LLMs in Security: Today + What’s Coming

UC Berkeley CS294/194-196: Responsible GenAI and Decentralized Intelligence
October 3rd, 2023

Leo Meyerovich, Graphistry Inc.
Thank you for feedback:

Fabrizio Cili, Dawn Song, Matei Zaharia, Dan Guido, Mahendra Kutare, & others

+ @ Graphistry: Alex Morrise, Alex Warren, Simon Abizmil, Cody Webb
Poll Everywhere:
← Who is he?
Bruce Schneier on Security

- Invented: Blowfish, twofish, …
- Startups: Counterpane, Resilient, …
- Advisory: EFF, …

PROMPT: Bruce schneier as an 80's action film star, happily riding a galloping unicorn and firing a large chain gun, dramatic and highly detailed. Oriented to the right.
The NSA is actually a backdoor inserted into government by Bruce Schneier

Submitted by Simidgey

permalink | vote

Bruce Schneier photos doctored by John Leach and Mortz Schlaboosk.
Bruce Schneier's secure handshake is so strong, you won't be able to exchange keys with anyone else for days.
Society would be safer if everyone had a SchneierGPT looking over their shoulder.
Bruce Schneier can read and understand Perl programs.

Submitted by Stefan Parvisinen

permalink | vote

Bruce Schneier photos donated by John Leech and Moritz Schallaböck.
Demo of louie autogpt explaining Perl

https://www.loom.com/spaces/All-Graphistry-1228153/folders/berkeley-194-d905f3da2df440c2b5e9e1c6824d6ab5
Bruce Schneier can calculate MD5 hashes in his head. For any length of data. In constant time. Drunk.

Submitted by Anonymous

permalink | vote

Bruce Schneier photos donated by John Leach and Marilla Schachterlock.
Demo of louie autogput generating an md5 sum

https://www.loom.com/spaces/All-Graphistry-1228153/folders/berkeley-194-d905f3da2df440c2b5e9e1c6824d6ab5
Day job: Analytics CEO

GRAPHISTRY
Visual graph AI platform:
First end-to-end GPU

Louie.AI: GenAI-first interactive
& autonomous analytics

Pentagon, UK/AU/US govs, startups, ...
**Find a professor:** Fun impactful ugrad + grad projects

- **Margrave:** XACML Verification and Change-Impact Analysis
  - Kathy Fader, WPI
  - Sheharn Krishnamurthi, Brown
  - Leo Meyerovich, Brown
  - Michael Cai Tocharas, Brown

- **AWS Cedar**
  - Explaining & verifying rich security policies

- **FLAPJAX**
  - Test of time award

- **Berkeley parallel browser + GPU DF.js**

- **Superconductor**
  - Information is Beautiful award

- **React**
  - Functional reactive web programming (FRP)

- **SERQO**
  - Multicore/GPU web

- **RAPIDS**
  - GPU dataframes
LLMs in security

- Escalating cyber attacks
- New era of cyber defense
- Business of security
- Compliance
- Market
- CISO Priorities
- The fight beyond IT
- Bots & misinformation
- Emergency services
- Intelligence community
- Controversial government operations & policy
- Military
- Free speech
- Authoritarian societies
- Today + next few years
- From technology first principles

Real-world data + tech demos
LLMs in security

Escalating cyber attacks

New era of cyber defense

Business of security
Compliance
Market
CISO Priorities

The fight beyond IT
Bots & misinformation
Emergency services
Intelligence community

Controversial government operations & policy
Military
Free speech
Authoritarian societies

Today + next few years
From technology first principles

Real-world data + tech demos
Attackers perspective: Hacking industrialized

Common attack types against common surfaces for common goals

Paid & OSS hacking tools
Phishkits, DDOS, …

Assembly line:
Drop in new 0-day exploits, payloads, …

https://therecord.media/microsoft-uncovers-giant-phishing-as-a-service-operation
Attackers double down on successful attack vectors & attack surfaces

Examples:

- Ransomware blew up because bitcoin
- Phishing => auto-draining e-banking accounts
- Malware => recruitment into botnet => sell DDoS
- Password stuffing => credit card testing => theft

Denial of service
Zayo: 2010-2021

Ransomware
Coveware: 2018-2020

Credential stuffing
Okta: 2022

Malware & phishing
Google Safe Browsing: 2007-2019
So where do LLMs fit into today’s cybersecurity attacker’s pipeline?

Phishkit generation
FraudGPT, WormGPT, … - helps foreign attackers writing convincingly text, generate web app code, …

Lure generation
Deepfakes, identities, websites, …

Malware generation
Translate CVEs ⇒ active exploits; rewrite code to defeat signature scans

Early days as existing alternatives easier
Ex: password stuffing from password DB leaks

Overall low use in classic popular IT attack types

Other categories are growing: later slides

Imminent threat of rapid escalation (months/years)
Initial vector: Impersonation
Escalation: Automation of API & social interactions
Figure: Twitter bots running on ChatGPT

Platform abuse plaguing “trust & safety” teams everywhere
Active non-IT categories: Platform abuse & identity theft - SEO spam

6 Ways SEO Pros Are Using ChatGPT Right Now

New AI-powered chatbot shows great potential for revolutionizing search engine optimization. Here are some ways it's already being used.

SEJ STAFF
Brian Frederick
January 17, 2023 - 9 min read

465 SHARES 110K READS

Quora uses Chat-GPT to appear on every query in Google

Quora uses Chat-GPT now to answer people's questions and to appear for all queries in Google, sometimes with completely nonsense answers and Google is forced to post messages to say that the content of the article is false. This is idiocracy.

Example of keywords: Can you melt an egg?

Google Search's guidance about AI-generated content

Wednesday, February 8, 2023

At Google, we've long believed in the power of AI to transform the ability to deliver helpful information. In this post, we'll share more about how AI-generated content fits into our long-standing approach to show helpful content to people on Search.

Rewarding high-quality content, however it is produced

Google's ranking systems aim to reward original, high-quality content that demonstrates qualities of what we call E-E-A-T: expertise, experience, authoritativeness, and trustworthiness. We share more about this in our How Search Works site.

Our focus on the quality of content, rather than how content is produced, is a useful guide that has helped us deliver reliable, high-quality results to users for years.

LLM trains on web pages => SEO generates web pages from LLM => Google rewards with ad $

Is this the end of search?
Active non-IT categories: Platform abuse & identity theft - Site spam

Job board ruined by GPT autoreplies - every candidate is “perfect”!

Dear Hiring Team,

I’m writing to express my interest in your Security Engineer position focusing on hardening your untrusted code execution sandbox system. My background perfectly aligns with the technical and collaborative aspects required for this role:

1. Sandbox Hardening: At Company XYZ, I led the security architecture overhaul of their container-based sandbox environment, successfully mitigating multiple known vulnerabilities. Specifically, I used Docker’s seccomp profiles and Linux’s cgroup policies to tighten sandbox configurations.

2. R&D Collaboration: In my role at ABC Corp, I served as the liaison between the R&D and Security teams, helping to incorporate security best practices into the development cycle without compromising agility.

3. Security Testing & CI/CD: I built a comprehensive security test suite at DEF Inc., which was fully integrated into the CI/CD pipeline. This suite caught 95% of security flaws before they hit production, a 40% improvement over the previous setup.

4. Technical Proficiency: I regularly use strace and similar tools for runtime analysis of containers. I’m also versed in Python, bash, and Linux—particularly Ubuntu and RHEL flavors.

5. Communication Skills: My experience collaborating in geographically dispersed teams makes me comfortable with asynchronous work. I’ve published security research papers and led workshops, showcasing my written and verbal English communication skills.

Considering the project’s 2-4 week timeline, I am confident in delivering a robust, hardened sandbox environment that satisfies your security requirements. I’m familiar with Github, Slack, and am flexible with meeting times for R&D project leads.

I look forward to the opportunity to contribute to this crucial project and potentially engage in long-term collaboration.

Best regards,
[Your Name]
Active non-IT categories: Platform abuse & identity theft - $ transfers

New
Few cases, but high $ ROI

Ex: Deepfakes subset of “CEO invoice scams”

Dangerous
Social engineering attacks are scary

Ex: Popular trusted Okta login vendor breached through helpdesk

Getting bad, fast
Audio, visual, & text technology improving + democratizing

Scalable & automatable attack vector
Need to protect against inauthentic activity

Increasingly hard to detect generated content, especially from determined attackers

Wreaking havoc on trust & safety teams: Reviews, comments, jobs, …

LLM-level solutions

- AI watermarks: Dying off? (OpenAI, …)
- Text analyzers: High FP (TurnItIn, …)

More effective today: Check hard-to-fake data

- Identity authentication
- Social network analysis: Group identity & individual behavior
- Digital forensics: Metadata mining
- Problem: Cyborgs
  Ex: QAnon Into the Storm (HBO)
  ○ Behavior analytics & metadata worked!
LLMs in security

- Escalating cyber attacks
- New era of cyber defense
- Business of security: Compliance, Market, CISO Priorities
- The fight beyond IT: Bots & misinformation, Emergency services, Intelligence community
- Controversial government operations & policy: Military, Free speech, Authoritarian societies
- Today + next few years: From technology first principles

Real-world data + tech demos
Defenders perspective: Representative Tier 2 SOC session

IP=10.16.0.8; msg=Malware.Object;
  time=2 Nov 2017 19:32:00 UTC;
  vendor=FireEye; Product=Web MPS NX
VirusTotal is a service that analyzes suspicious files and URLs and facilitates the quick detection of viruses, worms, trojans, and all kinds of malware detected by antivirus engines. More information.

<table>
<thead>
<tr>
<th>File name</th>
<th>cdbox_setup_1.3.7.2356.exe</th>
</tr>
</thead>
<tbody>
<tr>
<td>Submission date</td>
<td>2010-08-20 11:34:23 (UTC)</td>
</tr>
<tr>
<td>Current status</td>
<td>finished</td>
</tr>
<tr>
<td>Result</td>
<td>0/40 (0.0%)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Antivirus</th>
<th>Version</th>
<th>Last Update</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>AvgLab-V3</td>
<td>2010.06.20.01</td>
<td>2010.06.20</td>
<td>-</td>
</tr>
<tr>
<td>AntiVir</td>
<td>6.2.4.36</td>
<td>2010.06.20</td>
<td>-</td>
</tr>
<tr>
<td>Antiy-AVL</td>
<td>2.0.3.7</td>
<td>2010.06.16</td>
<td>-</td>
</tr>
<tr>
<td>Authentium</td>
<td>5.2.0.5</td>
<td>2010.06.20</td>
<td>-</td>
</tr>
<tr>
<td>Avast</td>
<td>4.8.1351.0</td>
<td>2010.06.20</td>
<td>-</td>
</tr>
</tbody>
</table>
… start over!
Slight (big) problem

With problems like alert fatigue, stress, and low levels of automation, most analysts burn out in their first year, and constant demand to “do more with less”
**Defenders perspective: $200B of defense in depth - surface keeps growing!**

<table>
<thead>
<tr>
<th>Shift-left</th>
<th>Security operations</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Vendor review</td>
<td>• Help desk</td>
</tr>
<tr>
<td>• Security design review</td>
<td>• Digital forensics &amp; incident response</td>
</tr>
<tr>
<td>• Threat modeling</td>
<td>• Patch management</td>
</tr>
<tr>
<td>• Code analysis</td>
<td>• Threat hunting &amp; threat intel</td>
</tr>
<tr>
<td></td>
<td>• Detection engineering</td>
</tr>
<tr>
<td></td>
<td>• Security awareness programs</td>
</tr>
<tr>
<td><strong>Production surface</strong></td>
<td></td>
</tr>
<tr>
<td>• Firewalls, load balancers, email, DNS</td>
<td></td>
</tr>
<tr>
<td>• Identity, access management, zero-trust</td>
<td></td>
</tr>
<tr>
<td>• Endpoint, network, cloud, shadow IT, SaaS</td>
<td></td>
</tr>
<tr>
<td>• Public site &amp; intranet</td>
<td></td>
</tr>
<tr>
<td>• Applications and their behavior</td>
<td></td>
</tr>
<tr>
<td>• Compliance &amp; audits</td>
<td></td>
</tr>
</tbody>
</table>

**Generative AI projects are rethinking every area listed here**
Blackbox tool vendors using LLMs for smarter detections. Ex: Phishing

URLs: DGA, lookalikes, ...

Phishing
- Fake websites
- False invoices
- Fake data requests: Legal, ...
- Fake text messages

Many vendors doing this

LLM classifier, but not necessarily generative (GPT)
SDLC & compliance examples

Security questionnaires
many, ex: vanta

Design review
ex: remysec

Code scans
many (see DARPA challenge)

Observation: Bots enable pushing smarter security discussions to end-users so they can help themselves and more autonomous workflows overall. Important for enterprise scale and SMB affordability & practicality.

REVERSE
Github Copilot generating ~half of new Github code, and question of code bugs & vulns
Top security analytics & automation platforms adding generative AI: SIEM, XDR, SOAR, DFIR, ...

**Siloed platforms**

**Cross-silo platforms**

**SOAR 2.0**
Data co-pilot; smart automation; autonomous investigation; threat intel handling; ...

**Detection engineering**
Easier, cover text columns, add smarter decision making; questions of scale, fidelity, model

**Upskilling:** Junior analysts, coding, ...
Drilldown: Smarter analyzers

Experiment - autonomous hunting by Graphistry / Louie.AI: US Gov challenge data
- ❌ GPT3-based autonomous strategy: ~0 findings within 10 queries
- 🥈 Automatic graph AI anomaly detector: ~80%
- 🏆 GPT4-based autonomous strategy: ~all hits in first 5 queries

Intuition: Similar to GPT passing the bar exam, GPT4 knows data source <> hunt strategies

[ Demo: https://www.loom.com/spaces/All-Graphistry-1228153/folders/berkeley-194-d905f3da2df440c2b5e9e1c6824d6ab5 ]

Frontier: Neurosymbolic synthesis
- Combining Z3 SMT solver and/or code execution with GPT4 is getting SOTA on various benchmarks
- ⇒ smarter dynamic code analysis?
- ⇒ crypto contract verification?

[ Demo: Louie <> Z3/PySym/HF? ]
Demo: Generating an OPA ABAC policy & verifying with Z3.. in 2 minutes

In 2 short prompts

- Autogenerate a reasonable OPA ABAC policy
- Autogenerate verification goals for the security policy

Converts to Python calling the Z3 theorem prover

Auto-runs, verifying & reporting whether the security policy obeys the verification goals

https://www.loom.com/spaces/All-Graphistry-1228153/folders/berkeley-194-d905f3da2df440c2b5e9e1c6824d6ab5
LLMs in security

- Escalating cyber attacks
- New era of cyber defense

- Business of security
  - Compliance
  - Market
  - CISO Priorities

- The fight beyond IT
  - Bots & misinformation
  - Emergency services
  - Intelligence community

- Controversial government operations & policy
  - Military
  - Free speech
  - Authoritarian societies

- Today + next few years
  - From technology first principles

Real-world data + tech demos
Regulatory environment is a mess

Safe result

- GDPR
  Sovereignty + deletion = 😂 + $$$

- Copyright
  Lawsuits: ChatGPT copying Sarah Silverman & reporters, while Adobe curates safe data

- Algorithmic discrimination
  Regulated technologies (ex: loan approvals) need to be free of bias

Safe process

- Explainability (XAI)
  Address risk management, consumer protection, etc. requirements

- Determinism
  Imagine using in a medical context and getting different answers!

- Chain of custody
  Similar to explainability

- Rules change by country, company & month
- People use ChatGPT even when prohibited
Topic for another day: Attacks against machine learning

Market sizes: Proxy for how leaders value mlsec

AI: Growing from $100B+ to $1T+ in 2030
Cybersecurity: $200B going to $500B in 2030

Attacks against machine learning: < 1% of cybersecurity market, and low ceiling?

Ex: Louie.AI focuses mostly on securing data, access control, & sandboxes.
LLM solves alignment, and most ML attacks outside our threat model

Challenging

Alignment: No one wants a racist clippy - users & businesses want social norms built in
CIA - Confidentiality, Integrity, Availability: Data poisoning, model inversion, …

Federated access control for federated data
Ex: When indexing Slack channels, and membership changes, ensuring RAG still right

Tool sandboxing
AutoGPT: Securing Python, DB, …

Further reading: See new guides by OWASP (light) and NIST (heavy)
CISO Priorities in practice

#1: Launch private LLMs - stop the bleeding
Internal ChatGPT UI, LLM API, & top genAI apps
First get data safe inside perimeter; AI guard rails laters

#2: Set & enforce policy
Identify policies employees will follow + top vendors

#3: Learn
Use private GPT + find responsive innovation partners enabling AI use cases (platform + solution): Fixie, Palantir, startups… including us!

#4: Low-hanging fruit
Productivity platform tools
Blackbox vendor tools

#5: Strategic initiatives
Ex: Database <> LLM enablement
Ex: Detections on traditionally hard data
Priority on indexing & using proprietary data for growing LLM use cases
LLMs in security

- Escalating cyber attacks
- New era of cyber defense
- Business of security
  - Compliance
  - Market
  - CISO Priorities
- The fight beyond IT
  - Bots & misinformation
  - Emergency services
  - Intelligence community
- Controversial government operations & policy
  - Military
  - Free speech
  - Authoritarian societies
- Today + next few years
  - From technology first principles

Real-world data + tech demos
Emergency services

Challenge: Continuous monitoring of social, news, & open media for threats & status

- Detect within 1 minute of posting:
- Hurricane damage, evacuation status, physical threat, …
- Decades of after action reports

⇒ LLM + vector DB + KG: Scalable & higher quality

Challenge: Misinformation

⇒ LLM meets graph: Label topics from misinfo networks

Challenge: Automated decision support

Fusion analysis, interactive analytics, alerting, briefs, …

⇒ LLM agents with many interfaces
Intelligence community

Intelligence gathering, analysis, & operations

- Detect specific threats: economic, democratic, physical, …
- Policy guidance: diplomatic & military
- Resource allocation clarity

Physical event → … → President’s daily brief

Raw analysis: sigint, osint, imint, …

Synthesis & higher-level reasoning

Decision science: Alerting, personalization, …

Need to rethink $100B/yr of US IC spending

OSINT Commoditization ⇒ Press, NGOs, threat intel sharing
## LLMs in security

<table>
<thead>
<tr>
<th>Category</th>
<th>Topics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Escalating cyber attacks</td>
<td></td>
</tr>
<tr>
<td>New era of cyber defense</td>
<td></td>
</tr>
<tr>
<td>Business of security</td>
<td>Compliance, Market, CISO Priorities</td>
</tr>
<tr>
<td>The fight beyond IT</td>
<td>Bots &amp; misinformation, Emergency services, Intelligence community</td>
</tr>
<tr>
<td>Controversial government operations &amp; policy</td>
<td>Military, Free speech, Authoritarian societies</td>
</tr>
<tr>
<td>Today + next few years</td>
<td>From technology first principles</td>
</tr>
</tbody>
</table>

Real-world data + tech demos
Military

Similar to intelligence community...
Focus on force, logistics, & combat
Real-time
Coordination & multi-lingual allies
Situational awareness
Training

Controversy: #autonomousweapons
Attacks against free speech societies: Election security, foreign policy, ...

Ex: President Zelensky of Ukraine faked in belly dancing video via Russian misinfo campaign

Responsibility?
Midjourney paused free tier … AFTER becoming successful

Challenges
- Vocal minority lies & artificially inflates
- Social platforms incentivized to amplify: clickbait ($$$), engagement, free speech
- Collaborating misinformation networks collaborate too: Ex - 5G, Russian propaganda, homeopathy
- GPT throwing fuel on the fire

(See also: my keynote at DefCon AI on analyzing collaborating covid medical misinfo communities)
Challenge: 1/3rd world authoritarian; controlled by tech suppressing free speech

Today: Authoritarian governments already use tech to
  ● **Fabricate** false reality
  ● **Suppress** violations
  ● **Surveil** for violators

Tools: State media, suppressed press, internet firewall, social & news censor lists

Ex: Russians believe Ukraine is not a war, but a special operation to fight Nazis; TikTok hides

Ex: Chinese deny US claims about Uyghur camps: genocide, forced labor, forced sterilization, ...

What are our responsibilities?
  ● LLMs accelerate, scale, & commoditize the capability
  ● Tech community: R&D, OSS, startups, & megacorps
  ● Government: Exports, visas, sanctions, ...
  ● Citizens, new hires, ...
LLMs in security

- Escalating cyber attacks
- New era of cyber defense
- Business of security - Compliance, Market, CISO Priorities
- The fight beyond IT - Bots & misinformation, Emergency services, Intelligence community
- Controversial government operations & policy - Military, Free speech, Authoritarian societies
- Today + next few years - From technology first principles

Real-world data + tech demos
Trajectory - Where we are today

Earlier: MS/PhD-level pipelines with low scores for HMM, RNN, … => limited use (Google, …)

2018: BERT - Basic NLP became easy via pretraining & quality => CyBERT, HuggingFace, ….

2022: GPT3 - Bigger generative models that fail typical exams

2023: ChatGPT + GPT3.5 + GPT4 - Generative models that pass typical exams + UI

Ecosystem & applications limited by GPT3/4 model limitations:

- Use cases: Any small data, slow data, and human speed (< 10 tok/sec)
- Model distillation to shrink models so it can run faster & cheaper with good quality
- Program & plan generation: Small snippets needing manual cleanup (Github CoPilot, SQL)
- Infra catchup: OSS, DB indexing, hooking up tools, … - early, see rise of vector DBs / RAG
Trajectory - Where we’re going in 2024, 2025: Coming technology waves

Current-generation tech - Application builders still catching up!

tok/s will increase - operational data
Unlocks bigger & faster data sources for tasks like alert & event monitoring

Data indexing - BYOData
KG integration, data catalog adoption, longer contexts ⇒ personal & org assistants, tools on top

Beginning of autoGPT and automation
Smarter planning & coding ⇒ autonomous hunting, ...

AutoAI
Agents learning how to use computational thinking: Python, solvers, analytics/ML/AI, DBs, KGs, ...
Thank you!

Cybersecurity

iPhone moment: Every area in security getting revisited

Attackers big on platform abuse; deep fakes coming; need stronger social/trust/identity solutions

New defender tools: Start simple - blackbox detections (ex: phishing) + co-pilots (investigation, …)

2024+ as tech improves: Autonomous (autoGPT) mode + economical line-rate monitoring

Broader opportunities & challenges

Also reshaping: Emergency management, IC community, …

Authoritarian govs: 1/3rd of population having tech replace their free speech with misinfo & surveillance

Democratic govs: Under attack from inauthentic information that gets digitally amplified

Hiring

Social & technical marketing manager,
Visualization engineer (WebGL/WebGPU)