September 2023

Search posts from September 2023 Search

45 posts: 5 entries, 35 links, 5 quotes

Sept. 4, 2023

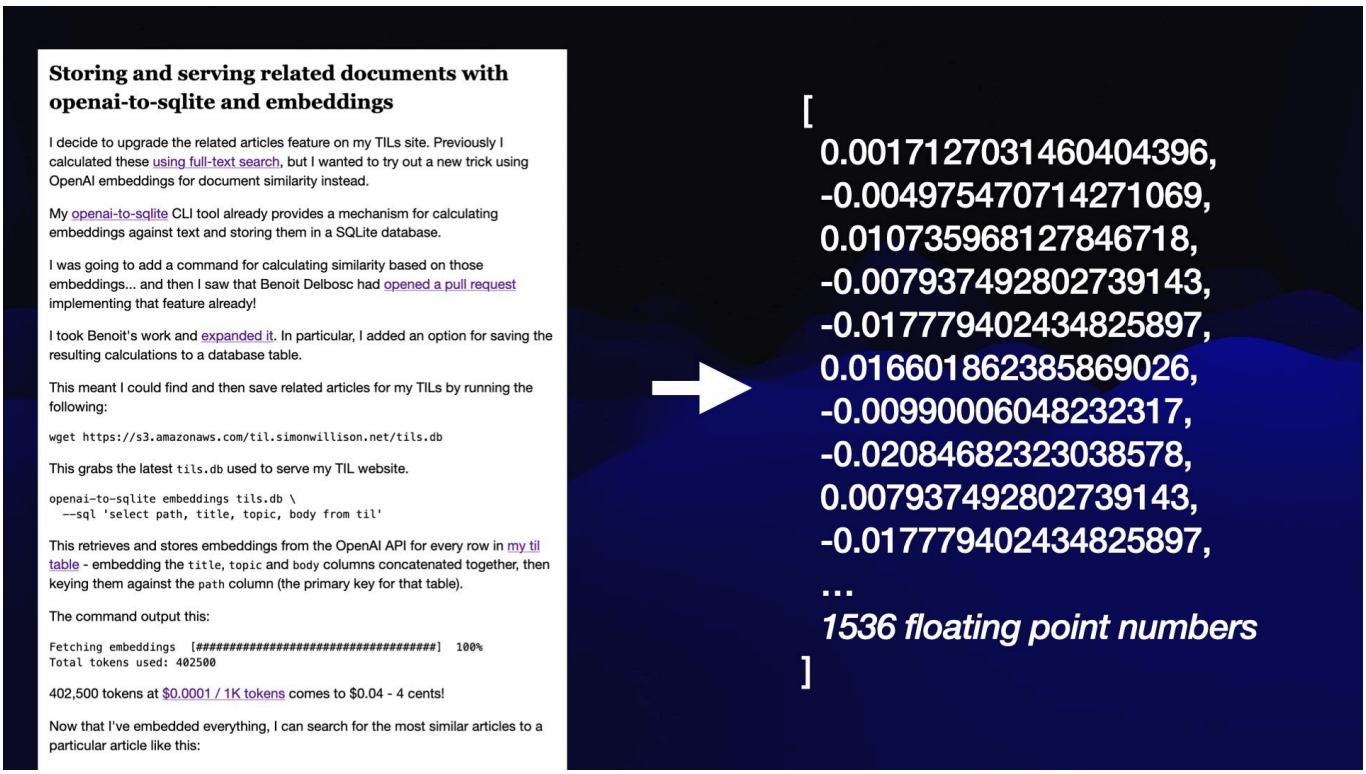
A practical guide to deploying Large Language Models Cheap, Good *and* Fast. Joel Kang's extremely comprehensive notes on what he learned trying to run Vicuna-13B-v1.5 on an affordable cloud GPU server (a T4 at $0.615/hour). The space is in so much flux right now—Joel ended up using MLC but the best option could change any minute.

Vicuna 13B quantized to 4-bit integers needed 7.5GB of the T4's 16GB of VRAM, and returned tokens at 20/second.

An open challenge running MLC right now is around batching and concurrency: "I did try making 3 concurrent requests to the endpoint, and while they all stream tokens back and the server doesn't OOM, the output of all 3 streams seem to actually belong to a single prompt."

1:43 pm / ai, generative-ai, llama, llms, mlc, vicuna

LLM now provides tools for working with embeddings



LLM is my Python library and command-line tool for working with language models. I just released LLM 0.9 with a new set of features that extend LLM to provide tools for working with *embeddings*.

[... 3,521 words]

8:32 pm / cli, open-source, projects, sqlite, ai, generative-ai, llms, embeddings, llm, rag

---

Wikipedia search-by-vibes through millions of pages offline (via) Really cool demo by Lee Butterman, who built embeddings of 2 million Wikipedia pages and figured out how to serve them directly to the browser, where they are used to implement “vibes based” similarity search returning results in 250ms. Lots of interesting details about how he pulled this off, using Arrow as the file format and ONNX to run the model in the browser.

9:13 pm / embedding, search, wikipedia, webassembly

Sept. 5, 2023

A token-wise likelihood visualizer for GPT-2. Linus Lee built a superb visualization to help demonstrate how Large Language Models work, in the form of a video essay where each word is coloured to show how “surprising” it is to the model. It’s worth carefully reading the text in the video as each term is highlighted to get the full effect.

3:39 am / ai, generative-ai, llms, gpt-2

Symbex 1.4. New release of my Symbex tool for finding symbols (functions, methods and classes) in a Python codebase. Symbex can now output matching symbols in JSON, CSV or TSV in addition to plain text.

I designed this feature for compatibility with the new “llm embed-multi” command—so you can now use Symbex to find every Python function in a nested directory and then pipe them to LLM to calculate embeddings for every one of them.

I tried it on my projects directory and embedded over 13,000 functions in just a few minutes! Next step is to figure out what kind of interesting things I can do with all of those embeddings.

5:29 pm / projects, ai, generative-ai, embeddings, symbex, llm

---

Sept. 6, 2023

Perplexity: interactive LLM visualization (via) I linked to a video of Linus Lee's GPT visualization tool the other day. Today he's released a new version of it that people can actually play with: it runs entirely in a browser, powered by a 120MB version of the GPT-2 ONNX model loaded using the brilliant Transformers.js JavaScript library.

3:33 am / javascript, ai, webassembly, generative-ai, llms, transformers-js

---

Using ChatGPT Code Intepreter (aka “Advanced Data Analysis”) to analyze your ChatGPT history. I posted a short thread showing how to upload your ChatGPT history to ChatGPT itself, then prompt it with “Build a dataframe of the id, title, create_time properties from the conversations.json JSON array of objects. Convert create_time to a date and plot it daily”.

3:42 pm / ai, generative-ai, chatgpt, llms

---

hubcap.php (via) This PHP script by Dave Hulbert delights me. It’s 24 lines of code that takes a specified goal, then calls my LLM utility on a loop to request the next shell command to execute in order to reach that goal... and pipes the output straight into `exec()` after a 3s wait so the user can panic and hit Ctrl+C if it’s about to do something dangerous!

3:45 pm / php, security, ai, generative-ai, llms, llm, ai-agents

---

Sept. 8, 2023

bpy—Blender on PyPI (via) TIL you can “pip install” Blender!

bpy “provides Blender as a Python module”—it’s part of the official Blender project, and ships with binary wheels ranging in size from 168MB to 319MB depending on your platform.

It only supports the version of Python used by the current Blender release though—right now that’s Python 3.10.

3:29 pm / pypi, python, blender

Dynamic linker tricks: Using LD_PRELOAD to cheat, inject features and investigate programs (via) This tutorial by Rafał Cieślak from 2013 filled in a bunch of gaps in my knowledge about how C works on Linux.

10:05 pm / c, linux

Sept. 9, 2023

Matthew Honnibal from spaCy on why LLMs have not solved NLP. A common trope these days is that the entire field of NLP has been effectively solved by Large Language Models. Here’s a lengthy comment from Matthew Honnibal, creator of the highly regarded spaCy Python NLP library, explaining in detail why that argument doesn’t hold up.

9:30 pm / nlp, ai, generative-ai, llms

Sept. 10, 2023

promptfoo: How to benchmark Llama2 Uncensored vs. GPT-3.5 on your own inputs. promptfoo is a CLI and library for “evaluating LLM output quality”. This tutorial in their documentation about using it to compare Llama 2 to gpt-3.5-turbo is a good illustration of how it works: it uses YAML files to configure the prompts, and more YAML to define assertions such as “not-icontains: AI language model”.

4:19 pm / cli, testing, ai, generative-ai, llms

The AI-assistant wars heat up with Claude Pro, a new ChatGPT Plus rival. I'm quoted in this piece about the new Claude Pro $20/month subscription from Anthropic:

> Willison has also run into problems with Claude's morality filter, which has caused him trouble by accident: "I tried to use it against a transcription of a podcast episode, and it processed most of the text before—right in front of my eyes—it deleted everything it had done! I eventually figured out that they had started talking about bomb threats against data centers towards the end of the episode, and Claude effectively got triggered by that and deleted the entire transcript."

5:07 pm / arstechnica, ai, generative-ai, llms, anthropic, claude, press-quotes

---

All models on Hugging Face, sorted by downloads (via) I realized this morning that “sort by downloads” against the list of all of the models on Hugging Face can work as a reasonably good proxy for “which of these models are easiest to get running on your own computer”.

5:24 pm / machine-learning, ai, hugging-face

Sept. 12, 2023

Build an image search engine with llm-clip, chat with models with llm chat

LLM is my combination CLI tool and Python library for working with Large Language Models. I just released LLM 0.10 with two significant new features: embedding support for binary files and the `llm chat` command.

[... 1,188 words]

8:33 pm / cli, projects, ai, annotated-release-notes, generative-ai, local-llms, llms, embeddings, llm, clip

---

Sept. 13, 2023

Simulating History with ChatGPT (via) Absolutely fascinating new entry in the using-ChatGPT-to-teach genre. Benjamin Breen teaches history at UC Santa Cruz, and has been developing a sophisticated approach to using ChatGPT to play out role-playing scenarios involving different periods of history. His students are challenged to participate in them, then pick them apart—fact-checking details from the scenario and building critiques of the perspectives demonstrated by the language model. There are so many quotable snippets in here, I recommend reading the whole thing.

3:36 am / education, teaching, ai, generative-ai, chatgpt, llms, benjamin-breen

---

> In the long term, I suspect that LLMs will have a significant positive impact on higher education. Specifically, I believe they will elevate the importance of the humanities. [...] LLMs are deeply, inherently

> textual. And they are reliant on text in a way that is directly linked to the skills and methods that we emphasize in university humanities classes.
>
> — **Benjamin Breen**

3:40 am / generative-ai, chatgpt, education, ai, llms, benjamin-breen

---

Some notes on Local-First Development (via) Local-First is the name that has been coined by the community of people who are interested in building apps where data is manipulated in a client application first (mobile, desktop or web) and then continually synchronized with a server, rather than the other way round. This is a really useful review by Kyle Mathews of how the space is shaping up so far—lots of interesting threads to follow here.

3:48 am / local-first

Introducing datasette-litestream: easy replication for SQLite databases in Datasette. We use Litestream on Datasette Cloud for streaming backups of user data to S3. Alex Garcia extracted out our implementation into a standalone Datasette plugin, which bundles the Litestream Go binary (for the relevant platform) in the package you get when you run "datasette install datasette-litestream"—so now Datasette has a very robust answer to questions about SQLite disaster recovery beyond just the Datasette Cloud platform.

7:28 pm / plugins, sqlite, datasette, datasette-cloud, litestream, alex-garcia

---

Sept. 14, 2023

CAISO Grid Status (via) CAISO is the California Independent System Operator, a non-profit managing 80% of California's electricity flow. This grid status page shows live data about the state of the grid and it's fascinating: right now (2pm local time) California is running 71.4% on renewables, having peaked at 80% three hours ago. The current fuel mix is 52% solar, 31% natural gas, 7% each large hydro and nuclear and 2% wind. The charts on this page show how solar turns off overnight and then picks up and peaks during daylight hours.

9:08 pm / energy, california

Sept. 16, 2023

How CPython Implements and Uses Bloom Filters for String Processing. Fascinating dive into Python string internals by Abhinav Upadhyay. It turns out CPython uses very simple bloom filters in several parts of the core string methods, to solve problems like splitting on newlines where there are actually eight codepoints that could represent a newline, and a tiny bloom filter can help filter a character in a single operation before performing all eight comparisons only if that first check failed.

10:32 pm / bloom-filters, performance, python

Notes on using a single-person Mastodon server. Julia Evans experiences running a single-person Mastodon server (on masto.host—the same host I use for my own) pretty much exactly match what I've learned so far as well. The biggest disadvantage is the missing replies issue, where your server only shows replies to posts that come from people who you follow—so it's easy to reply to something in a way that duplicates other replies that are invisible to you.

10:35 pm / julia-evans, mastodon

Sept. 17, 2023

Weeknotes: Embeddings, more embeddings and Datasette Cloud

Since my last weeknotes, a flurry of activity. LLM has embeddings support now, and Datasette Cloud has driven some major improvements to the wider Datasette ecosystem.

[... 2,427 words]

5:10 am / plugins, projects, datasette, weeknotes, datasette-cloud, sqlite-utils, alex-garcia, embeddings, llm

---

Sept. 18, 2023

> Note that there have been no breaking changes since the [SQLite] file format was designed in 2004. The changes shows in the version history above have all be one of (1) typo fixes, (2) clarifications, or (3) filling in the "reserved for future extensions" bits with descriptions of those extensions as they occurred.
>
> — **D. Richard Hipp**

6:02 pm / d-richard-hipp, sqlite

Sept. 19, 2023

LLM 0.11. I released LLM 0.11 with support for the new gpt-3.5-turbo-instruct completion model from OpenAI.

The most interesting feature of completion models is the option to request “log probabilities” from them, where each token returned is accompanied by up to 5 alternatives that were considered, along with their scores.

3:28 pm / projects, ai, openai, generative-ai, llms, llm

The WebAssembly Go Playground (via) Jeff Lindsay has a full Go 1.21.1 compiler running entirely in the browser.

7:53 pm / go, jeff-lindsay, webassembly

Sept. 23, 2023

TG: Polygon indexing (via) TG is a brand new geospatial library by Josh Baker, author of the Tile38 in-memory spatial server (kind of a geospatial Redis). TG is written in pure C and delivered as a single C file, reminiscent of the SQLite amalgamation.

TG looks really interesting. It implements almost the exact subset of geospatial functionality that I find most useful: point-in-polygon, intersect, WKT, WKB, and GeoJSON—all with no additional dependencies.

The most interesting thing about it is the way it handles indexing. In this documentation Josh describes two approaches he uses to speeding up point-in-polygon and intersection using a novel approach that goes beyond the usual RTree implementation.

I think this could make the basis of a really useful SQLite extension—a lighter-weight alternative to SpatiaLite.

4:32 am / c, geospatial, gis, spatialite, sqlite, geojson, tg

---

Sept. 24, 2023

Should you give candidates feedback on their interview performance? Jacob provides a characteristically nuanced answer to the question of whether you should provide feedback to candidates you have interviewed. He suggests offering the candidate the option to email asking for feedback early in the interview process to avoid feeling pushy later on, and proposes the phrase “you failed to demonstrate...” as a useful framing device.

10:25 pm / jacob-kaplan-moss, management

Sept. 25, 2023

A Hackers’ Guide to Language Models. Jeremy Howard’s new 1.5 hour YouTube introduction to language models looks like a really useful place to catch up if you’re an experienced Python programmer looking to start experimenting with LLMs. He covers what they are and how they work, then shows how to build against the OpenAI API, build a Code Interpreter clone using OpenAI functions, run models from Hugging Face on your own machine (with NVIDIA cards or on a Mac) and finishes with a demo of fine-tuning a Llama 2 model to perform text-to-SQL using an open dataset.

12:24 am / python, ai, openai, generative-ai, llama, llms, jeremy-howard, fine-tuning, nvidia

---

> We already know one major effect of AI on the skills distribution: AI acts as a skills leveler for a huge range of professional work. If you were in the bottom half of the skill distribution for writing, idea generation, analyses, or any of a number of other professional tasks, you will likely find that, with the help of AI, you have become quite good.
>
> — **Ethan Mollick**

4:37 pm / llms, ai, ethan-mollick, generative-ai

---

page 1 / 2 next »

2023 » September

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**	