Developments in Networks Biology Robotics and More

The Security section has grown almost as large as AI (and longer than Programming)–and that's not including some security issues specific to AI, like model leeching. Does that mean that AI is cooling down? Or that security is heating up? It's really impossible for security issues to get too much attention. The biggest news in AI arrived on the last day of October, and it wasn't technical at all: the Biden administration's executive order on AI. It will take some time to digest this, and even longer to see whether vendors follow the order's recommendations. In itself, it's evidence of an important ongoing trend: in the next year, many of the most important developments in AI will be legal rather than technical.

Artificial Intelligence

• In an executive order, the US has issued a set of rules covering the development of advanced AI systems. The regulations encourage the development of watermarks (specifically the C2PA initiative) to authenticate communication; they attempt to set standards for testing; and they call for agencies to develop rules to protect consumers and workers.

• Nightshade is another tool that artists can use to prevent generative AI systems from using their work. It makes unnoticeable modifications to the image that cause the AI model to misinterpret it and create incorrect output.

• Stanford's Institute for Human-Centered Artificial Intelligence has issued a report on transparency for large language models: whether the creators of LLMs are disclosing essential data about their models. No model scores well, and transparency appears to be declining as the field grows more competitive.

• Chatbots perpetuate false and racially biased information in medical care. Debunked ideas about pain tolerance, kidney function, and other factors are included in training data, causing models to repeat those ideas.

• An AI Bill of Materials (AIBOM) would document all of the materials that go into the creation of an AI system. This documentation would be essential to building AI that is capable of complying with regulation.

• GPT-4 does Stephenson: GPT simulates the Young Lady's Illustrated Primer (from The Diamond Age). With illustrations from DALL-E.

• Step-Back Prompting is another prompting technique in which you ask a question, but before getting an answer, you ask the LLM to provide background information that will help it answer the question.

• Prompt injection just got scarier. GPT-4V, which allows users to include images in conversations, is vulnerable to prompt injection through the images themselves; text in the images can be interpreted as prompts. Malicious prompts can even be hidden in images.

• Google joins Microsoft, Adobe, and others in indemnifying users of their AI against copyright lawsuits.

• Model leeching is a new attack against large lan-

guage models. In model leeching, a carefully constructed set of prompts allows attackers to generate a smaller model that behaves similarly. The smaller model can then be used to construct other attacks against the original model.

• Open source language models are proliferating. Replit Code v1.5 3B is now available on Hugging Face. This model is designed for code completion, and has been trained on permissively licensed code so there should be minimal legal issues.

• Anthropic appears to have made significant progress in making large language models interpretable. The key is understanding the behavior of groups of neurons, which they call "features," rather than individual neurons.

• Mistral 7B is an open source large language model with impressive performance. It was developed independently. (It is not related to LLaMA.) Its performance is claimed to be better than equivalently sized models.

• AMD may be able to challenge NVIDIA's dominance of the GPU market. NVIDIA's dominance relies on the widely used CUDA language for programming GPUs. AMD has developed a version of PyTorch that has been tuned for use on AMD GPUs, eliminating the need for low-level GPU programming.

• Larger training datasets leads to more biased and hateful output, not less.

• LangStream (unrelated to LangChain) is an open source platform for building streaming applications that use generative AI.

• GPT-4 and Claude have proven useful in translating 16th century demonology texts written in Medieval Latin. Claude's 100K context window appears to be a big help. (And Medieval Latin is much different from the Latin you probably didn't learn in school.)

• A vulnerability called ShellTorch allows attackers to gain access to AI servers using TorchServe, a tool for deploying and scaling AI models using PyTorch.

• Reservoir computing is another kind of neural network that has promise for understanding chaotic systems.

• Perhaps not surprisingly, language models can do an excellent job of lossless compression better than standards like FLAC. (This doesn't mean that language models store a compressed copy of the web.)

• An artist makes the case that training generative models not to "hallucinate" has made them less interesting and less useful for creative applications.

• Can you melt eggs? Quora has included a feature that generates answers using an older GPT model. This model answered "yes," and aggressive SEO managed to get that "yes" to the top of a Google search.

Programming

• Harpoon is a no-code, drag and drop tool for Kubernetes deployment.

• Cackle is a new tool for the Rust tool chain. It checks access control lists and is used to make software supply chain attacks more difficult.

• Correctness SLOs (Service-Level Objectives) are a way to specify the statistical properties of a program's output if it is running properly. They could become important as AI is integrated into more applications.

• Cilium is a tool for cloud native network observability. It provides a layer on top of eBPF that solves security and observability problems for Docker and Kubernetes workloads.

• The Six Pillars of Platform Engineering is a great

start for any organization that is serious about developer experience. The pillars are Security, Pipelines, Provisioning, Connectivity, Orchestration, and Observability. One article in this series is devoted to each.

• Adam Jacob, creator of Chef Software, is out to reimagine DevOps. System Initiative is an open source tool for managing infrastructure that stresses collaboration between engineers and operations staff–something that was always the goal of DevOps but rarely achieved.

• Unreal engine, a game development platform that had been free for users outside of the gaming industry, will now have a subscription fee. It will remain free for students and educators.

• CRDTs (conflict-free replicated data types) are a data structure that is designed for resolving concurrent changes in collaborative applications (like Google Docs). Here's a good interactive tutorial and a project: building a collaborative pixel editor.

• Ambient is a purely web-based platform for multiplayer games, built with Wasm, WebGPU, and Rust. Instant deployment, no servers.

• Google has open sourced its graph mining library. Graphs are becoming increasingly important in data mining and machine learning.

• Microsoft has released a binary build of OpenJDK 21, presumably optimized for Azure. Shades of Embrace and Extend? That doesn't appear to be happening.

• Polystores can store many different kinds of data–relational data, vector data, unstructured data, graph data–in a single data management system.

Security

• The EFF has posted an excellent introduction to passkeys, which are the next step past passwords in user authentication.

• Microsoft has started an early access program for Security Copilot, a chatbot based on GPT-4 that has been tuned to answer questions about computer security. It can also summarize data from security incidents, analyze data from new attacks, and suggest responses.

• Google is planning to test IP protection in Chrome. IP protection hides users' IP addresses by routing traffic to or from specific domains through proxies. Address hiding prevents a number of common attacks, including cross-site scripting.

• While the European Cyber Resilience Act (CRA) has many good ideas about making software more secure, it puts liability for software flaws on open source developers and companies funding open source development.

• A new attack against memory, called RowPress, can cause bitflips even in DDR4 memory, which already incorporates protections against the RowHammer attack.

• August and September's distributed denial of service attacks (DDOS) against Cloudflare and Google took advantage of a newly discovered vulnerability in HTTP/2. Attackers open many streams per request, creating extremely high utilization with relatively few connections.

• Mandiant has provided a fascinating analysis of the Russian military intelligence's (GRU's) playbook in Ukraine.

• Mozilla and Fastly are developing OHTTP (Oblivious HTTP), a successor to HTTP that has been designed for privacy. OHTTP separates information about the requestor from the request itself, so no single party ever has both pieces of information.

• A newly discovered backdoor to WordPress allows attackers to take over websites. The malware is disguised as a WordPress plug-in that appears legitimate.

• While standards are still developing, decentralized identity and verifiable credentials are starting to appear outside of the cryptocurrency world. When adopted, these technologies will significantly enhance both privacy and security.

• To improve its ability to detect unwanted and harmful email, Gmail will be requiring bulk email senders (over 5,000 messages per day) to implement SPF, DKIM, and DMARC authentication records in DNS or risk having their messages marked as spam.

• Genetic data has been stolen from 23andMe. The attack was quite simple: the attackers just used usernames and passwords that were in circulation and had been reused.

• The time required to execute a ransomware attack has reduced from 10 days to 2 days, and it's increasingly common for victims to be hit with a second attack against systems that have already been compromised.

Networks

• Toxiproxy is a tool for chaos network engineering. It is a proxy server that simulates many kinds of network misbehavior.

• Network neutrality rises again: The chair of the FCC has proposed returning to Obama-era network neutrality rules, in which carriers couldn't prioritize traffic from some users in exchange for payment. Laws in some states, such as California, have largely prevented traffic prioritization, but a return of network neutrality would provide a uniform regulatory framework.

• Most VPNs (even VPNs that don't log traffic) track user activity. Obscura is a new VPN that was designed for privacy, and that cannot track activity.

Biology

• The US Fish & Wildlife Service is creating a biodiversity library. The library's goal is to preserve tissue samples from all endangered species in the US. The animals' DNA will be sequenced and uploaded to GenBank, a collection of all publicly available DNA sequences.

Quantum Computing

• Atom Computing claims to have built a 1,000 qubit quantum computer. While this is still too small to do real work, it's the largest quantum computer we know about; it looks like it can scale to (somewhat) larger sizes; and it doesn't require extreme cold.

• Two research teams have made progress in quantum error correction. Lately, we've seen several groups reporting progress in QEC, which is key to making quantum computing practical. Will this soon be a solved problem?

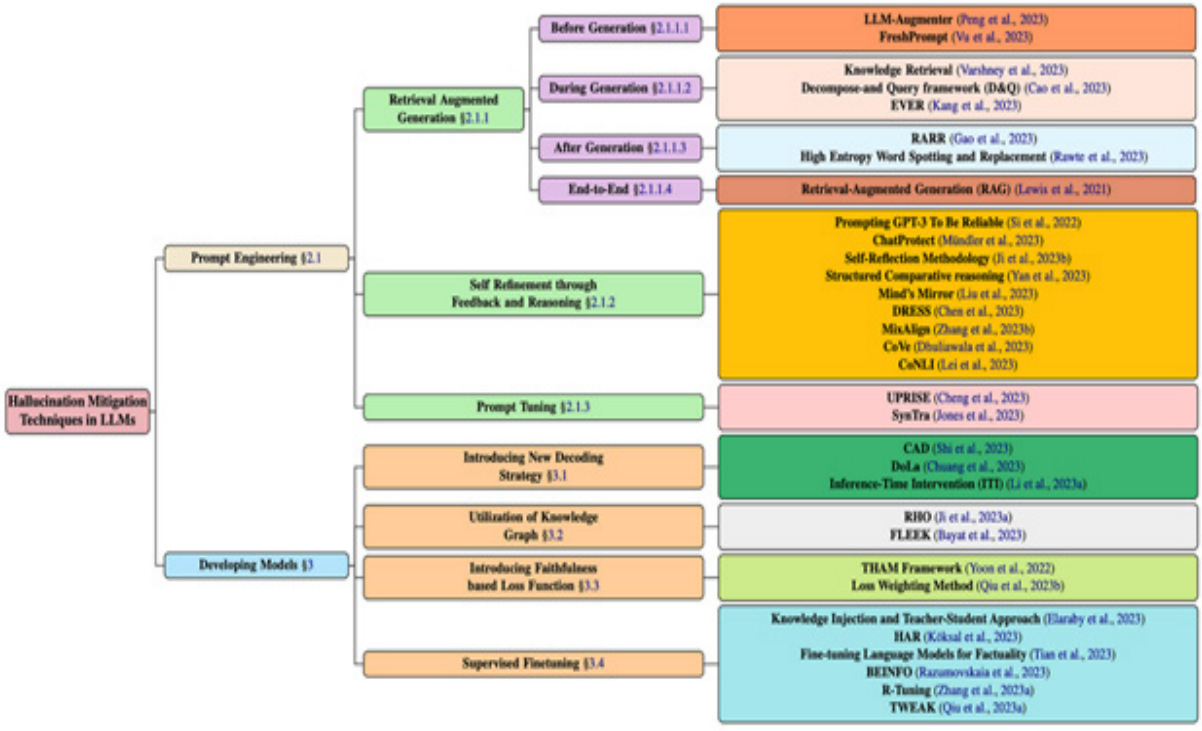
Robotics

• This article's title is all you need: Boston Dynamics turned its robotic dog into a walking tour guide using ChatGPT. It can give a tour of Boston Dynamics' facilities in which it answers questions, using data from its cameras to provide added context. And it has a British accent.

• Another autonomous robotic dog can plan and execute actions in complex environments. While its agility is impressive, what sets it apart is the ability to plan actions to achieve a goal, taking into account the objects that it sees.

• A tetrahedral robot is able to change its shape and size, use several different styles of walking, and adapt itself to different tasks

Large Language Model Hallucination Mitigation Techniques



This recently released study is a comprehensive survey of 32+ mitigation techniques to address hallucination.

1. Introduction

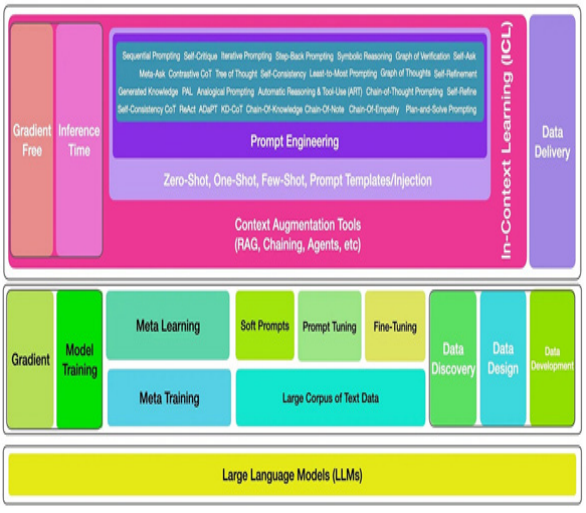
The techniques are broken down into two main streams, gradient and non-gradient approaches. Gradient approaches refers to fine-tuning the base LLM. While non-gradient approaches involves prompt engineering techniques which are delivered at inference.

Most notable are the inclusion of:

1. Retrieval-Augmented Generation (RAG)
2. Knowledge Retrieval
3. CoNLI
4. CoVe

Hallucination mitigation in LLMs represents a multifaceted challenge addressed through a spectrum of innovative techniques.

LLMs & Data Delivery



Unlike traditional AI systems focused on limited tasks, LLMs have been exposed to vast amounts of online text data during training.

This allows LLMs to display impressive language fluency, it also means they are capable of:

1. Extrapolating information from the biases in training data,

2. Misinterpreting ambiguous prompts, or modifying the information to align superficially with the input.

This becomes hugely alarming when language generation capabilities are used for sensitive applications, such as:

1. Summarising medical records,

2. Customer support conversations,

3. Financial analysis reports, and providing erroneous legal advice.

2. Hallucination Mitigation Taxonomy

The study includes very insightful taxonomy of hallucination mitigation techniques for LLMs; both gradient and non-gradient.

Gradient approaches include complex and opaque decoding strategies, knowledge graphs, fine-tuning strategies and more.

Non-gradient approaches include RAG, Self-Refinement and prompt tuning.

Notably the RAG approaches are segmented into four parts;

1. Before Generation

2. During Generation

3. After Generation

4. End-to-End

The power of prompt engineering to mitigate hallucination lies in defining:

1. Specific context &

2. Expected Outcomes

3. The Best Defence

The best defence against hallucination is not one single approach or method, but a combined approach defending against hallucination.

Seamlessly integrating numerous mitigation approaches, is the most important takeaway.

3. The factors which any organisation should keep in mind are:

1. To what extent is there a reliance on labeled data?

2. What are the possibilities of introducing unsupervised or weak-supervised learning techniques to improve scalability and flexibility?

3. Consideration of gradient and non-gradient approaches to produce coherent and contextually relevant information.

4. The collected works on hallucination mitigation reveal a diverse array of strategies, each contributing uniquely to address the nuances of hallucination in LLMs.

5. Self-refinement through feedback and reasoning brings forth impactful strategies.

6. Structured Comparative reasoning introduces a structured approach to text preference prediction, enhancing coherence and reducing hallucination.

7. Supervised fine-tuning can be explored via Knowledge Injection and Teacher-Student Approaches.

Domain-specific knowledge is injected into weaker LLMs and approaches that employ counter factual datasets for improved factuality.