# January 2024

Search posts from January 2024 Search

75 posts: 8 entries, 57 links, 10 quotes

## Jan. 1, 2024

**After ten years, it's time to stop making videos.** Ten years ago, my friend Tom Scott started a deliberate streak of posting YouTube videos - initially about one a day before settling into a cadence of one a week. He kept that up for the full ten years, growing his subscribers to over 6 million in the process.

Today he's ending that streak, in unparalleled style.

(I'm proud to have made an appearance in video number 13, talking about Zeppelins.)

# 10:59 pm / tom-scott, youtube, zeppelins, streaks

---

## Jan. 2, 2024

> Since the advent of ChatGPT, and later by using LLMs that operate locally, I have made extensive use of this new technology. The goal is to accelerate my ability to write code, but that's not the only purpose. There's also the intent to not waste mental energy on aspects of programming that are not worth the effort.
>
> [...] Current LLMs will not take us beyond the paths of knowledge, but if we want to tackle a topic we do not know well, they can often lift us from our absolute ignorance to the point where we know enough to move forward on our own.
>
> — **Salvatore Sanfilippo**

# 2:50 pm / salvatore-sanfilippo, llms, ai, generative-ai, chatgpt

---

**NPM: modele-social** (via) This is a fascinating open source package: it's an NPM module containing an implementation of the rules for calculating social security contributions in France, maintained by a team at Urssaf, the not-quite-government organization in France that manages the collection of social security contributions there.

The rules themselves can be found in the associated GitHub repository, encoded in a YAML-like declarative language called Publicodes that was developed by the French government for this and similar purposes.

# 5:55 pm / government, open-source, npm

---

## Tom Scott, and the formidable power of escalating streaks


Ten years ago yesterday, Tom Scott posted this video to YouTube about "Special Crossings For Horses In Britain". It was the first in his Things You Might Not Know series, but more importantly it was the start of a streak.

[... 1,352 words]

---

8:32 pm / inspiring, productivity, tom-scott, youtube, streaks, duolingo

---

## Jan. 3, 2024

**Fastest Way to Read Excel in Python** (via) Haki Benita produced a meticulously researched and written exploration of the options for reading a large Excel spreadsheet into Python. He explored Pandas, Tablib, Openpyxl, shelling out to LibreOffice, DuckDB and python-calamine (a Python wrapper of a Rust library). Calamine was the winner, taking 3.58s to read 500,00 rows—compared to Pandas in last place at 32.98s.

# 8:04 pm / excel, pandas, python, rust, duckdb, haki-benita

---

**container2wasm** (via) "Converts a container to WASM with emulation by Bochs (for x86_64 containers) and TinyEMU (for riscv64 containers)"—effectively letting you take a Docker container and turn it into a WebAssembly blob that can then run in any WebAssembly host environment, including the browser.

Run "c2w ubuntu:22.04 out.wasm" to output a WASM binary for the Ubuntu 22:04 container from Docker Hub, then "wasmtime out.wasm uname -a" to run a command.

Even better, check out the live browser demos linked fro the README, which let you do things like run a Python interpreter in a Docker container directly in your browser.

# 11:21 pm / docker, webassembly

---

## Jan. 4, 2024

**My blog's year archive pages now have tag clouds** (via) Inspired by the tag cloud I used in my recent 2023 AI roundup post, I decided to add a tag cloud to the top of every one of my archive-by-year pages showing what topics I had spent the most time with that year.

I already had old code for this, so I pasted it into GPT-4 along with an example of the output of my JSON endpoint from Django SQL Dashboard and had it do most of the work for me.

# 9:02 pm / projects, ai, django-sql-dashboard, chatgpt, llms

---

## Jan. 5, 2024

> If you learn something the hard way, share your findings with others. You have blazed a new trail; now you must mark it for your fellow travellers. Sharing knowledge is an unreasonably effective way of helping others.
>
> — **Nicolas Bouliane**

# 10:32 pm / documentation

---

## Jan. 6, 2024

**Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations** (via) NIST—the National Institute of Standards and Technology, a US government agency, released a 106 page report on attacks against modern machine learning models, mostly covering LLMs.

Prompt injection gets two whole sections, one on direct prompt injection (which incorporates jailbreaking as well, which they misclassify as a subset of prompt injection) and one on indirect prompt injection.

They talk a little bit about mitigations, but for both classes of attack conclude: "Unfortunately, there is no comprehensive or foolproof solution for protecting models against adversarial prompting, and future work will need to be dedicated to investigating suggested defenses for their efficacy."

# 4:08 am / jailbreaking, ai, prompt-injection, generative-ai, llms

---

**Microsoft Research relicense Phi-2 as MIT** (via) Phi-2 was already an interesting model—really strong results for its size—made available under a non-commercial research license. It just got significantly more interesting: Microsoft relicensed it as MIT open source.

# 6:06 am / microsoft, open-source, mitlicense, ai, generative-ai, llms, phi

---

**GPT in 500 lines of SQL** (via) Utterly brilliant piece of PostgreSQL hackery by Alex Bolenok, who implements a full GPT-2 style language model in SQL on top of pg_vector. The final inference query is 498 lines long!

# 10:55 pm / postgresql, sql, ai, generative-ai, llms, gpt-2

---

## Jan. 7, 2024

## It's OK to call it Artificial Intelligence

***Update 9th January 2024**: This post was clumsily written and failed to make the point I wanted it to make. I've published a follow-up, What I should have said about the term Artificial Intelligence which you should read instead.*

[... 1,818 words]

---

12:01 am / ai, generative-ai, llms

---

## Weeknotes: Page caching and custom templates for Datasette Cloud

My main development focus this week has been adding public page caching to Datasette Cloud, and exploring what custom template support might look like for that service.

[... 924 words]

---

8:45 pm / caching, security, varnish, xss, datasette, cloudflare, weeknotes, datasette-cloud

---

## Jan. 8, 2024

**Text Embeddings Reveal (Almost) As Much As Text**. Embeddings of text—where a text string is converted into a fixed-number length array of floating point numbers—are demonstrably reversible: "a multi-step method that iteratively corrects and re-embeds text is able to recover 92% of 32-token text inputs exactly".

This means that if you're using a vector database for embeddings of private data you need to treat those embedding vectors with the same level of protection as the original text.

# 5:22 am / privacy, security, ai, embeddings

---

**Does GPT-2 Know Your Phone Number?** (via) This report from Berkeley Artificial Intelligence Research in December 2020 showed GPT-3 outputting a full page of chapter 3 of Harry Potter and the Philosopher's Stone—similar to how the recent suit from the New York Times against OpenAI and Microsoft demonstrates memorized news articles from that publication as outputs from GPT-4.

# 5:26 am / microsoft, new-york-times, ai, gpt-3, openai, generative-ai, llms, gpt-2

---

> We believe that AI tools are at their best when they incorporate and represent the full diversity and breadth of human intelligence and experience. [...] Because copyright today covers virtually every sort of human expression– including blog posts, photographs, forum posts, scraps of software code, and government documents–it would be impossible to train today's leading AI models without using copyrighted materials. Limiting training data to public domain books and drawings created more than a century ago might yield an interesting experiment, but would not provide AI systems that meet the needs of today's citizens.
>
> — **OpenAI to the Lords Select Committee on LLMs**

# 5:33 pm / copyright, generative-ai, openai, ai, llms, politics, training-data

---

**OpenAI and journalism**. Bit of a misleading title here: this is OpenAI's first public response to the lawsuit filed by the New York Times concerning their use of unlicensed NYT content to train their models.

# 6:33 pm / copyright, new-york-times, ai, openai, generative-ai, llms

---

## Jan. 9, 2024

**Mixtral of Experts**. The Mixtral paper is out, exactly a month after the release of the Mixtral 8x7B model itself. Thanks to the paper I now have a reasonable understanding of how a mixture of experts model works: each layer has 8 available blocks, but a router model selects two out of those eight for each token passing through that layer and combines their output. "As a result, each token has access to 47B parameters, but only uses 13B active parameters during inference."

The Mixtral token context size is an impressive 32k, and it compares extremely well against the much larger Llama 70B across a whole array of benchmarks.

Unsurprising but disappointing: there's nothing in the paper at all about what it was trained on.

# 4:03 am / ai, generative-ai, local-llms, llms, mistral

---

## What I should have said about the term Artificial Intelligence

With the benefit of hindsight, I did a bad job with my post, It's OK to call it Artificial Intelligence a few days ago.

[... 376 words]

9:13 pm / ai, llms

---

**Python 3.13 gets a JIT**. "In late December 2023 (Christmas Day to be precise), CPython core developer Brandt Bucher submitted a little pull-request to the Python 3.13 branch adding a JIT compiler."

Anthony Shaw does a deep dive into this new experimental JIT, explaining how it differs from other JITs. It's an implementation of a copy-and-patch JIT, an idea that only emerged in 2021. This makes it architecturally much simpler than a traditional JIT, allowing it to compile faster and take advantage of existing LLVM tools on different architectures.

So far it's providing a 2-9% performance improvement, but the real impact will be from the many future optimizations it enables.

# 9:25 pm / jit, llvm, python, anthony-shaw

---

**WikiChat: Stopping the Hallucination of Large Language Model Chatbots by Few-Shot Grounding on Wikipedia**. This paper describes a really interesting LLM system that runs Retrieval Augmented Generation against Wikipedia to help answer questions, but includes a second step where facts in the answer are fact-checked against Wikipedia again before returning an answer to the user. They claim "97.3% factual accuracy of its claims in simulated conversation" on a GPT-4 backed version, and also see good results when backed by LLaMA 7B.

The implementation is mainly through prompt engineering, and detailed examples of the prompts they used are included at the end of the paper.

# 9:30 pm / wikipedia, ai, prompt-engineering, generative-ai, llms, rag, hallucinations

---

**The Eight Golden Rules of Interface Design** (via) By HCI researcher Ben Shneiderman. I particularly like number 4, "Design dialogs to yield closure", which encourages feedback at the completion of a group of actions that "gives users the satisfaction of accomplishment, a sense of relief."

# 9:37 pm / usability, ux

---

**ooh.directory: A page for every blog**. I hadn't checked in on Phil Gyford's ooh.directory blog directory since it first launched in November 2022. I'm delighted to see that it's thriving—2,117 blogs have now been carefully curated, and the latest feature is a page for each blog showing its categories, description, an activity graph and the most recent posts syndicated via RSS/Atom.

# 10:15 pm / atom, blogs, phil-gyford, rss, syndication

---

## Jan. 10, 2024

**The Random Transformer** (via) "Understand how transformers work by demystifying all the math behind them"—Omar Sanseviero from Hugging Face meticulously implements the transformer architecture behind LLMs from scratch using Python and numpy. There's a lot to take in here but it's all very clearly explained.

# 5:09 am / python, transformers, ai, numpy, generative-ai, llms

---

**You Can Build an App in 60 Minutes with ChatGPT, with Geoffrey Litt** (via) YouTube interview between Dan Shipper and Geoffrey Litt. They talk about how ChatGPT can build working React applications and how this means you can build extremely niche applications that you woudn't have considered working on before—then to demonstrate that idea, they collaborate to build a note-taking app to be used just during that specific episode recording, pasting React code from ChatGPT into Replit.

Geoffrey: "I started wondering what if we had a world where everybody could craft software tools that match the workflows they want to have, unique to themselves and not just using these pre-made tools. That's what malleable software means to me."

# 11:41 pm / ai, react, generative-ai, chatgpt, llms, geoffrey-litt

---

**AI versus old-school creativity: a 50-student, semester-long showdown** (via) An interesting study in which 50 university students "wrote, coded, designed, modeled, and recorded creations with and without AI, then judged the results".

This study seems to explore the approach of incremental prompting to produce an AI-driven final results. I use GPT-4 on a daily basis but my usage patterns are quite different: I very rarely let it actually write anything for me, instead using it as brainstorming partner, or to provide feedback, or as API reference or a thesaurus.

# 11:49 pm / education, ai, generative-ai, chatgpt, llms

---

## Jan. 11, 2024

**Budgeting with ChatGPT** (via) Jon Callahan describes an ingenious system he set up to categorize his credit card transactions using GPT 3.5. He has his bank email him details of any transaction over $0, then has an email filter to forward those to Postmark, which sends them via a JSON webhook to a custom Deno Deploy app which cleans the transaction up with a GPT 3.5 prompt (including guessing the merchant) and submits the results to a base in Airtable.

# 4:40 am / email, ai, airtable, openai, llms

---

## Jan. 12, 2024

**Where is all of the fediverse?** (via) Neat piece of independent research by Ben Cox, who used the /api/v1/instance/peers Mastodon API endpoint to get a list of "peers" (instances his instance knows about), then used their DNS records to figure out which hosting provider they were running on.

Next Ben combined that with active users from the /nodeinfo/2.0 API on each instance to figure out the number of users on each of those major hosting providers.

Cloudflare and Fastly were heavily represented, but it turns out you can unveil the underlying IP for most instances by triggering an HTTP Signature exchange with them and logging the result.

Ben's conclusion: Hertzner and OVH are responsible for hosting a sizable portion of the fediverse as it exists today.

# 6:54 pm / dns, hosting, mastodon, fediverse

---

**Marimo** (via) This is a really interesting new twist on Python notebooks.

The most powerful feature is that these notebooks are reactive: if you change the value or code in a cell (or change the value in an input widget) every other cell that depends on that value will update automatically. It's the same pattern implemented by Observable JavaScript notebooks, but now it works for Python.

There are a bunch of other nice touches too. The notebook file format is a regular Python file, and those files can be run as "applications" in addition to being edited in the notebook interface. The interface is very nicely built, especially for such a young project—they even have GitHub Copilot integration for their CodeMirror cell editors.

# 9:17 pm / open-source, python, jupyter, observable, github-copilot, marimo

---

## Jan. 13, 2024

**More than an OpenAI Wrapper: Perplexity Pivots to Open Source**. I'm increasingly impressed with Perplexity.ai—I'm using it on a daily basis now. It's by far the best implementation I've seen of LLM-assisted search—beating Microsoft Bing and Google Bard at their own game.

A year ago it was implemented as a GPT 3.5 powered wrapper around Microsoft Bing. To my surprise they've now evolved way beyond that: Perplexity has their own search index now and is running their own crawlers, and they're using variants of Mistral 7B and Llama 70B as their models rather than continuing to depend on OpenAI.

# 6:12 am / crawling, search, ai, generative-ai, llms, perplexity, ai-assisted-search

---



2024 » January

| 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 | | | | |