Simon Willison's Weblog Subscribe

February 2024

Search posts from February 2024 Search

79 posts: 4 entries, 62 links, 13 quotes

Feb. 1, 2024

teknium/OpenHermes-2.5 (via) The Nous-Hermes and Open Hermes series of LLMs, fine-tuned on top of base models like Llama 2 and Mistral, have an excellent reputation and frequently rank highly on various leaderboards.

The developer behind them, Teknium, just released the full set of fine-tuning data that they curated to build these models. It's a 2GB JSON file with over a million examples of high quality prompts, responses and some multi-prompt conversations, gathered from a number of different sources and described in the data card.

# 4:18 am / ai, generative-ai, llms, fine-tuning, nous-research, llm-release

---

Feb. 2, 2024

ChunkViz (via) Handy tool by Greg Kamradt to help understand how different text chunking mechanisms work by visualizing them. Chunking is an important part of preparing text to be embedded for semantic search, and thanks to this tool I've finally got a solid mental model of what recursive character text splitting does.

# 2:23 am / ai, embeddings

---

unstructured. Relatively new but impressively capable Python library (Apache 2 licensed) for extracting information from unstructured documents, such as PDFs, images, Word documents and many other formats.

I got some good initial results against a PDF by running "pip install 'unstructured[pdf]'" and then using the "unstructured.partition.pdf.partition_pdf(filename)" function.

There are a lot of moving parts under the hood: pytesseract, OpenCV, various PDF libraries, even an ONNX model—but it installed cleanly for me on macOS and worked out of the box.

# 2:47 am / ocr, pdf, python

> For many people in many organizations, their measurable output is words - words in emails, in reports, in presentations. We use words as proxy for many things: the number of words is an indicator of effort, the quality of the words is an indicator of intelligence, the degree to which the words are error-free is an indicator of care.
>
> [...] But now every employee with Copilot can produce work that checks all the boxes of a formal report without necessarily representing underlying effort.

— **Ethan Mollick**

# 3:34 am / ethan-mollick, ethics, generative-ai, ai, llms, ai-ethics

---

Samattical (via) Automattic (the company behind WordPress) have a benefit that's provided to all 1,900+ of their employees: a paid three month sabbatical every five years.

CEO Matt Mullenweg is taking advantage of this for the first time, and here shares an Ignite talk in which he talks about the way the benefit encourages the company to plan for 5% of the company to be unavailable at any one time, helping avoid any single employee becoming a bottleneck.

# 3:42 am / automattic, matt-mullenweg

> LLMs may offer immense value to society. But that does not warrant the violation of copyright law or its underpinning principles. We do not believe it is fair for tech firms to use rightsholder data for commercial purposes without permission or compensation, and to gain vast financial rewards in the process. There is compelling evidence that the UK benefits economically, politically and societally from upholding a globally respected copyright regime.

— **UK House of Lords report on Generative AI**

# 3:54 am / politics, ethics, generative-ai, ai, llms, ai-ethics, law

---

Open Language Models (OLMos) and the LLM landscape (via) OLMo is a newly released LLM from the Allen Institute for AI (AI2) currently available in 7b and 1b parameters (OLMo-65b is on the way) and trained on a fully openly published dataset called Dolma.

The model and code are Apache 2, while the data is under the “AI2 ImpACT license”.

From the benchmark scores shared here by Nathan Lambert it looks like this may be the highest performing model currently available that was built using a fully documented training set.

What’s in Dolma? It’s mainly Common Crawl, Wikipedia, Project Gutenberg and the Stack.

# 4:11 am / ai, generative-ai, llms, training-data, ai2, llm-release

---

Feb. 3, 2024

The Engineering behind Figma’s Vector Networks (via) Fascinating post by Alex Harri (in 2019) describing FIgma’s unique approach to providing an alternative to the classic Bézier curve pen tool. It includes a really clear explanation of Bézier curves, then dives into the alternative, recent field of vector networks which support lines and curves between any two points rather than enforcing a single path.

# 11:08 pm / graphics, bezier

Introducing Nomic Embed: A Truly Open Embedding Model. A new text embedding model from Nomic AI which supports 8192 length sequences, claims better scores than many other models (including OpenAI’s new text-embedding-3-small) and is available as both a hosted API and a run-yourself model. The model is Apache 2 licensed and Nomic have released the full set of training data and code.

From the accompanying paper: “Full training of nomic-embed-text-v1 can be conducted in a single week on one 8xH100 node.”

# 11:13 pm / ai, embeddings, nomic

Feb. 4, 2024

> Rye lets you get from no Python on a computer to a fully functioning Python project in under a minute with linting, formatting and everything in place.
>
> [...] Because it was demonstrably designed to avoid interference with any pre-existing Python configurations, Rye allows for a smooth and gradual integration and the emotional barrier of picking it up even for people who use other tools was shown to be low.

— **Armin Ronacher**

# 3:12 pm / armin-ronacher, python, rye

---

llm-sentence-transformers 0.2. I added a new --trust-remote-code option when registering an embedding model, which means LLM can now run embeddings through the new Nomic AI nomic-embed-text-v1 model.

# 7:39 pm / plugins, projects, transformers, ai, embeddings, llm, nomic

---

> Sometimes, performance just doesn't matter. If I make some codepath in Ruff 10x faster, but no one ever hits it, I'm sure it could get some likes on Twitter, but the impact on users would be meaningless.
>
> And yet, it's good to care about performance everywhere, even when it doesn't matter. Caring about performance is cultural and contagious. Small wins add up. Small losses add up even more.

— **Charlie Marsh**

# 7:41 pm / performance, ruff, charlie-marsh

Feb. 5, 2024

How does Sidekiq really work? (via) I really like this category of blog post: Dan Svetlov took the time to explore the Sidekiq message queue’s implementation and then wrote it up in depth.

# 5:20 pm / queues, ruby, sidekiq

---

shot-scraper 1.4. I decided to add HTTP Basic authentication support to shot-scraper today and found several excellent pull requests waiting to be merged, by Niel Thiart and mhalle.

1.4 adds support for HTTP Basic auth, custom --scale-factor shots, additional --browser-arg arguments and a fix for --interactive mode.

# 11:11 pm / projects, shot-scraper

Feb. 6, 2024

scriptisto (via) This is really clever. “scriptisto is tool to enable writing one file scripts in languages that require compilation, dependencies fetching or preprocessing.”

You start your file with a “#!/usr/bin/env scriptisto” shebang line, then drop in a specially formatted block that tells it which compiler (if any) to use and how to build the tool. The rest of the file can then be written in any of the dozen-plus included languages... or you can create your own template to support something else.

The end result is you can now write a one-off tool in pretty much anything and have it execute as if it was a single built executable.

# 4:49 am / programming

The power of two random choices, visualized. Grant Slatton shares a visualization illustrating “a favorite load balancing technique at AWS”: pick two nodes at random and then send the task to whichever of those two has the lowest current load score.

Why just two nodes? “The function grows logarithmically, so it’s a big jump from 1 to 2 and then tapers off *real* quick.”

# 10:21 pm / aws, load-balancing, scaling

SQL for Data Scientists in 100 Queries. New comprehensive SQLite SQL tutorial from Greg Wilson, author of Teaching Tech Together and founder of The Carpentries.

# 11:08 pm / greg-wilson, sql, sqlite

Feb. 7, 2024

Datasette 1.0a8: JavaScript plugins, new plugin hooks and plugin configuration in datasette.yaml

I just released Datasette 1.0a8. These are the annotated release notes.

[... 1,709 words]

4:37 pm / plugins, projects, datasette, annotated-release-notes

> If your only way of making a painting is to actually dab paint laboriously onto a canvas, then the result might be bad or good, but at least it’s the result of a whole lot of micro-decisions you made as an artist. You were exercising editorial judgment with every paint stroke. That is absent in the output of these programs.
>
> — **Neal Stephenson**

# 5:04 pm / neal-stephenson, generative-ai

Feb. 8, 2024

Google’s Gemini Advanced: Tasting Notes and Implications. Ethan Mollick reviews the new Google Gemini Advanced—a rebranded Bard, released today, that runs on the GPT-4 competitive Gemini Ultra model.

“GPT-4 [...] has been the dominant AI for well over a year, and no other model has come particularly close. Prior to Gemini, we only had one advanced AI model to look at, and it is hard drawing conclusions with a dataset of one. Now there are two, and we can learn a few things.”

I like Ethan’s use of the term “tasting notes” here. Reminds me of how Matt Webb talks about being a language model sommelier.

# 3:10 pm / google, ai, generative-ai, gpt-4, bard, llms, ethan-mollick, gemini

The first four Val Town runtimes (via) Val Town solves one of my favourite technical problems: how to run untrusted code in a safe sandbox. They're on their fourth iteration of this now, currently using a Node.js application that launches Deno sub-processes using the node-deno-vm npm package and runs code in those, taking advantage of the Deno sandboxing mechanism and terminating processes that take too long in order to protect against `while(true)` style attacks.

# 6:38 pm / javascript, nodejs, sandboxing, deno, tom-macwright, val-town

Feb. 9, 2024

“Wherever you get your podcasts” is a radical statement. Anil Dash points out that podcasts are one of the few cases where the dream really did work out:

“[...] what it represents is the triumph of exactly the kind of technology that’s supposed to be impossible: open, empowering tech that’s not owned by any one company, that can’t be controlled by any one company, and that allows people to have ownership over their work and their relationship with their audience.”

# 5:18 am / anil-dash, podcasts, rss, web-standards

Figure out who’s leaving the company: dump, diff, repeat (via) Rachel Kroll describes a neat hack for companies with an internal LDAP server or similar machine-readable employee directory: run a cron somewhere internal that grabs the latest version and diffs it against the previous to figure out who has joined or left the company.

I suggest using Git for this - a form of Git scraping - as then you get a detailed commit log of changes over time effectively for free.

I really enjoyed Rachel's closing thought:

> Incidentally, if someone gets mad about you running this sort of thing, you probably don't want to work there anyway. On the other hand, if you're able to build such tools without IT or similar getting "threatened" by it, then you might be somewhere that actually enjoys creating interesting and useful stuff. Treasure such places. They don't tend to last.

# 5:44 am / git, git-scraping, rachel-kroll

How I write HTTP services in Go after 13 years (via) Useful set of current best practices for deploying HTTP servers written in Go. I guess Go counts as boring technology these days, which is high praise in my book.

# 8:40 pm / go

Weeknotes: a Datasette release, an LLM release and a bunch of new plugins

I wrote extensive annotated release notes for Datasette 1.0a8 and LLM 0.13 already. Here’s what else I’ve been up to this past three weeks.

[... 1,074 words]

11:59 pm / projects, datasette, weeknotes, shot-scraper, llm, quickjs, enrichments

Feb. 10, 2024

(Almost) Every infrastructure decision I endorse or regret after 4 years running infrastructure at a startup (via) Absolutely fascinating post by Jack Lindamood talking about services, tools and processes used by his startup and which ones turned out to work well v.s. which ones are now regretted.

I’d love to see more companies produce lists like this.

# 5:51 am / architecture, infrastructure, startups

> Reality is that LLMs are not AGI -- they're a big curve fit to a very large dataset. They work via memorization and interpolation. But that interpolative curve can be tremendously useful, if you want to automate a known task that's a match for its training data distribution.
>
> Memorization works, as long as you don't need to adapt to novelty. You don't *need* intelligence to achieve usefulness across a set of known, fixed scenarios.

— **François Chollet**

# 6:39 am / llms, ai, generative-ai, francois-chollet

Rye: Added support for marking virtualenvs ignored for cloud sync (via) A neat feature in the new Rye 0.22.0 release. It works by using an xattr Rust crate to set the attributes “com.dropbox.ignored” and “com.apple.fileprovider.ignore#P” on the folder.

# 6:50 am / python, dropbox, rust, rye

Feb. 11, 2024

Python Development on macOS Notes: pyenv and pyenv-virtualenvwrapper (via) Jeff Triplett shares the recipe he uses for working with pyenv (initially installed via Homebrew) on macOS.

I really need to start habitually using this. The benefit of pyenv over Homebrew’s default Python is that pyenv managed Python versions are forever—your projects won’t suddenly stop working in the future when Homebrew changes its default Python version.

# 4:41 am / macos, python, jeff-triplett

> One consideration is that such a deep ML system could well be developed outside of Google-- at Microsoft, Baidu, Yandex, Amazon, Apple, or even a startup. My impression is that the Translate team experienced this. Deep ML reset the translation game; past advantages were sort of wiped out. Fortunately, Google's huge investment in deep ML largely paid off, and we excelled in this new game. Nevertheless, our new ML-based translator was still beaten on benchmarks by a small startup. The risk that Google could similarly be beaten in relevance by another company is highlighted by a startling conclusion from BERT: huge amounts of user feedback can be largely replaced by unsupervised learning from raw text. That could have heavy implications for Google.

— **Eric Lehman,** internal Google email in 2018

# 10:59 pm / machine-learning, translation, google, generative-ai, ai, llms, bert

---

page 1 / 3 next » last »»

2024 » February

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			