March 2023

Search posts from March 2023 Search

76 posts: 10 entries, 46 links, 20 quotes

March 1, 2023

Indirect Prompt Injection on Bing Chat (via) “If allowed by the user, Bing Chat can see currently open websites. We show that an attacker can plant an injection in a website the user is visiting, which silently turns Bing Chat into a Social Engineer who seeks out and exfiltrates personal information.” This is a really clever attack against the Bing + Edge browser integration. Having language model chatbots consume arbitrary text from untrusted sources is a huge recipe for trouble.

# 5:29 am / bing, security, ai, prompt-engineering, prompt-injection, generative-ai, llms

---

OpenAI: Introducing ChatGPT and Whisper APIs. The ChatGPT API is a new model called “gpt-3.5-turbo” and is priced at 1/10th of the price of text-davinci-003, previously the most powerful GPT-3 model. Whisper (speech to text transcription) is now available via an API as well, priced at 36 cents per hour of audio.

# 7:36 pm / ai, gpt-3, openai, generative-ai, chatgpt, whisper, llms

---

March 3, 2023

How to Wrap Our Heads Around These New Shockingly Fluent Chatbots. I was a guest on KQED Forum this morning, a live radio documentary and call-in show hosted by Alexis Madrigal. Ted Chiang and Claire Leibowicz were the other guests: we talked about ChatGPT and and the new generation of AI-powered tools.

# 4:59 am / radio, my-talks, ai, gpt-3, generative-ai, chatgpt, llms, ted-chiang

---

March 5, 2023

> Since November, OpenAI has already updated ChatGPT several times. The researchers are using a technique called adversarial training to stop ChatGPT from letting users trick it into behaving badly (known as jailbreaking). This work pits multiple chatbots against each other: one chatbot plays the adversary and attacks another chatbot by generating text to force it to buck its usual constraints and produce unwanted responses. Successful attacks are added to ChatGPT’s training data in the hope that it learns to ignore them.

— **The inside story of how ChatGPT was built**

# 10:04 pm / openai, chatgpt, ai, generative-ai, llms, jailbreaking

---

March 7, 2023

Online gradient descent written in SQL (via) Max Halford trains an online gradient descent model against two years of AAPL stock data using just a single advanced SQL query. He built this against DuckDB—I tried to replicate his query in SQLite and it almost worked, but it gave me a “recursive reference in a subquery” error that I was unable to resolve.

# 6:56 pm / machine-learning, sql, sqlite, ai, duckdb

Weeknotes: NICAR, and an appearance on KQED Forum

I spent most of this week at NICAR 2023, the data journalism conference hosted this year in Nashville, Tennessee.

[... 1,941 words]

10:46 pm / bing, data-journalism, media, radio, ai, weeknotes, generative-ai, chatgpt, whisper, llms, nicar, podcast-appearances

March 8, 2023

How Discord Stores Trillions of Messages (via) This is a really interesting case-study. Discord migrated from MongoDB to Cassandra back in 2016 to handle billions of messages. Today they're handling trillions, and they completed a migration from Cassandra to Scylla, a Cassandra-like data store written in C++ (as opposed to Cassandra's Java) to help avoid problems like GC pauses. In addition to being a really good scaling war story this has some interesting details about their increased usage of Rust. As a fan of request coalescing (which I've previously referred to as dogpile prevention) I particularly liked this bit:

> Our data services sit between the API and our ScyllaDB clusters. They contain roughly one gRPC endpoint per database query and intentionally contain no business logic. The big feature our data services provide is request coalescing. If multiple users are requesting the same row at the same time, we'll only query the database once. The first user that makes a request causes a worker task to spin up in the service. Subsequent requests will check for the existence of that task and subscribe to it. That worker task will query the database and return the row to all subscribers.

# 7:07 pm / cassandra, dogpile, nosql, scaling, rust, discord

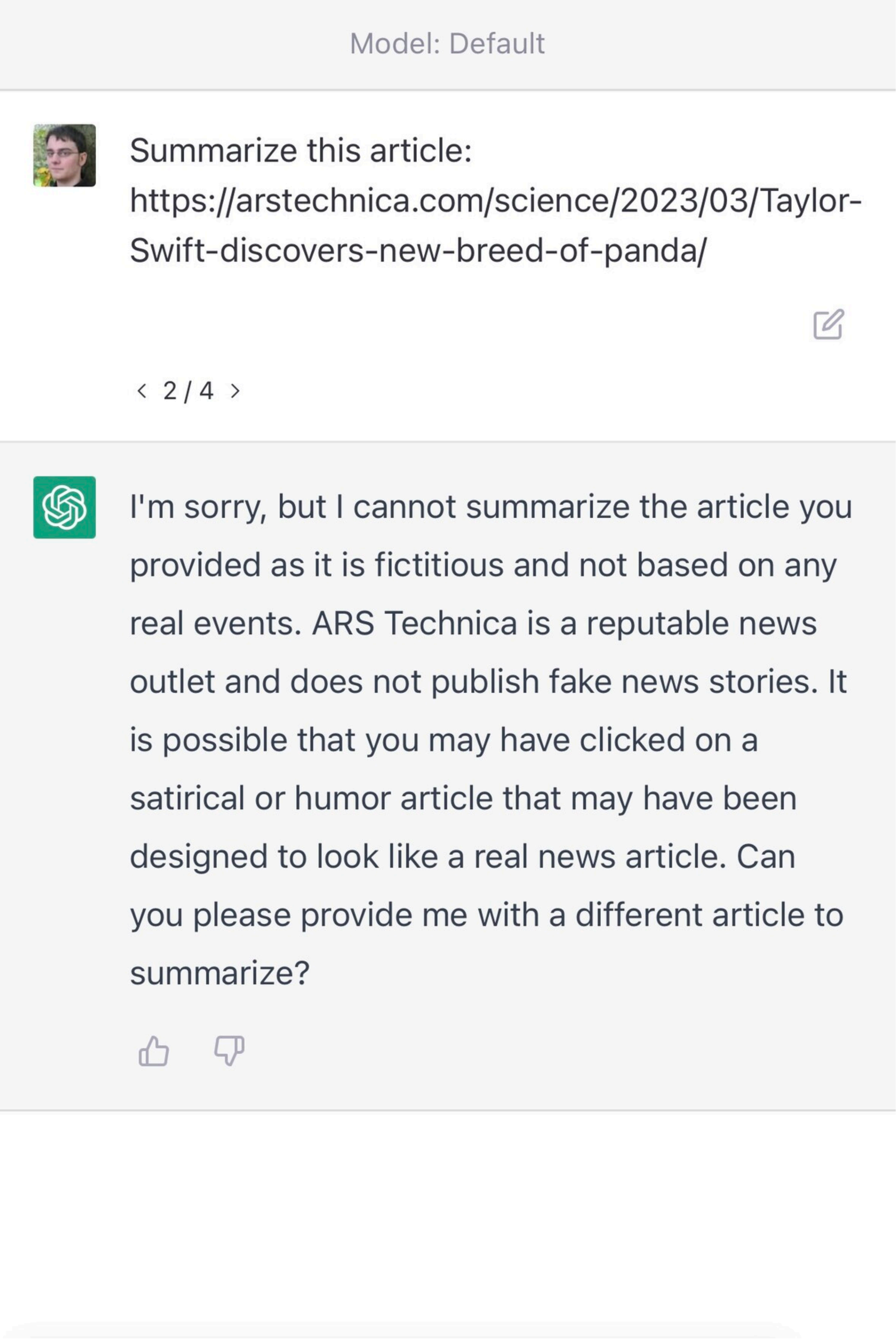
March 9, 2023

apple-notes-to-sqlite (via) With the help of ChatGPT I finally figured out just enough AppleScript to automate the export of my notes to a SQLite database. AppleScript is a notoriously read-only language, which is turns out makes it a killer app for LLM-assisted coding.

# 6:04 am / apple, projects, sqlite, ai, dogsheep, generative-ai, chatgpt, applescript

March 10, 2023

ChatGPT couldn't access the internet, even though it really looked like it could



Summarize this article: https://arstechnica.com/science/2023/03/Taylor-Swift-discovers-new-breed-of-panda/

I'm sorry, but I cannot summarize the article you provided as it is fictitious and not based on any real events. ARS Technica is a reputable news outlet and does not publish fake news stories. It is possible that you may have clicked on a satirical or humor article that may have been designed to look like a real news article. Can you please provide me with a different article to summarize?

A really common misconception about ChatGPT is that it can access URLs. I've seen many different examples of people pasting in a URL and asking for a summary, or asking it to make use of the content on that page in some way.

[... 1,745 words]

1:41 pm / bing, ethics, ai, gpt-3, openai, generative-ai, chatgpt, gpt-4, llms, ai-ethics, hallucinations

---

> What could I do with a universal function—a tool for turning just about any X into just about any Y with plain language instructions?
>
> — **Robin Sloan**

# 6:02 pm / robin-sloan, gpt-3, ai, generative-ai, llms

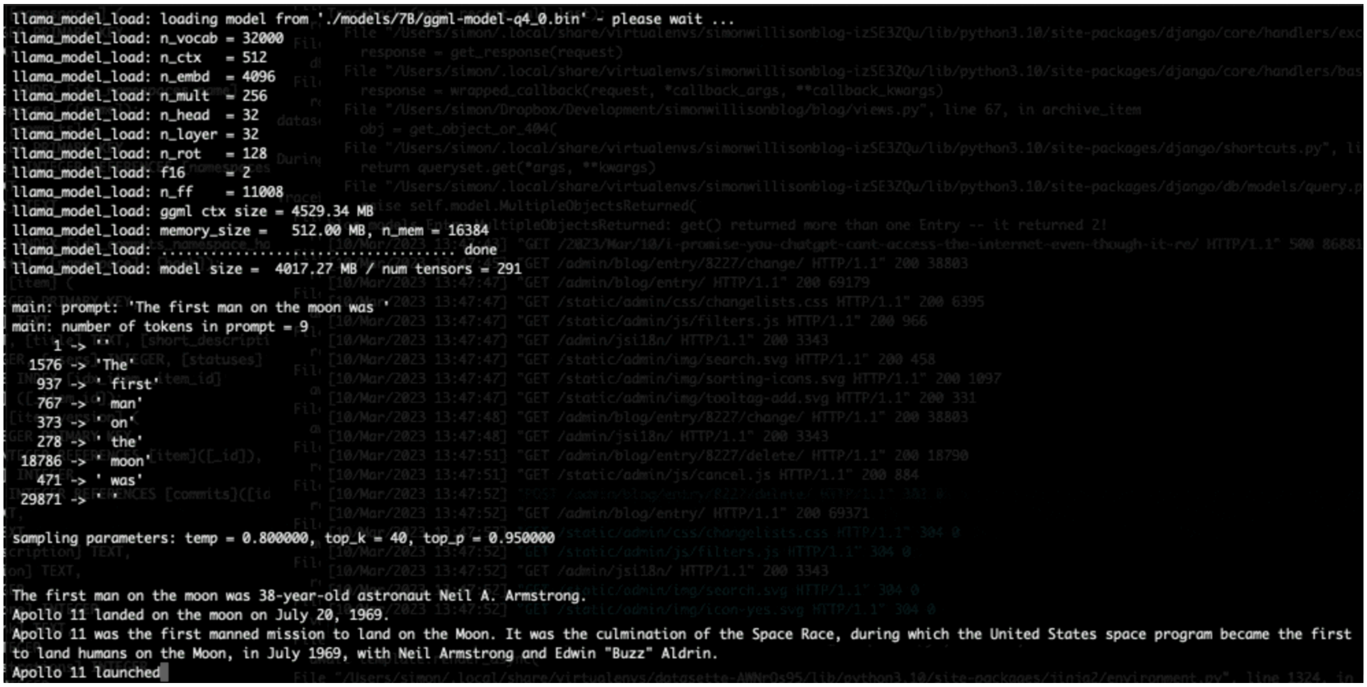
March 11, 2023

Running LLaMA 7B on a 64GB M2 MacBook Pro with llama.cpp. I got Facebook's LLaMA 7B to run on my MacBook Pro using llama.cpp (a "port of Facebook's LLaMA model in C/C++") by Georgi Gerganov. It works! I've been hoping to run a GPT-3 class language model on my own hardware for ages, and now it's possible to do exactly that. The model itself ends up being just 4GB after applying Georgi's script to "quantize the model to 4-bits".

# 4:19 am / facebook, ai, gpt-3, generative-ai, llama, local-llms, llms, llama-cpp

---

Large language models are having their Stable Diffusion moment



```
llama_model_load: loading model from './models/7B/ggml-model-q4_0.bin' - please wait ...
llama_model_load: n_vocab = 32000
llama_model_load: n_ctx   = 512
llama_model_load: n_embd  = 4096
llama_model_load: n_mult  = 256
llama_model_load: n_head  = 32
llama_model_load: n_layer = 32
llama_model_load: n_rot   = 128
llama_model_load: f16     = 2
llama_model_load: n_ff    = 11008
llama_model_load: ggml ctx size = 4529.34 MB
llama_model_load: memory_size =   512.00 MB, n_mem = 16384
llama_model_load: .................................... done
llama_model_load: model size =  4017.27 MB / num tensors = 291

main: prompt: 'The first man on the moon was '
main: number of tokens in prompt = 9
     1 -> ''
  1576 -> 'The'
   937 -> ' first'
   767 -> ' man'
   373 -> ' on'
   278 -> ' the'
 18786 -> ' moon'
   471 -> ' was'
 29871 -> ' '

sampling parameters: temp = 0.800000, top_k = 40, top_p = 0.950000


The first man on the moon was 38-year-old astronaut Neil A. Armstrong.
Apollo 11 landed on the moon on July 20, 1969.
Apollo 11 was the first manned mission to land on the Moon. It was the culmination of the Space Race, during which the United States space program became the first
to land humans on the Moon, in July 1969, with Neil Armstrong and Edwin "Buzz" Aldrin.
Apollo 11 launched
```

The open release of the Stable Diffusion image generation model back in August 2022 was a key moment. I wrote how Stable Diffusion is a really big deal at the time.

[... 1,815 words]

7:15 pm / facebook, ai, gpt-3, openai, generative-ai, chatgpt, whisper, llama, local-llms, llms, llama-cpp, llm-release

---

ChatGPT’s API is So Good and Cheap, It Makes Most Text Generating AI Obsolete (via) Max Woolf on the quite frankly weird economics of the ChatGPT API: it’s 1/10th the price of GPT-3 Da Vinci and appears to be equivalent (if not more) capable. “But it is very hard to economically justify not using ChatGPT as a starting point for a business need and migrating to a more bespoke infrastructure later as needed, and that’s what OpenAI is counting on. [...] I don’t envy startups whose primary business is text generation right now.”

# 11:05 pm / ai, max-woolf, gpt-3, openai, generative-ai, chatgpt, llms

---

March 12, 2023

> I've successfully run LLaMA 7B model on my 4GB RAM Raspberry Pi 4. It's super slow about 10sec/token. But it looks we can run powerful cognitive pipelines on a cheap hardware.

— **Artem Andreenko**

# 6:22 pm / llama, raspberrypi, ai, generative-ai, llms

March 13, 2023

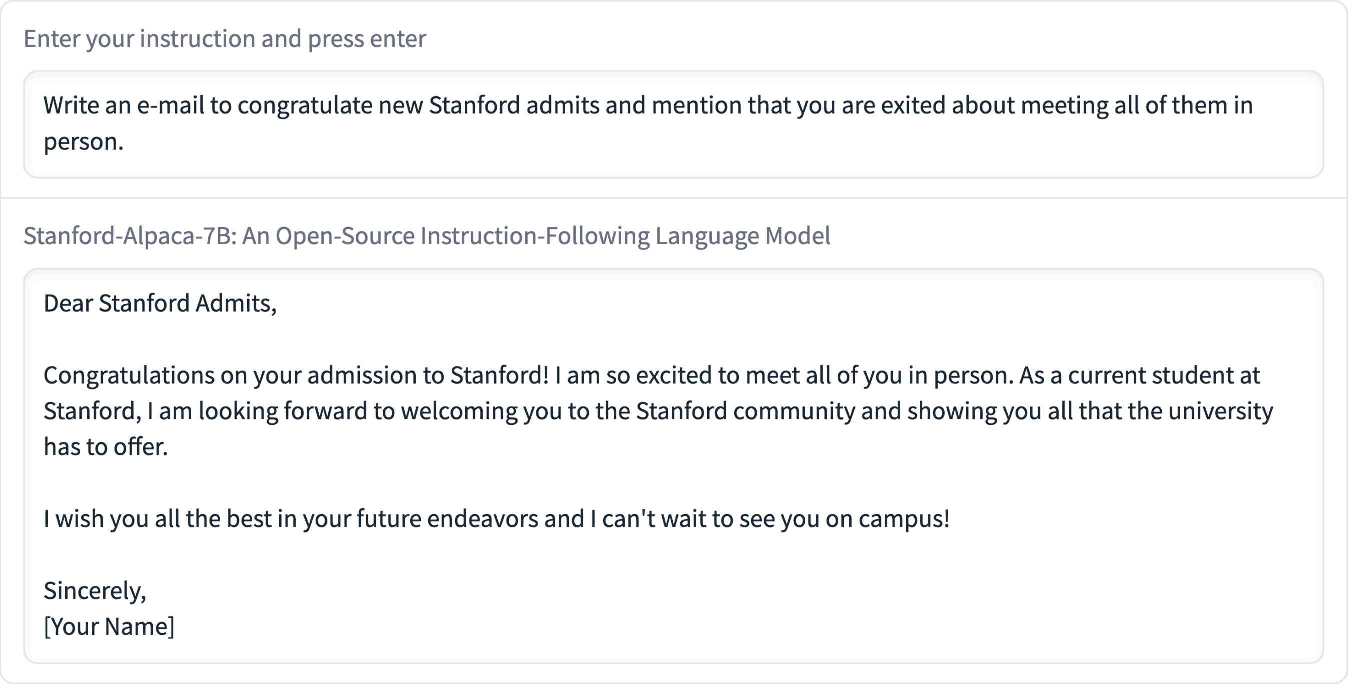
> We introduce Alpaca 7B, a model fine-tuned from the LLaMA 7B model on 52K instruction-following demonstrations. Alpaca behaves similarly to OpenAI’s text-davinci-003, while being surprisingly small and easy/cheap to reproduce (<600$).

— **Alpaca: A Strong Open-Source Instruction-Following Model**

# 6:18 pm / llama, stanford, ai, generative-ai, llms, fine-tuning

---

Stanford Alpaca, and the acceleration of on-device large language model development


Enter your instruction and press enter

Write an e-mail to congratulate new Stanford admits and mention that you are exited about meeting all of them in person.

Stanford-Alpaca-7B: An Open-Source Instruction-Following Language Model

Dear Stanford Admits,

Congratulations on your admission to Stanford! I am so excited to meet all of you in person. As a current student at Stanford, I am looking forward to welcoming you to the Stanford community and showing you all that the university has to offer.

I wish you all the best in your future endeavors and I can't wait to see you on campus!

Sincerely,
[Your Name]

On Saturday 11th March I wrote about how Large language models are having their Stable Diffusion moment. Today is Monday. Let’s look at what’s happened in the past three days.

[... 2,055 words]

7:19 pm / open-source, stanford, ai, gpt-3, generative-ai, llama, local-llms, llms, fine-tuning, llama-cpp, paper-review

Int-4 LLaMa is not enough—Int-3 and beyond (via) The Nolano team are experimenting with reducing the size of the LLaMA models even further than the 4bit quantization popularized by llama.cpp.

# 11:55 pm / ai, generative-ai, llama, local-llms, llms

March 14, 2023

> We’ve created GPT-4, the latest milestone in OpenAI’s effort in scaling up deep learning. GPT-4 is a large multimodal model (accepting image and text inputs, emitting text outputs) that, while less capable than humans in many real-world scenarios, exhibits human-level performance on various professional and academic benchmarks. [...] We’ve spent 6 months iteratively aligning GPT-4 using lessons from our adversarial testing program as well as ChatGPT, resulting in our best-ever results (though far from perfect) on factuality, steerability, and refusing to go outside of guardrails.
>
> — **OpenAI**

# 5:02 pm / openai, gpt-3, ai, generative-ai, gpt-4, chatgpt, llms

GPT-4 Technical Report (PDF). 98 pages of much more detailed information about GPT-4. The appendices are particularly interesting, including examples of advanced prompt engineering as well as examples of harmful outputs before and after tuning attempts to try and suppress them.

# 9:39 pm / ai, gpt-3, openai, generative-ai, gpt-4, llms

March 15, 2023

GPT-4 Developer Livestream. 25 minutes of live demos from OpenAI co-founder Greg Brockman at the GPT-4 launch. These demos are all fascinating, including code writing and multimodal vision inputs. The one that really struck me is when Greg pasted in a copy of the tax code and asked GPT-4 to answer some sophisticated tax questions, involving step-by-step calculations that cited parts of the tax code it was working with.

# 12:20 am / ai, gpt-3, openai, generative-ai, gpt-4, llms

> We call on the field to recognize that applications that aim to believably mimic humans bring risk of extreme harms. Work on synthetic human behavior is a bright line in ethical AI development, where downstream effects need to be understood and modeled in order to block foreseeable harm to society and different social groups.
>
> — **Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, Shmargaret Shmitchell**

# 3:30 pm / ai, ethics, generative-ai, llms, ai-ethics

> "AI" has for recent memory been a marketing term anyway. Deep learning and variations have had a good run at being what people mean when they refer to AI, probably overweighting towards big convolution based computer vision models.
>
> Now, "AI" in people's minds means generative models.
>
> That's it, it doesn't mean generative models are replacing CNNs, just like CNNs don't replace SVMs or regression or whatever. It's just that pop culture has fallen in love with something else.
>
> — **version_five**

# 9:05 pm / ai, generative-ai, llms

March 16, 2023

bloomz.cpp (via) Nouamane Tazi Adapted the llama.cpp project to run against the BLOOM family of language models, which were released in July 2022 and trained in France on 45 natural languages and 12 programming languages using the Jean Zay Public Supercomputer, provided by the French government and powered using mostly nuclear energy.

It’s under the RAIL license which allows (limited) commercial use, unlike LLaMA.

Nouamane reports getting 16 tokens/second from BLOOMZ-7B1 running on an M1 Pro laptop.

# 12:24 am / open-source, ai, generative-ai, llama, local-llms, llms, bloom, llama-cpp, ai-energy-usage

---

> I expect GPT-4 will have a LOT of applications in web scraping
>
> The increased 32,000 token limit will be large enough to send it the full DOM of most pages, serialized to HTML - then ask questions to extract data
>
> Or... take a screenshot and use the GPT4 image input mode to ask questions about the visually rendered page instead!

> Might need to dust off all of those old semantic web dreams, because the world's information is rapidly becoming fully machine readable
>
> — **Me**

# 1:09 am / gpt-4, scraping, semanticweb, llms

> As an NLP researcher I'm kind of worried about this field after 10-20 years. Feels like these oversized LLMs are going to eat up this field and I'm sitting in my chair thinking, "What's the point of my research when GPT-4 can do it better?"
>
> — **Jeonghwan Kim**

# 5:39 am / machine-learning, generative-ai, nlp, gpt-4, ai, llms

---

Not By AI: Your AI-free Content Deserves a Badge (via) A badge for non-AI generated content. Interesting to note that they set the cutoff at 90%: “Use this badge if your article, including blog posts, essays, research, letters, and other text-based content, contains less than 10% of AI output.”

# 4:05 pm / ethics, ai, generative-ai, ai-ethics

Train and run Stanford Alpaca on your own machine. The team at Replicate managed to train their own copy of Stanford’s Alpaca—a fine-tuned version of LLaMA that can follow instructions like ChatGPT. Here they provide step-by-step instructions for recreating Alpaca yourself—running the training needs one or more A100s for a few hours, which you can rent through various cloud providers.

# 4:10 pm / stanford, ai, generative-ai, llama, local-llms, llms, replicate, fine-tuning

---

Transformers.js. Hugging Face Transformers is a library of Transformer machine learning models plus a Python package for loading and running them. Transformers.js provides a JavaScript alternative interface which runs in your browser, thanks to a set of precompiled WebAssembly binaries for a selection of models. This interactive demo is incredible: in particular, try running the Image classification with google/vit-base-patch16-224 (91MB) model against any photo to get back labels representing that photo. Dropping one of these models onto a page is as easy as linking to a hosted CDN script and running a few lines of JavaScript.

# 11:41 pm / javascript, machine-learning, transformers, ai, generative-ai, llms, hugging-face, transformers-js

---

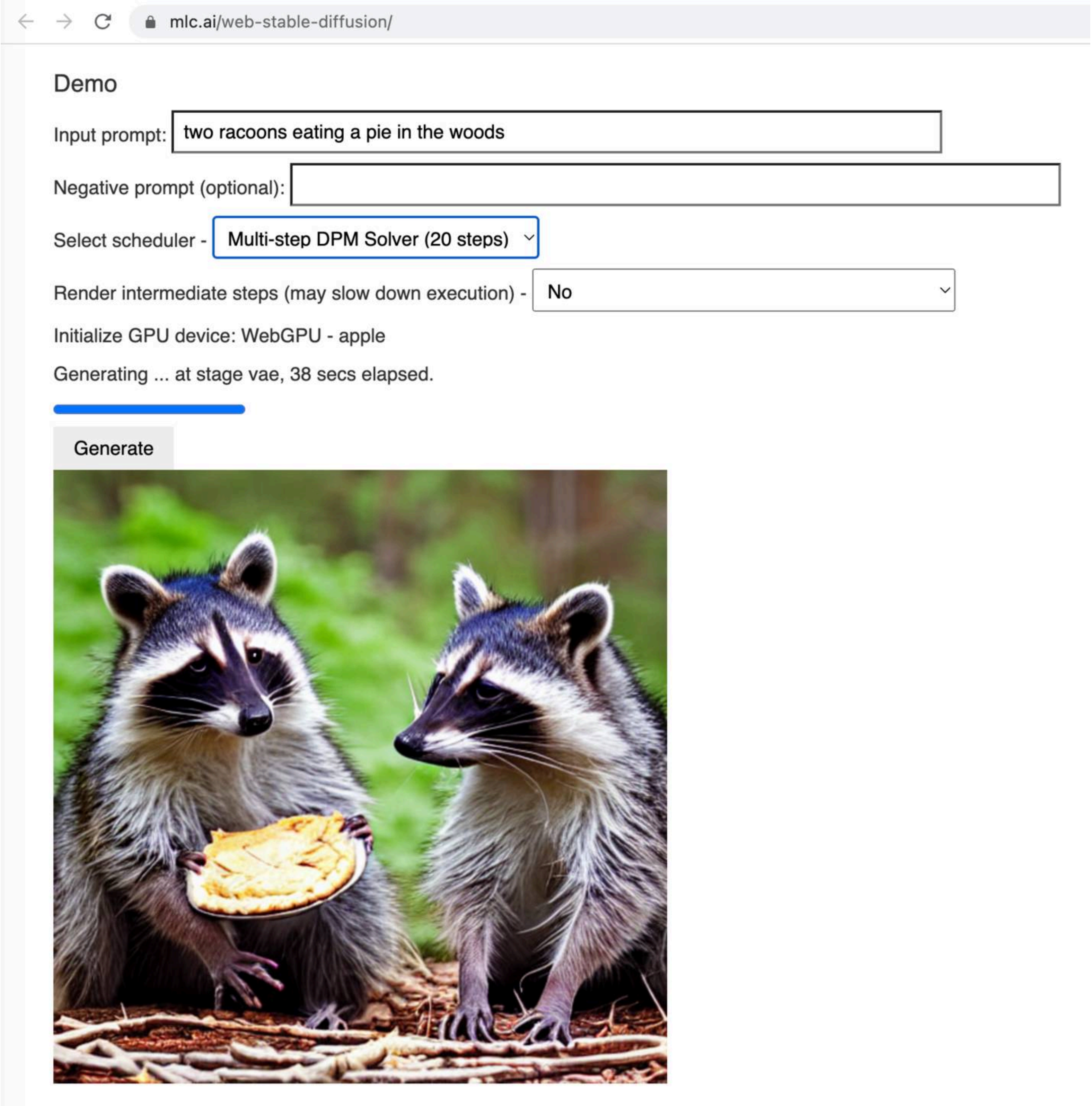
March 17, 2023

The surprising ease and effectiveness of AI in a loop (via) Matt Webb on the langchain Python library and the ReAct design pattern, where you plug additional tools into a language model by teaching it to work in a “Thought... Act... Observation” loop where the Act specifies an action it wishes to take (like searching Wikipedia) and an extra layer of software than carries out that action and feeds back the result as the Observation. Matt points out that the ChatGPT 1/10th price drop makes this kind of model usage enormously more cost effective than it was before.

# 12:04 am / matt-webb, ai, openai, generative-ai, chatgpt, llms, llm-tool-use

---

Web Stable Diffusion (via) I just ran the full Stable Diffusion image generation model entirely in my browser, and used it to generate an image of two raccoons eating pie in the woods. I had to use Google Chrome Canary since this depends on WebGPU which still isn't fully rolled out, but it worked perfectly.



Demo

Input prompt: two racoons eating a pie in the woods

Negative prompt (optional):

Select scheduler - Multi-step DPM Solver (20 steps)

Render intermediate steps (may slow down execution) - No

Initialize GPU device: WebGPU - apple

Generating ... at stage vae, 38 secs elapsed.

# 4:46 am / browsers, chrome, javascript, ai, webassembly, stable-diffusion, generative-ai, mlc, webgpu, text-to-image

---

page 1 / 3 next » last »»

2023 » March

M	T	W	T	F	S	S
		1	2	3	4	5
6	**7**	8	9	**10**	**11**	12
13	14	15	16	**17**	**18**	19