May 2023

Search posts from May 2023 Search

70 posts: 12 entries, 44 links, 14 quotes

May 1, 2023

Weeknotes: Miscellaneous research into Rye, ChatGPT Code Interpreter and openai-to-sqlite

I gave myself some time off stressing about my core responsibilities this week after PyCon, which meant allowing myself to be distracted by some miscellaneous research projects.

[... 891 words]

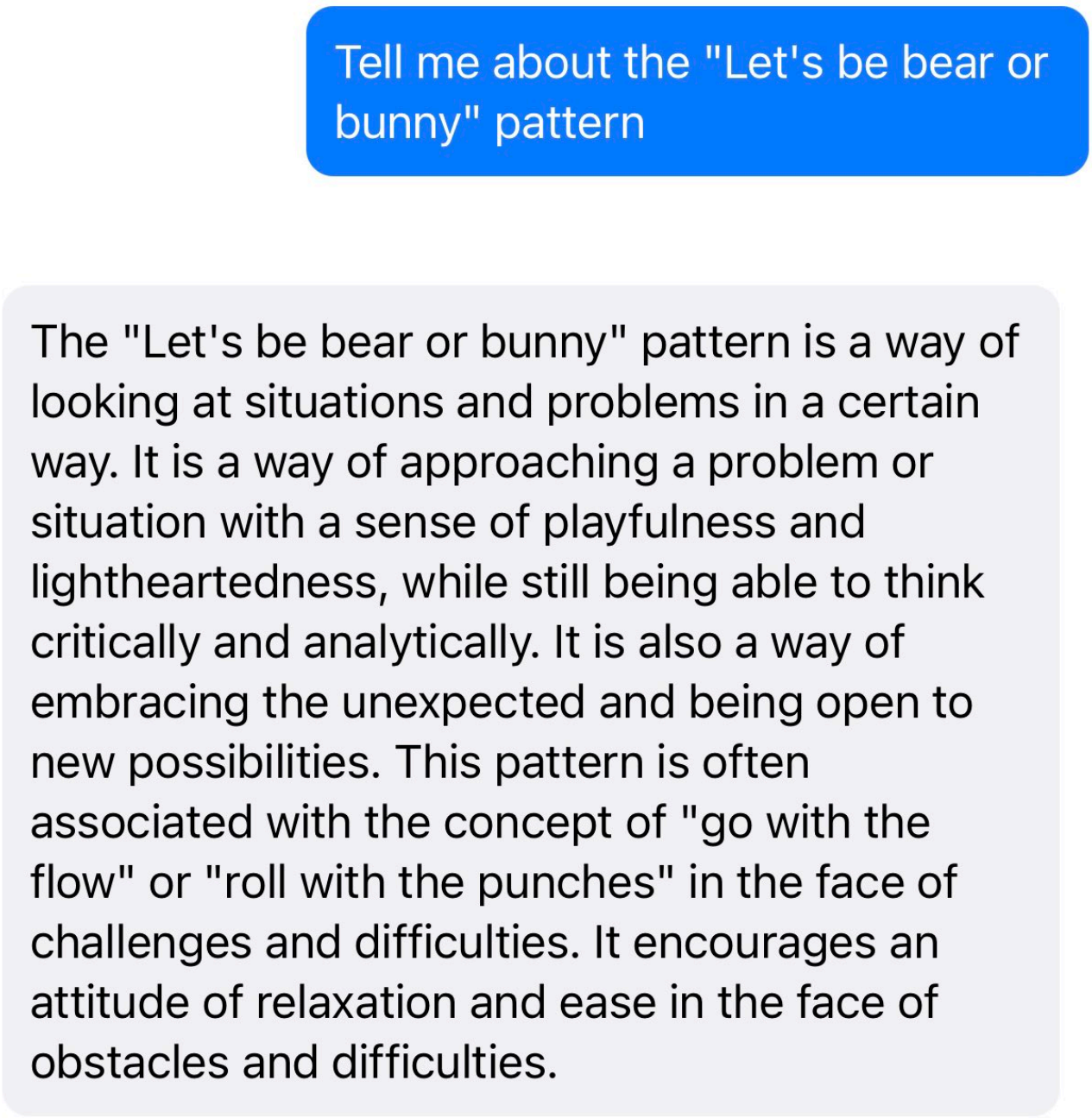
5:12 am / projects, weeknotes, prompt-injection, chatgpt, rye, code-interpreter, coding-agents

---

A new onboarding experience on Mastodon. Reassuring to see this commitment to resolving some of the biggest pain points preventing people from adopting Mastodon, especially given it has meaningful competition as a federated social network in the form of Bluesky now.

5:44 pm / decentralisation, mastodon, bluesky

Let's be bear or bunny



The Machine Learning Compilation group (MLC) are my favourite team of AI researchers at the moment.

[... 599 words]

6:37 pm / ai, generative-ai, llama, local-llms, llms, mlc, vicuna, hallucinations

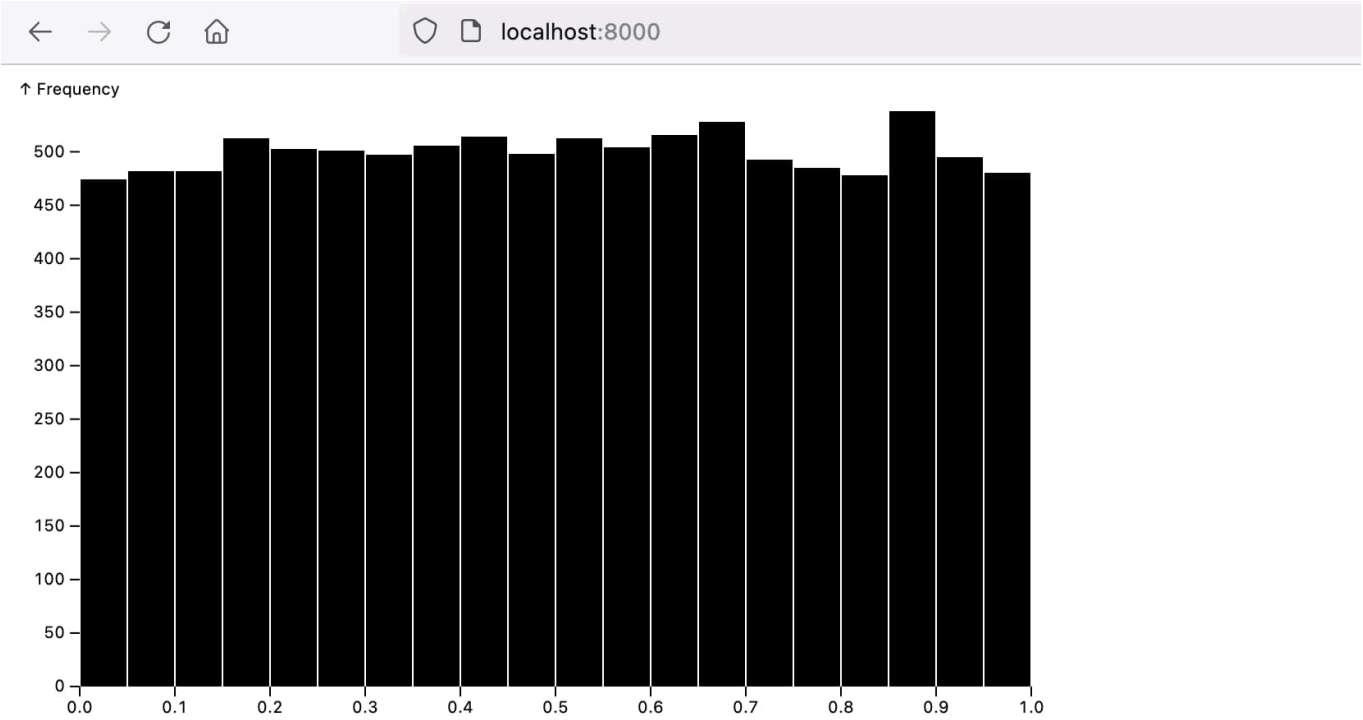
---

Amnesty Uses Warped, AI-Generated Images to Portray Police Brutality in Colombia. I saw massive backlash against Amnesty Norway for this on Twitter, where people argued that using AI-generated images to portray human rights violations like this undermines Amnesty's credibility. I agree: I think this is a very risky move. An Amnesty spokesperson told VICE Motherboard that they did this to provide coverage "without endangering anyone who was present", since many protestors who participated in the national strike covered their faces to avoid being identified.

9:32 pm / ethics, ai, generative-ai, ai-ethics

May 2, 2023

download-esm: a tool for downloading ECMAScript modules



I've built a new CLI tool, download-esm, which takes the name of an npm package and will attempt to download the ECMAScript module version of that package, plus all of its dependencies, directly from the jsDelivr CDN—and then rewrite all of the import statements to point to those local copies.

[... 1,240 words]

4:47 am / cli, ecmascript, javascript, projects, npm, ai-assisted-programming

---

Prompt injection explained, with video, slides, and a transcript

I participated in a webinar this morning about prompt injection, organized by LangChain and hosted by Harrison Chase, with Willem Pienaar, Kojin Oshiba (Robust Intelligence), and Jonathan Cohen and Christopher Parisien (Nvidia Research).

[... 3,120 words]

8:22 pm / security, my-talks, ai, prompt-engineering, prompt-injection, generative-ai, llms, annotated-talks, exfiltration-attacks

---

May 3, 2023

> We show for the first time that large-scale generative pretrained transformer (GPT) family models can be pruned to at least 50% sparsity in one-shot, without any retraining, at minimal loss of accuracy. [...] We can execute SparseGPT on the largest available open-source models, OPT-175B and BLOOM-176B, in under 4.5 hours, and can reach 60% unstructured sparsity with negligible increase in perplexity: remarkably, more than 100 billion weights from these models can be ignored at inference time.
>
> — **SparseGPT,** by Elias Frantar and Dan Alistarh

7:48 pm / llms, ai, generative-ai, bloom, local-llms

---

replit-code-v1-3b (via) As promised last week, Replit have released their 2.7b "Causal Language Model", a foundation model trained from scratch in partnership with MosaicML with a focus on code completion. It's licensed CC BY-SA-4.0 and is available for commercial use. They repo includes a live demo and initial experiments with it look good—you could absolutely run a local GitHub Copilot style editor on top of this model.

8:09 pm / ai, generative-ai, local-llms, llms, llm-release

OpenLLaMA. The first openly licensed model I've seen trained on the RedPajama dataset. This initial release is a 7B model trained on 200 billion tokens, but the team behind it are promising a full 1 trillion token model in the near future. I haven't found a live demo of this one running anywhere yet.

8:58 pm / ai, generative-ai, llama, local-llms, llms, redpajama, llm-release

---

> At this point the lawsuits seem a bit far-fetched: “You should have warned us months ago that artificial intelligence would hurt your business” is unfair given how quickly ChatGPT has exploded from nowhere to become a cultural and business phenomenon. But now everyone is on notice! If you are not warning your shareholders now about how AI could hurt your business, and then it does hurt your business, you’re gonna get sued.
>
> — **Matt Levine**

9:04 pm / chatgpt, ai, generative-ai, matt-levine

May 4, 2023

Mojo may be the biggest programming advance in decades (via) Jeremy Howard makes a very convincing argument for why the new programming language Mojo is a big deal.

Mojo is a superset of Python designed by a team lead by Chris Lattner, who previously created LLVM, Clang and and Swift.

Existing Python code should work unmodified, but it also adds features that enable performant low-level programming—like “fn” for creating typed, compiled functions and “struct” for memory-optimized alternatives to classes.

It’s worth watching Jeremy’s video where he uses these features to get more than a 2000x speed up implementing matrix multiplication, while still keeping the code readable and easy to follow.

Mojo isn’t available yet outside of a playground preview environment, but it does look like an intriguing new project.

4:41 am / programming-languages, python, ai, mojo, jeremy-howard

---

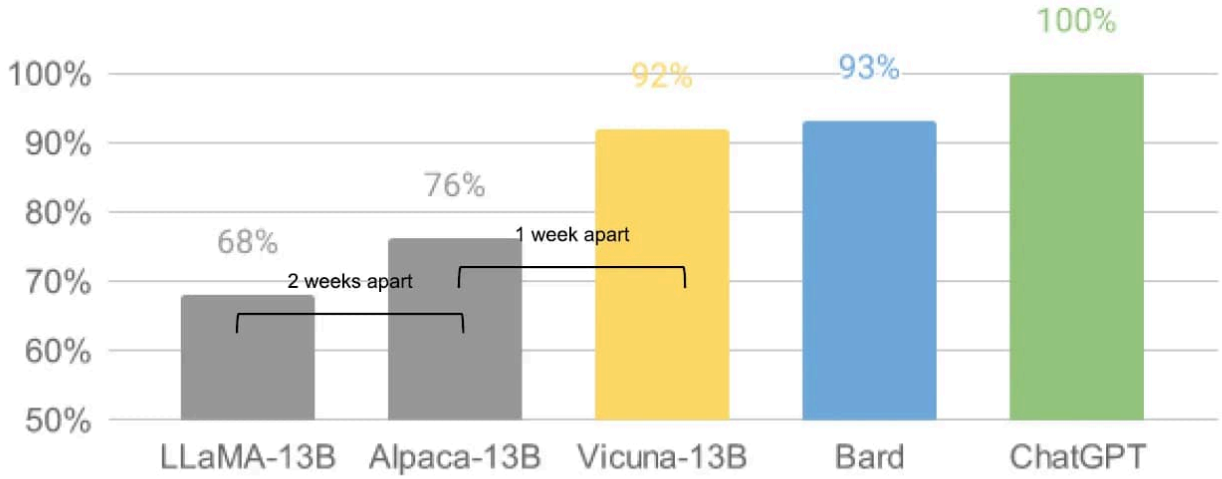
Midjourney 5.1

Midjourney released version 5.1 of their image generation model on Tuesday. Here's their announcement on Twitter—if you have a Discord account there's a more detailed Discord announcement here.

[... 396 words]

3:42 pm / ai, generative-ai, midjourney, text-to-image

Leaked Google document: "We Have No Moat, And Neither Does OpenAI"



*GPT-4 grades LLM outputs. Source: https://vicuna.lmsys.org/

SemiAnalysis published something of a bombshell leaked document this morning: Google "We Have No Moat, And Neither Does OpenAI".

[... 1,073 words]

4:05 pm / google, open-source, openai, generative-ai, local-llms, llms, paper-review

---

May 5, 2023

No Moat: Closed AI gets its Open Source wakeup call — ft. Simon Willison (via) I joined the Latent Space podcast yesterday (on short notice, so I was out and about on my phone) to talk about the leaked Google memo about open source LLMs. This was a Twitter Space, but swyx did an excellent job of cleaning up the audio and turning it into a podcast.

6:17 pm / podcasts, speaking, ai, generative-ai, local-llms, llms, podcast-appearances

---

Introducing MPT-7B: A New Standard for Open-Source, Commercially Usable LLMs (via) There's a lot to absorb about this one. Mosaic trained this model from scratch on 1 trillion tokens, at a cost of $200,000 taking 9.5 days. It's Apache-2.0 licensed and the model weights are available today.

They're accompanying the base model with an instruction-tuned model called MPT-7B-Instruct (licensed for commercial use) and a non-commercially licensed MPT-7B-Chat trained using OpenAI data. They also announced MPT-7B-StoryWriter-65k+—"a model designed to read and write stories with super long context lengths"—with a previously unheard of 65,000 token context length.

They're releasing these models mainly to demonstrate how inexpensive and powerful their custom model training service is. It's a very convincing demo!

7:05 pm / open-source, ai, generative-ai, local-llms, llms

May 8, 2023

Künstliche Intelligenz: Es rollt ein Tsunami auf uns zu (via) A column on AI in Der Spiegel, with a couple of quotes from my blog translated to German.

12:47 am / ai

Big Opportunities in Small Data

I gave an invited keynote at Citus Con 2023, the PostgreSQL conference. Below is the abstract, video, slides and links from the presentation.

[... 385 words]

3:06 am / postgresql, sqlite, my-talks, datasette, small-data

---

> Because we do not live in the Star Trek-inspired rational, humanist world that Altman seems to be hallucinating. We live under capitalism, and under that system, the effects of flooding the market with technologies that can plausibly perform the economic tasks of countless working people is not that those people are suddenly free to become philosophers and artists. It means that those people will find themselves staring into the abyss – with actual artists among the first to fall.
>
> — **Naomi Klein**

3:09 pm / ai, ethics, generative-ai, ai-ethics

> What Tesla is contending is deeply troubling to the Court. Their position is that because Mr. Musk is famous and might be more of a target for deep fakes, his public statements are immune. In other words, Mr. Musk, and others in his position, can simply say whatever they like in the public domain, then hide behind the potential for their recorded statements being a deep fake to avoid taking ownership of what they did actually say and do. The Court is unwilling to set such a precedent by condoning Tesla's approach here.
>
> — **Judge Evette Pennypacker**

4:46 pm / ai, ethics, generative-ai, ai-ethics

Seashells. This is a really useful tool for monitoring the status of a long-running CLI script on another device. You can run any command and pipe its output to “nc seashells.io 1337”—which will then return the URL to a temporary web page which

you can view on another device (including a mobile phone) to see the constantly updating output of that command.

5:20 pm / cli

GitHub code search is generally available. I've been a beta user of GitHub's new code search for a year and a half now and I wouldn't want to be without it. It's spectacularly useful: it provides fast, regular-expression-capable search across every public line of code hosted by GitHub—plus code in private repos you have access to.

I mainly use it to compensate for libraries with poor documentation—I can usually find an example of exactly what I want to do somewhere on GitHub.

It's also great for researching how people are using libraries that I've released myself—to figure out how much pain deprecating a method would cause, for example.

6:52 pm / github, open-source, search

Jsonformer: A Bulletproof Way to Generate Structured JSON from Language Models. This is such an interesting trick. A common challenge with LLMs is getting them to output a specific JSON shape of data reliably, without occasionally messing up and generating invalid JSON or outputting other text.

Jsonformer addresses this in a truly ingenious way: it implements code that interacts with the logic that decides which token to output next, influenced by a JSON schema. If that code knows that the next token after a double quote should be a comma it can force the issue for that specific token.

This means you can get reliable, robust JSON output even for much smaller, less capable language models.

It's built against Hugging Face transformers, but there's no reason the same idea couldn't be applied in other contexts as well.

11:02 pm / json, ai, generative-ai, llms, hugging-face

> When trying to get your head around a new technology, it helps to focus on how it challenges existing categorizations, conventions, and rule sets. Internally, I've always called this exercise, "dealing with the platypus in the room." Named after the category-defying animal; the duck-billed, venomous, semi-aquatic, egg-laying mammal. [...] AI is the biggest platypus I've ever seen. Nearly every notable quality of AI and LLMs challenges our conventions, categories, and rulesets.
>
> — **Drew Breunig**

11:14 pm / ai, generative-ai, drew-breunig

May 9, 2023

Language models can explain neurons in language models (via) Fascinating interactive paper by OpenAI, describing how they used GPT-4 to analyze the concepts tracked by individual neurons in their much older GPT-2 model. "We generated cluster labels by embedding each neuron explanation using the OpenAI Embeddings API, then clustering them and asking GPT-4 to label each cluster."

5:35 pm / ai, explorables, openai, generative-ai, gpt-4, llms, embeddings, gpt-2

ImageBind. New model release from Facebook/Meta AI research: "An approach to learn a joint embedding across six different modalities—images, text, audio, depth, thermal, and IMU (inertial measurement units) data". The non-interactive

demo shows searching audio starting with an image, searching images starting with audio, using text to retrieve images and audio, using image and audio to retrieve images (e.g. a barking sound and a photo of a beach to get dogs on a beach) and using audio as input to an image generator.

7:04 pm / facebook, ai, generative-ai, embeddings

May 10, 2023

Thunderbird Is Thriving: Our 2022 Financial Report (via) Astonishing numbers: in 2022 the Thunderbird project received $6,442,704 in donations from 300,000 users. These donations are now supporting 24 staff members. Part of their success is credited to an “in-app donations appeal” that they launched at the end of 2022.

12:14 am / mozilla, open-source, thunderbird

See this page fetch itself, byte by byte, over TLS (via) George MacKerron built a TLS 1.3 library in TypeScript and used it to construct this amazing educational demo, which performs a full HTTPS request for its own source code over a WebSocket and displays an annotated byte-by-byte representation of the entire exchange. This is the most useful illustration of how HTTPS actually works that I’ve ever seen.

1:58 pm / encryption, http, https, tls, websockets, explorables

---

> The largest model in the PaLM 2 family, PaLM 2-L, is significantly smaller than the largest PaLM model but uses more training compute. Our evaluation results show that PaLM 2 models significantly outperform PaLM on a variety of tasks, including natural language generation, translation, and reasoning. These results suggest that model scaling is not the only way to improve performance. Instead, performance can be unlocked by meticulous data selection and efficient architecture/objectives. Moreover, a smaller but higher quality model significantly improves inference efficiency, reduces serving cost, and enables the model’s downstream application for more applications and users.
>
> — **PaLM 2 Technical Report**, PDF

6:43 pm / google, generative-ai, bard, ai, llms

Hugging Face Transformers Agent. Fascinating new Python API in Hugging Face Transformers version v4.29.0: you can now provide a text description of a task—e.g. “Draw me a picture of the sea then transform the picture to add an island”—and a LLM will turn that into calls to Hugging Face models which will then be installed and used to carry out the instructions. The Colab notebook is worth playing with—you paste in an OpenAI API key and a Hugging Face token and it can then run through all sorts of examples, which tap into tools that include image generation, image modification, summarization, audio generation and more.

7:50 pm / ai, generative-ai, llms, hugging-face

Weeknotes: sqlite-utils 3.31, download-esm, Python in a sandbox

A couple of speaking appearances last week—one planned, one unplanned. Plus `sqlite-utils 3.31`, `download-esm` and a new TIL.

[... 608 words]

10:07 pm / projects, python, weeknotes, deno, sqlite-utils, pyodide

---

2023 » May

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**	31				