April 2024

Search posts from April 2024 Search

90 posts: 5 entries, 59 links, 26 quotes

April 1, 2024

OpenAI: Start using ChatGPT instantly. ChatGPT no longer requires signing in with an account in order to use the GPT-3.5 version, at least in some markets. I can access the service without login in an incognito browser window here in California.

The login-free free version includes "additional content safeguards for this experience, such as blocking prompts and generations in a wider range of categories", with no more details provided as to what that means.

Interestingly, even logged out free users get the option (off by default) to opt-out of having their conversations used to "improve our models for everyone".

OpenAI say that this initiative is to support "the aim to make AI accessible to anyone curious about its capabilities." This makes sense to me: there are still a huge number of people who haven't tried any of the LLM chat tools due to the friction of creating an account.

7:31 pm / ai, openai, generative-ai, chatgpt, llms

Diving Deeper into AI Package Hallucinations. Bar Lanyado noticed that LLMs frequently hallucinate the names of packages that don't exist in their answers to coding questions, which can be exploited as a supply chain attack.

He gathered 2,500 questions across Python, Node.js, Go, .NET and Ruby and ran them through a number of different LLMs, taking notes of any hallucinated packages and if any of those hallucinations were repeated.

One repeat example was "pip install huggingface-cli" (the correct package is "huggingface[cli]"). Bar then published a harmless package under that name in January, and observebd 30,000 downloads of that package in the three months that followed.

10:51 pm / security, ai, generative-ai, llms, supply-chain, hallucinations

---

PEP 738 – Adding Android as a supported platform (via) The BeeWare project got PEP 730—Adding iOS as a supported platform—accepted by the Python Steering Council in December, now it's Android's turn. Both iOS and Android will be supported platforms for CPython 3.13.

It's been possible to run custom compiled Python builds on those platforms for years, but official support means that they'll be included in Python's own CI and release process.

11:57 pm / android, python, ios, beeware

April 2, 2024

> LLMs are like a trained circus bear that can make you porridge in your kitchen. It's a miracle that it's able to do it at all, but watch out because no matter how well they can act like a human on some tasks, they're still a wild animal. They might ransack your kitchen, and they could kill you, accidentally or intentionally!

— **Alex Komoroske**

3:19 pm / llms, ai, generative-ai, alex-komoroske

Bringing Python to Workers using Pyodide and WebAssembly (via) Cloudflare Workers is Cloudflare's serverless hosting tool for deploying server-side functions to edge locations in their CDN.

They just released Python support, accompanied by an extremely thorough technical explanation of how they got that to work. The details are fascinating.

Workers runs on V8 isolates, and the new Python support was implemented using Pyodide (CPython compiled to WebAssembly) running inside V8.

Getting this to work performantly and ergonomically took a huge amount of work.

There are too many details in here to effectively summarize, but my favorite detail is this one:

"We scan the Worker's code for import statements, execute them, and then take a snapshot of the Worker's WebAssembly linear memory. Effectively, we perform the expensive work of importing packages at deploy time, rather than at runtime."

4:09 pm / python, serverless, cloudflare, webassembly, pyodide

---

Cally: Accessibility statement (via) Cally is a neat new open source date (and date range) picker Web Component by Nick Williams.

It's framework agnostic and weighs less than 9KB grilled, but the best feature is this detailed page of documentation covering its accessibility story, including how it was tested—in JAWS, NVDA and VoiceOver.

I'd love to see other open source JavaScript libraries follow this example.

7:38 pm / accessibility, javascript, open-source, web-components

---

April 3, 2024

Enforcing conventions in Django projects with introspection (via) Luke Plant shows how to use the Django system checks framework to introspect models on startup and warn if a DateTime or Date model field has been added that doesn't conform to a specific naming convention.

Luke also proposes "*_at" as a convention for DateTimes, contrasting with "*_on" or "*_date" (I prefer the latter) for Dates.

2:58 pm / django, luke-plant

April 4, 2024

Kobold letters (via) Konstantin Weddige explains a sophisticated HTML email phishing vector he calls Kobold emails.

When you forward a message, most HTML email clients will indent the forward by nesting it inside another element.

This means CSS rules within the email can be used to cause an element that was invisible in the original email to become visible when it is forwarded—allowing tricks like a forwarded innocuous email from your boss adding instructions for wiring money from the company bank account.

Gmail strips style blocks before forwarding—which it turns out isn't protection against this, because you can put a style block in the original email to hide the attack text which will then be stripped for you when the email is forwarded.

12:43 pm / css, email, security

The cost of AI reasoning over time (via) Karina Nguyen from Anthropic provides a fascinating visualization illustrating the cost of different levels of LLM over the past few years, plotting their cost-per-token against their scores on the MMLU benchmark.

Claude 3 Haiku currently occupies the lowest cost to score ratio, over on the lower right hand side of the chart.

12:51 pm / ai, generative-ai, llms, anthropic, claude

---

llm-command-r. Cohere released Command R Plus today—an open weights (non commercial/research only) 104 billion parameter LLM, a big step up from their previous 35 billion Command R model.

Both models are fine-tuned for both tool use and RAG. The commercial API has features to expose this functionality, including a web-search connector which lets the model run web searches as part of answering the prompt and return documents and citations as part of the JSON response.

I released a new plugin for my LLM command line tool this morning adding support for the Command R models.

In addition to the two models it also adds a custom command for running prompts with web search enabled and listing the referenced documents.

5:38 pm / plugins, projects, ai, generative-ai, llms, llm, cohere, command-r, rag, llm-tool-use, llm-release

---

> Before Google Reader was shut down, they were internally looking for maintainers. It turned out you have to deal with three years of infra migrations if you sign up to be the new owner of Reader. No one wanted that kind of job for a product that is not likely to grow 10x.
>
> — **Jaana Dogan**

8:51 pm / google, google-reader

April 5, 2024

s3-credentials 0.16. I spent entirely too long this evening trying to figure out why files in my new supposedly public S3 bucket were unavailable to view. It turns out these days you need to set a `PublicAccessBlockConfiguration` of `{"BlockPublicAcls": false, "IgnorePublicAcls": false, "BlockPublicPolicy": false, "RestrictPublicBuckets": false}`.

The `s3-credentials --create-bucket --public` option now does that for you. I also added a `s3-credentials debug-bucket name-of-bucket` command to help figure out why a bucket isn't working as expected.

5:35 am / aws, projects, s3, s3-credentials

Everything I Know About the XZ Backdoor (via) Evan Boehs provides the most detailed timeline I've seen of the recent xz story, where a backdoor was inserted into the xz compression library in an attempt to compromise OpenSSH.

10:58 pm / security

April 6, 2024

datasette-import. A new plugin for importing data into Datasette. This is a replacement for datasette-paste, duplicating and extending its functionality. datasette-paste had grown beyond just dealing with pasted CSV/TSV/JSON data—it handles file uploads as well now—which inspired the new name.

10:40 pm / plugins, projects, datasette

April 7, 2024

The lifecycle of a code AI completion (via) Philipp Spiess provides a deep dive into how Sourcegraph's Cody code completion assistant works. Lots of fascinating details in here:

"One interesting learning was that if a user is willing to wait longer for a multi-line request, it usually is worth it to increase latency slightly in favor of quality. For our production setup this means we use a more complex language model for multi-line completions than we do for single-line completions."

This article is from October 2023 and talks about Claude Instant. The code for Cody is open source so I checked to see if they have switched to Haiku yet and found a commit from March 25th that adds Haiku as an A/B test.

7:37 pm / ai, generative-ai, llms, ai-assisted-programming, anthropic, claude

---

April 8, 2024

> in July 2023, we [Hugging Face] wanted to experiment with a custom license for this specific project [text-generation-inference] in order to protect our commercial solutions from companies with bigger means than we do, who would just host an exact copy of our cloud services.
>
> The experiment however wasn't successful.
>
> It did not lead to licensing-specific incremental business opportunities by itself, while it did hamper or at least complicate the community contributions, given the legal uncertainty that arises as soon as you deviate from the standard licenses.
>
> — **Julien Chaumond**

6:35 pm / open-source, hugging-face

Introducing Enhance WASM (via) “Backend agnostic server-side rendering (SSR) for Web Components”—fascinating new project from Brian LeRoux and Begin.

The idea here is to provide server-side rendering of Web Components using WebAssembly that can run on any platform that is supported within the Extism WASM ecosystem.

The key is the enhance-ssr.wasm bundle, a 4.1MB WebAssembly version of the enhance-ssr JavaScript library, compiled using the Extism JavaScript PDK (Plugin Development Kit) which itself bundles a WebAssembly version of QuickJS.

7:44 pm / javascript, web-components, webassembly, quickjs

Building files-to-prompt entirely using Claude 3 Opus

```
files-to-prompt files_to_prompt tests | llm -m opus --system '
rewrite the tests to cover the ability to pass multiple files and
folders to the tool'

files-to-prompt files_to_prompt tests | llm -m opus --system '
add one last test which tests .gitignore and include_hidden against
an example that mixes single files and directories of files together
in one invocation'
```

files-to-prompt is a new tool I built to help me pipe several files at once into prompts to LLMs such as Claude and GPT-4.

[... 3,235 words]

8:40 pm / cli, projects, ai, prompt-engineering, generative-ai, llms, ai-assisted-programming, llm, anthropic, claude, files-to-prompt

---

April 9, 2024

Hello World (via) Lennon McLean dives deep down the rabbit hole of what happens when you execute the binary compiled from “Hello world” in C on a Linux system, digging into the details of ELF executables, objdump disassembly, the C standard library, stack frames, null-terminated strings and taking a detour through musl because it’s easier to read than Glibc.

1:06 am / c, linux

---

llm.c (via) Andrej Karpathy implements LLM training—initially for GPT-2, other architectures to follow—in just over 1,000 lines of C on top of CUDA. Includes a tutorial about implementing LayerNorm by porting an implementation from Python.

3:24 pm / c, ai, andrej-karpathy, generative-ai, llms, gpt-2

Command R+ now ranked 6th on the LMSYS Chatbot Arena. The LMSYS Chatbot Arena Leaderboard is one of the most interesting approaches to evaluating LLMs because it captures their ever-elusive “vibes”—it works by users voting on the best responses to prompts from two initially hidden models

Big news today is that Command R+—the brand new open weights model (Creative Commons non-commercial) by Cohere—is now the highest ranked non-proprietary model, in at position six and beating one of the GPT-4s.

(Linking to my screenshot on Mastodon.)

4:19 pm / ai, generative-ai, llms, cohere, command-r, chatbot-arena

---

A solid pattern to build LLM Applications (feat. Claude) (via) Hrishi Olickel is one of my favourite prompt whisperers. In this YouTube video he walks through his process for building quick interactive applications with the assistance of Claude 3, spinning up an app that analyzes his meeting transcripts to extract participants and mentioned organisations, then presents a UI for exploring the results built with Next.js and shadcn/ui.

An interesting tip I got from this: use the weakest, not the strongest models to iterate on your prompts. If you figure out patterns that work well with Claude 3 Haiku they will have a significantly lower error rate with Sonnet or Opus. The speed of

the weaker models also means you can iterate much faster, and worry less about the cost of your experiments.

6:39 pm / ai, generative-ai, llms, ai-assisted-programming, claude

---

Extracting data from unstructured text and images with Datasette and GPT-4 Turbo. Datasette Extract is a new Datasette plugin that uses GPT-4 Turbo (released to general availability today) and GPT-4 Vision to extract structured data from unstructured text and images.

I put together a video demo of the plugin in action today, and posted it to the Datasette Cloud blog along with screenshots and a tutorial describing how to use it.

11:03 pm / projects, ai, datasette, datasette-cloud, openai, generative-ai, gpt-4, llms, vision-llms, structured-extraction

---

April 10, 2024

Mistral tweet a magnet link for mixtral-8x22b. Another open model release from Mistral using their now standard operating procedure of tweeting out a raw torrent link.

This one is an 8x22B Mixture of Experts model. Their previous most powerful openly licensed release was Mixtral 8x7B, so this one is a whole lot bigger (a 281GB download)—and apparently has a 65,536 context length, at least according to initial rumors on Twitter.

2:31 am / ai, generative-ai, local-llms, llms, mistral, llm-release

---

Gemini 1.5 Pro public preview (via) Huge release from Google: Gemini 1.5 Pro—the GPT-4 competitive model with the incredible 1 million token context length—is now available without a waitlist in 180+ countries (including the USA but not Europe or the UK as far as I can tell)... and the API is free for 50 requests/day (rate limited to 2/minute).

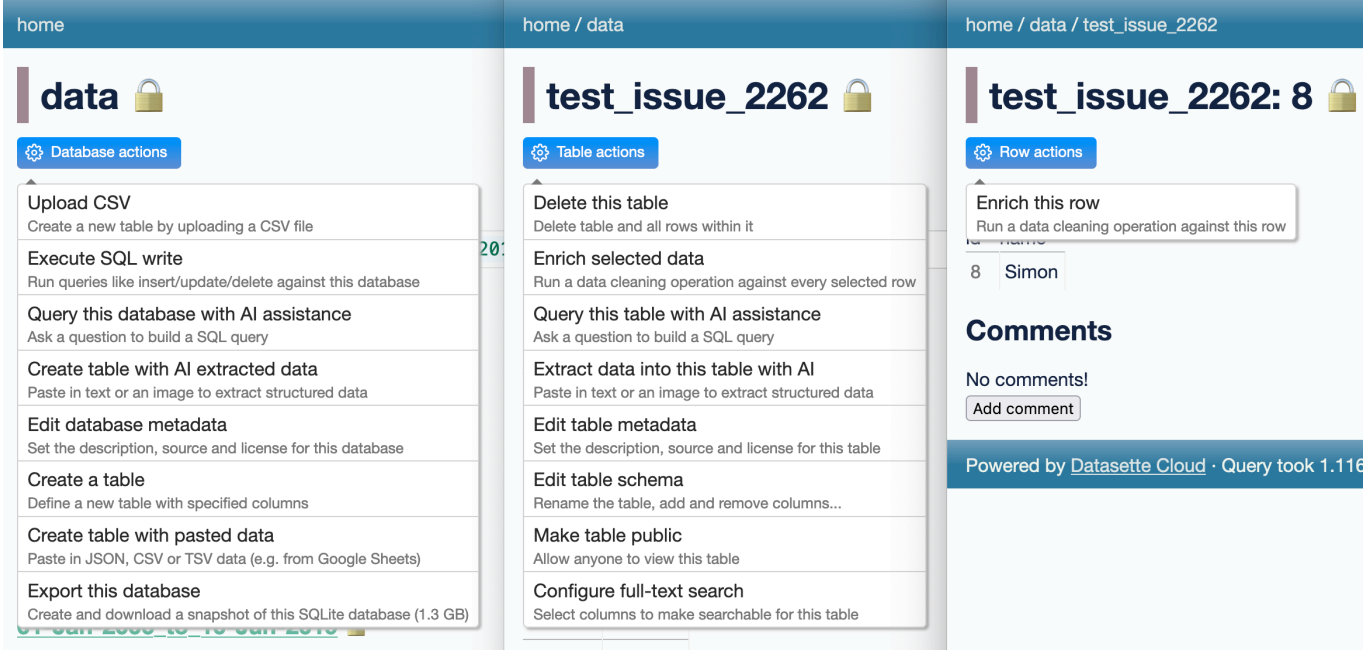
Beyond that you'll need to pay—$7/million input tokens and $21/million output tokens, which is slightly less than GPT-4 Turbo and a little more than Claude 3 Sonnet.

They also announced audio input (up to 9.5 hours in a single prompt), system instruction support and a new JSON mode.

2:38 am / google, ai, generative-ai, llms, gemini, vision-llms, llm-pricing, llm-release

---

Three major LLM releases in 24 hours (plus weeknotes)



I'm a bit behind on my weeknotes, so there's a lot to cover here. But first... a review of the last 24 hours of Large Language Model news. All times are in US Pacific on April 9th 2024.

[... 1,401 words]

5:09 am / projects, ai, weeknotes, datasette-cloud, openai, generative-ai, llms, gemini, llm-release

---

> The challenge [with RAG] is that most corner-cutting solutions look like they're working on small datasets while letting you pretend that things like search relevance don't matter, while in reality relevance significantly impacts quality of responses when you move beyond prototyping (whether they're literally search relevance or are better tuned SQL queries to retrieve more appropriate rows). This creates a false expectation of how the prototype will translate into a production capability, with all the predictable consequences: underestimating timelines, poor production behavior/performance, etc.
>
> — **Will Larson**

11:09 pm / generative-ai, will-larson, search, ai, llms, rag

Notes on how to use LLMs in your product. A whole bunch of useful observations from Will Larson here. I love his focus on the key characteristic of LLMs that "you cannot know whether a given response is accurate", nor can you calculate a dependable confidence score for a response—and as a result you need to either "accept potential inaccuracies (which makes sense in many cases, humans are wrong sometimes too) or keep a Human-in-the-Loop (HITL) to validate the response."

11:14 pm / will-larson, ai, generative-ai, llms

Shell History Is Your Best Productivity Tool (via) Martin Heinz drops a wealth of knowledge about ways to configure zsh (the default shell on macOS these days) to get better utility from your shell history.

11:17 pm / shell, zsh

April 11, 2024

[on GitHub Copilot] It's like insisting to walk when you can take a bike. It gets the hard things wrong but all the easy things right, very helpful and much faster. You have to learn what it can and can't do.

— **Andrej Karpathy**

1:27 am / andrej-karpathy, ai-assisted-programming, generative-ai, ai, llms, github-copilot

---

2024 » April

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	20	21
22	**23**	24	25	26	27	28
29	30					