Developments in AI, Programming, Web, Security, Virtual and Augmented Reality, and Quantum Computing

More large language models. Always more large language models. Will the new year be any different? But there is a difference in this month's AI news: there's an emphasis on tools that make it easy for users to use models. Whether it's just tweaking a URL so you can ask questions of a paper on arXiv or using LLamafile to run a model on your laptop (make sure you have a lot of memory!) or using the Notebook Language Model to query your own documents, AI is becoming widely accessible–and not just a toy with a web interface.

Artificial Intelligence

- Adding talk2 to the start of any arXiv URL (e.g., talk2arxiv.org) loads the paper into an AI chat application so you can talk to it. This is a very clever application of the RAG pattern.
- Google's Autonomous Vehicle startup, Waymo, has reported a total of three minor injuries to humans in over 7 million miles of driving. This is clearly not Tesla, not Uber, not Cruise.
- Google's DeepMind has used a large language model to solve a previously unsolved problem in mathematics. This is arguably the first time a language model has created information that didn't previously exist.
- The creator of llamafile has offered a set of one-line bash scripts for laptop-powered AI.
- Microsoft has released a small language model named Phi-2. Phi-2 is a 2.7B parameter model that has been trained extensively on "textbook-quality data." Without naming names, they claim performance superior to Llama 2.
- Claude, Anthropic's large language model, can be used in Google Sheets via a browser extension.
- The Notebook Language Model is a RAG implementation designed for individuals. It is a Google notebook (similar to Colab or Jupyter) that allows you to upload documents and then ask questions about those documents.
- The European Union is about to pass its AI Act, which will be the world's most significant attempt to regulate artificial intelligence.
- Mistral has released Mixtral 8x7B, a mixture-of-experts model in which the model first determines which of eight sets of 7 billion parameters will generate the best response to a prompt. The results compare well to Llama 2. Mistral 7B and Mixtral can be run with Llamafile.
- Meta has announced Purple Llama, a project around trust and safety for large language models. They have released a set of benchmarks for evaluating model safety, along with a classifier for filtering unsafe input (prompts) and model output.
- The Switch Kit is an open source software development kit that allows you to replace OpenAI with an open source language model easily.
- Google has announced that its multimodal Gemini AI model is available to software developers via their AI Studio and Vertex AI.
- Progressive upscaling is a technique for starting with a low-resolution image and using AI to increase the resolution. It reduces the computational power needed to generate high-resolution images. It has

been implemented as a plug-in to Stable Diffusion called DemoFusion.

- The internet enabled mass surveillance, but that still leaves you with exabytes of data to analyze. According to Bruce Schneier, AI's ability to analyze and draw conclusions from that data enables "mass spying."
- A group of over 50 organizations, including Meta, IBM, and Hugging Face, has formed the AI Alliance to focus on the development of open source models.
- DeepMind has built an AI system that demonstrates social learning: the ability to learn how to solve a problem by observing an expert.
- Are neural networks the only way to build artificial intelligence? Hivekit is building tools for a distributed spatial rules engine that can provide the communications layer for hives, swarms, and colonies.
- The proliferation of AI testing tools continues with Gaia, a benchmark suite intended to determine whether AI systems are, indeed, intelligent. The benchmark consists of a set of questions that are easy for humans to answer but difficult for computers.
- Meta has just published a suite of multilingual spoken language models called Seamless. The models are capable of near real-time translation and claim to be more faithful to natural human expression.
- In an experiment simulating a stock market, a stock-trading AI system engaged in "insider trading" after being put under pressure to show greater returns and receiving "tips" from company "employees."
- What's the best way to run a large language model on your laptop? Simon Willison recommends llamafile, which packages a model together with the weights as a single (large) executable that works on multiple operating systems.
- Further work on extracting training data from ChatGPT, this time against the production model, shows that these systems may be opaque, but they aren't quite "black boxes."
- Amazon Q is a new large language model that includes a chatbot and other tools to aid office workers. It can be customized by individual businesses that subscribe to the service so that it has access to their proprietary data.

Programming

- A new language superset: Pluto is a superset of Lua. Supersetting may be the "new thing" in language design: TypeScript, Mojo, and a few others (including the first versions of C++) come to mind.
- Virtualization within containers orchestrated by Kubernetes: Can you imagine a Kubernetes cluster running within a Docker container? Is that a good thing or evidence of how a stack's complexity can grow without bounds?
- Google engineers propose an alternative to microservices: limited monoliths that are deployed by an automated runtime that determines where and when to instantiate them. As Kelsey Hightower said, deployment architecture becomes an implementation detail.
- The OpenBao project is intended to be an open source fork of HashiCorp's Vault, analogous to the OpenTofu fork of Terraform. There is speculation that IBM will back both projects.
- Biscuit authorization is a distributed authorization protocol that is relatively small, flexible, and is designed for use in distributed systems. Any node can validate a Biscuit token using only public information.
- gokrazy is a minimal Go runtime environment for the Raspberry Pi and (some) PCs. It minimizes maintenance by eliminating everything that isn't needed to compile and run Go programs.
- You very clearly don't need this: A Brainfuck interpreter written in PostScript. (If you really must know, Brainfuck is arguably the world's most uncomfortable programming language, and PostScript is the lan-

guage your computer sends to a printer.)

- Baserow is a no-code, open source tool that combines a spreadsheet with a database. It's similar to Airtable.
- New programming language of the month: Onyx is a new programming language designed to generate WebAssembly (Wasm), using Wasmer as the underlying runtime.

Web

- Anil Dash predicts that the internet is about to get weird again–the way it should be. Power is shifting from the entrenched, heavily funded "walled gardens" and back to people who just want to be creative.
- Meta's Threads has begun to test integration with ActivityPub, which will make it accessible to Mastodon servers.
- The HTML Energy movement attempts to reclaim the creativity of the early web by building sites from scratch with HTML and abandoning high-powered web frameworks.
- The best WebAssembly runtime might be no runtime at all: just transpile it to C.

Security

- Researchers have discovered a man-in-the-middle attack against SSH, one of the foundations of cybersecurity.
- A new version of SSH (SSH3) promises to be faster and more feature-rich. It is based on HTTP/3 and written in Go.
- Security researchers have demonstrated two important vulnerabilities in OpenAI's custom GPTs. Malicious actors can extract system prompts, and they can force it to leak uploaded files and other data.
- Meta has made end-to-end encryption (E2EE) the default for all users of Messenger and Facebook messaging. Their E2EE implementation is based on Signal's. They have built a new storage and retrieval service for encrypted messages.
- A chatbot driven by a jailbroken language model can be used to jailbreak other chatbots. Language models are very good at coming up with prompts that get other models to go outside their boundaries, with success rates of 40% to 60%. AI security will be a key topic this year.

Quantum Computing

- IBM has developed a 1121 qubit quantum processor, along with a system built from three 133 qubit processor chips that greatly improves the accuracy of quantum gates. Working quantum computers will probably require over a million qubits, but this is a big step forward.
- A research group has announced that it can perform computations on 48 logical (i.e., error-corrected) qubits. While there are a number of limitations to their work, it's an important step toward practical quantum computing.
- Two posts about post-quantum cryptography explain what it's about.

Brains

- Researchers have developed a noninvasive system that can turn human thought into text. Users wear a cap with sensors that generates EEG data. Accuracy isn't very high yet, but it is already superior to other thought-to-speech technologies.
- Artificial neural networks with brains: Researchers connected cultured human brain cells (organoids) to an interface that allowed them to give the organoids audio data. They found that it was able to recognize vowel sounds.

Virtual and Augmented Reality

- OpenUSD is an open source standard for scene representation that could enable a real metaverse, not the proprietary walled garden imagined by last year's metaverse advocates