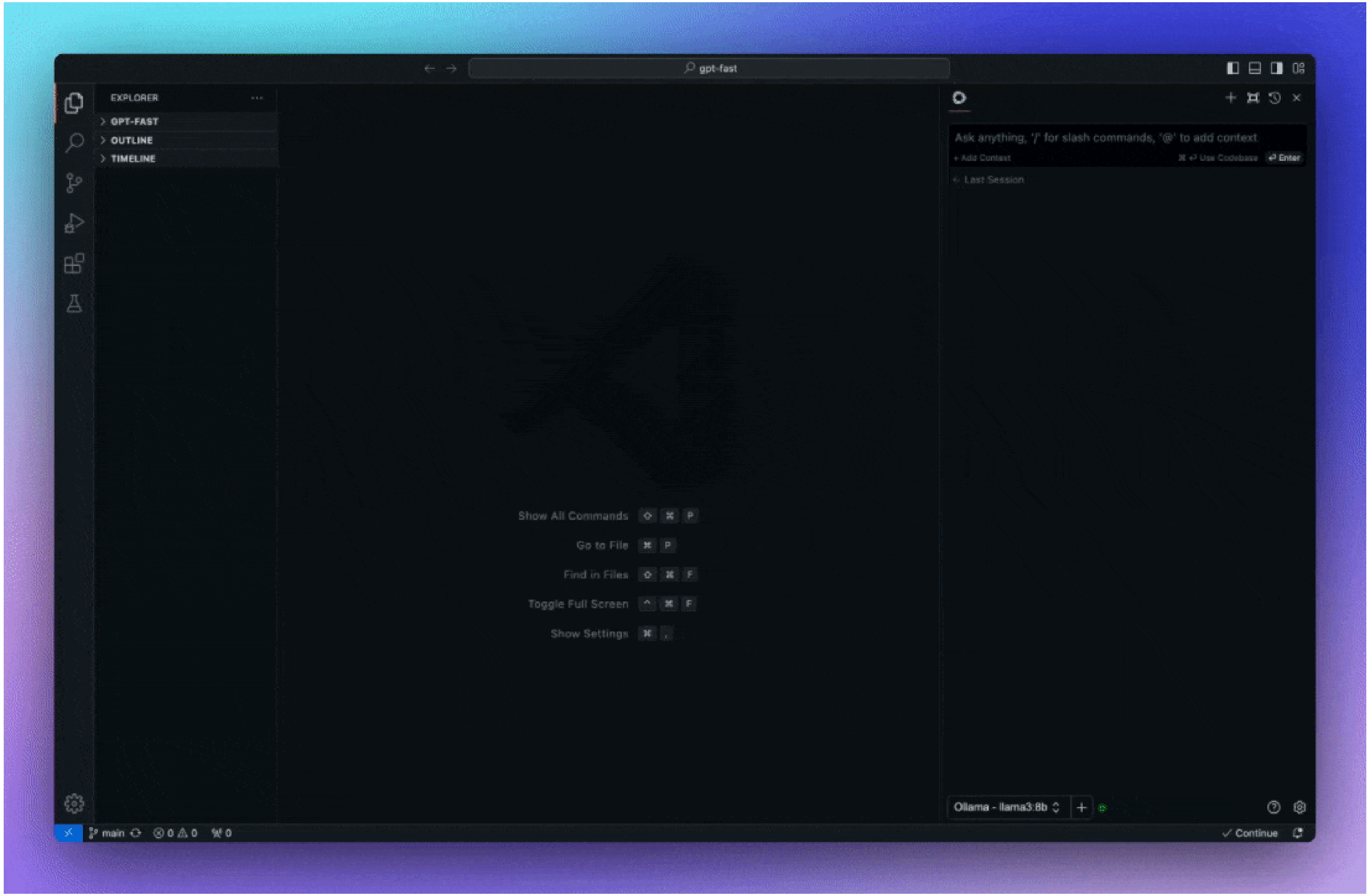


```python
if __name__ == '__main__':
    import argparse
    parser = argparse.ArgumentParser(description='Your CLI description.')

    parser.add_argument('--checkpoint_path', type=Path, default=Path("checkpoints/meta-llama/Llama-2-7b-cha
    parser.add_argument('--compile', action='store_true', help='Whether to compile the model.')
    parser.add_argument('--tasks', nargs='+', type=str, default=["hellaswag"], help='list of lm-eluther tas
    parser.add_argument('--limit', type=int, default=None, help='number of samples to evalulate')
    parser.add_argument('--max_seq_length'

    args = parser.parse_args()
    main(
        Path(args.checkpoint_path), args.compile, args.tasks, args.limit, args.max_seq_length,
    )
```

Chat with your codebase



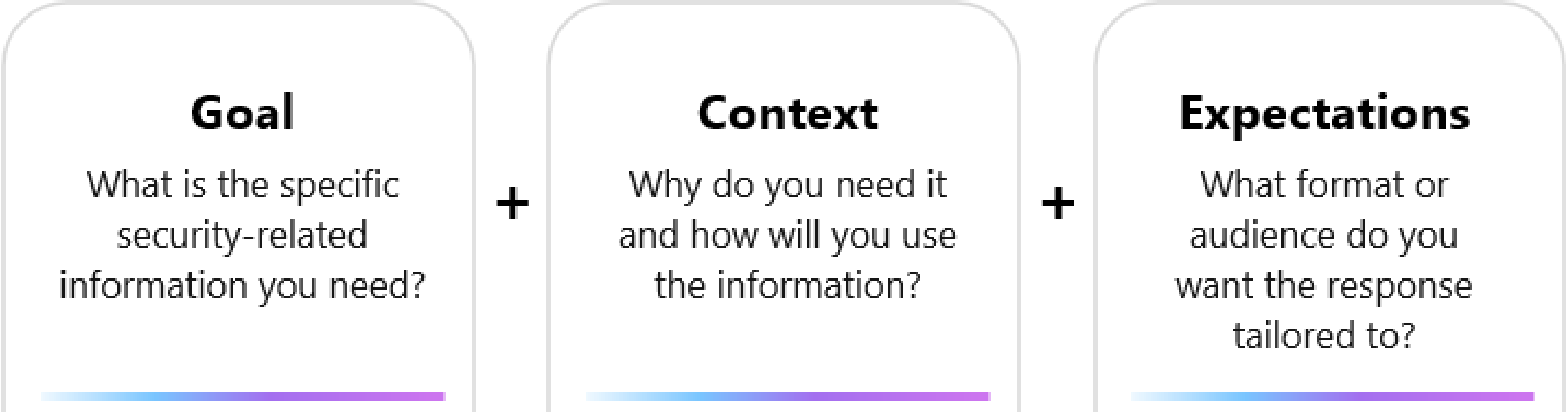
Aks your documentation



Context is King

- Reference classes, files and more to context in your prompt
  @Classname
  @Filename
  @DocumentationLink
- Ask to solve help with Terminal Errors CTRL+SHIFT+R

Good Prompting



Goal

What is the specific security-related information you need?

+

Context

Why do you need it and how will you use the information?

+

Expectations

What format or audience do you want the response tailored to?

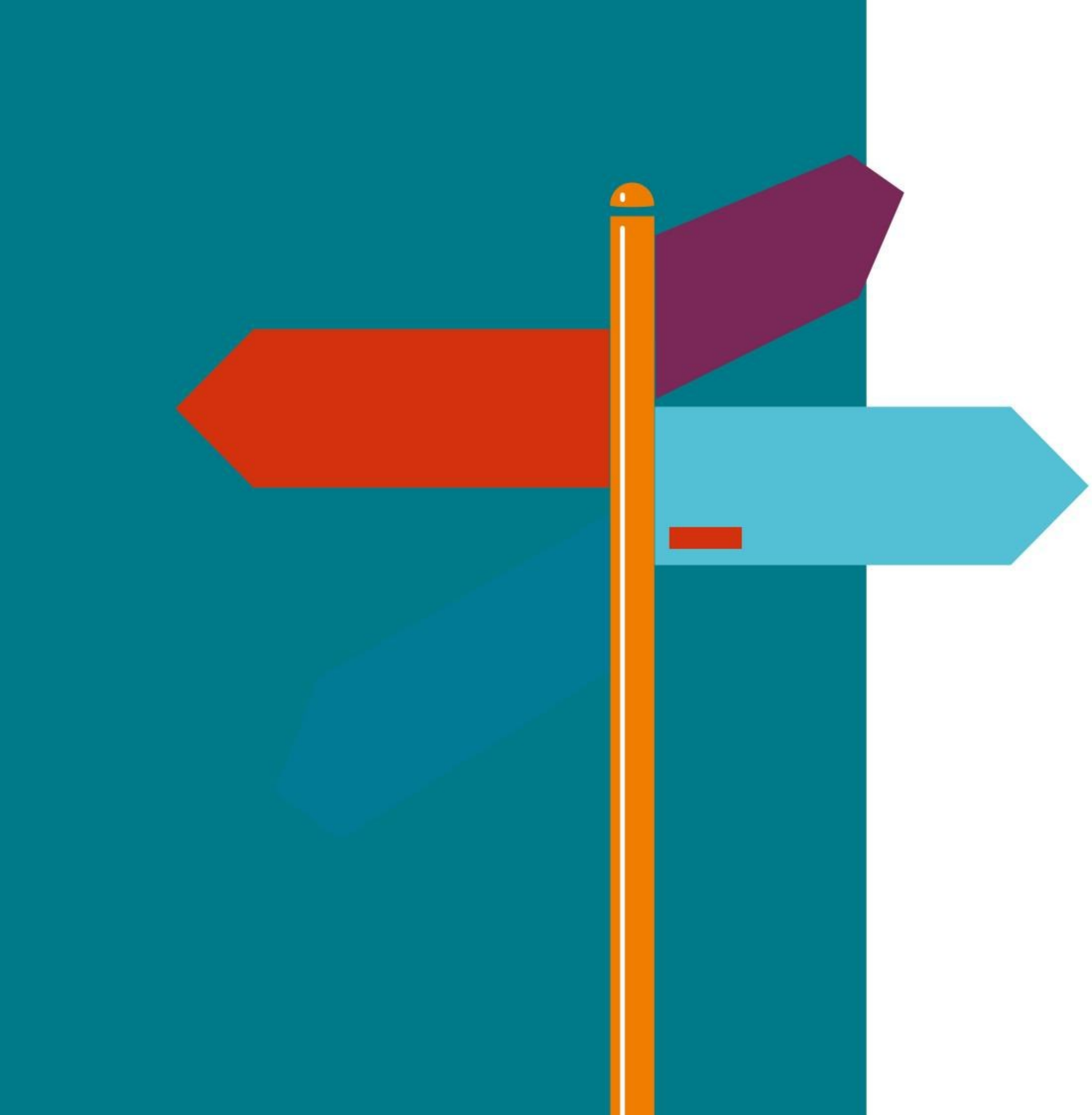
Add a document chunking function

with @LangChain

so all text files in the folder "/sources" get split into 256 character chunks and stored in a list

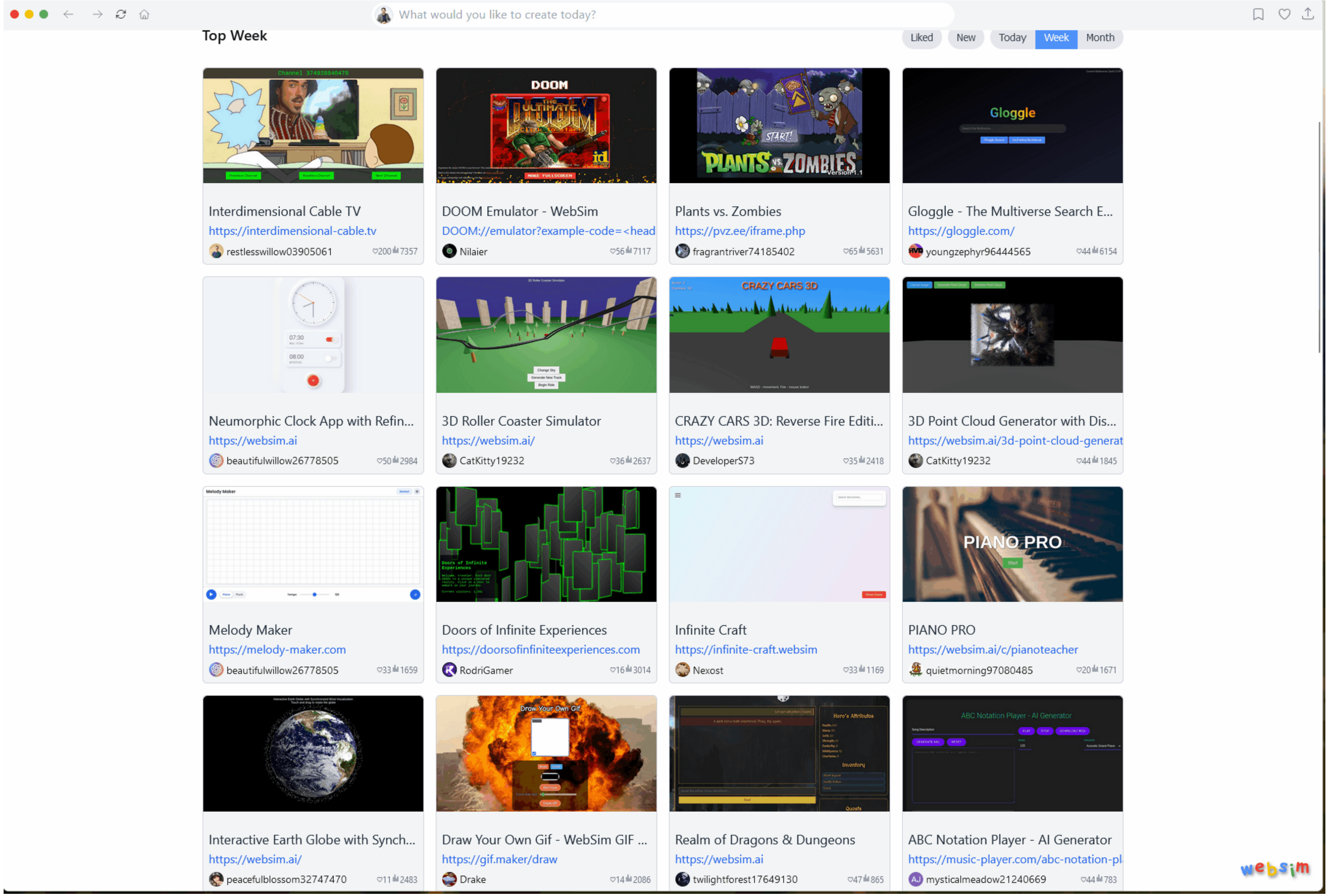
Coding with Copilots

- Context is King
- Instruct vs. Completion
- Good prompting practice



Trends

Promising Trends



Top Week

Interdimensional Cable TV
https://interdimensional-cable.tv
restlesswillow03905061 ♡200 7357

DOOM Emulator - WebSim
DOOM://emulator?example-code=<head
Nilaier ♡56 7117

Plants vs. Zombies
https://pvz.ee/iframe.php
fragrantriver74185402 ♡65 5631

Gloggle - The Multiverse Search E...
https://gloggle.com/
youngzephyr96444565 ♡44 6154

Neumorphic Clock App with Refin...
https://websim.ai
beautifulwillow26778505 ♡50 2984

3D Roller Coaster Simulator
https://websim.ai/
CatKitty19232 ♡36 2637

CRAZY CARS 3D: Reverse Fire Editi...
https://websim.ai
DeveloperS73 ♡35 2418

3D Point Cloud Generator with Dis...
https://websim.ai/3d-point-cloud-generat
CatKitty19232 ♡44 1845

Melody Maker
https://melody-maker.com
beautifulwillow26778505 ♡33 1659

Doors of Infinite Experiences
https://doorsofinfiniteexperiences.com
RodriGamer ♡16 3014

Infinite Craft
https://infinite-craft.websim
Nexost ♡33 1169

PIANO PRO
https://websim.ai/c/pianoteacher
quietmorning97080485 ♡20 1671

Interactive Earth Globe with Synch...
https://websim.ai/
peacefulblossom32747470 ♡11 2483

Draw Your Own Gif - WebSim GIF ...
https://gif.maker/draw
Drake ♡14 2086

Realm of Dragons & Dungeons
https://websim.ai
twilightforest17649130 ♡47 865

ABC Notation Player - AI Generator
https://music-player.com/abc-notation-pl
mysticalmeadow21240669 ♡44 783

Promising Trends

- Agents for Software Engineering
  - OpenDevin (Video)
  - Devika
- GenAI based Websites
  - https://websim.ai
- Test Generation
  - TestGen-LLM
- Compilation / Decompilation
  - Meta LLM Compiler

Q & A