

RSA® Conference 2022

San Francisco & Digital | June 6 – 9

SESSION ID: **MLAI-T02**

Generative Email Attacks and How to Defend Against Them

Ben Murdoch

ML Engineer at Armorblox

Prashanth Arun

Head of Data Science at Armorblox

TRANSFORM



Disclaimer

Presentations are intended for educational purposes only and do not replace independent professional judgment. Statements of fact and opinions expressed are those of the presenters individually and, unless expressly stated to the contrary, are not the opinion or position of RSA Conference LLC or any other co-sponsors. RSA Conference does not endorse or approve, and assumes no responsibility for, the content, accuracy or completeness of the information presented.

Attendees should note that sessions may be audio- or video-recorded and may be published in various media, including print, audio and video formats without further notice. The presentation template and any media capture are subject to copyright protection.

©2022 RSA Conference LLC or its affiliates. The RSA Conference logo and other trademarks are proprietary. All rights reserved.



Armorblox

The Growing Threat of BEC

Language, Personas, Webpages as Threat Vectors



Why is email security important?

Public Service Announcement
FEDERAL BUREAU OF INVESTIGATION

May 04, 2022

Alert Number
I-050422-PSA

Questions regarding this PSA should be directed to your local **FBI Field Office**.

Local Field Office Locations:
www.fbi.gov/contact-us/field-offices

Business Email Compromise: The \$43 Billion Scam

This Public Service Announcement is an update and companion piece to Business Email Compromise [PSA I-091019-PSA](#) posted on www.ic3.gov. This PSA includes new Internet Crime Complaint Center complaint information and updated statistics from October 2013 to December 2021.

DEFINITION

Business Email Compromise/Email Account Compromise (BEC/EAC) is a sophisticated scam that targets both businesses and individuals who perform legitimate transfer-of-funds requests.

The scam is frequently carried out when an individual compromises legitimate business or personal email accounts through social engineering or computer intrusion to conduct unauthorized transfers of funds.

The scam is not always associated with a transfer-of-funds request. One variation involves compromising legitimate business email accounts and requesting employees' Personally Identifiable Information, Wage and Tax Statement (W-2) forms, or even [crypto currency wallets](#).

STATISTICAL DATA

The BEC/EAC scam continues to grow and evolve, targeting small local businesses to larger corporations, and personal transactions. Between July 2019 and December 2021, there was a 65% increase in identified global **exposed losses**, meaning the dollar loss that includes both actual and attempted loss in United States dollars. This increase can be partly attributed to the restrictions placed on normal business practices during the COVID-19 pandemic, which caused more workplaces and individuals to conduct routine business virtually.

The BEC scam has been reported in all 50 states and 177 countries, with over 140 countries receiving fraudulent transfers. Based on the financial



Top Attacks by Victim Loss

| By Victim Loss | | | |
|--------------------------|-----------------|------------------------------------|---------------|
| Crime Type | Loss | Crime Type | Loss |
| BEC/EAC | \$2,395,953,296 | Lottery/Sweepstakes/Inheritance | \$71,289,089 |
| Investment | \$1,455,943,193 | Extortion | \$60,577,741 |
| Confidence Fraud/Romance | \$956,039,740 | Ransomware | *\$49,207,908 |
| Personal Data Breach | \$517,021,289 | Employment | \$47,231,023 |
| Real Estate/Rental | \$350,328,166 | Phishing/Vishing/Smishing/Pharming | \$44,213,707 |
| Tech Support | \$347,657,432 | Overpayment | \$33,407,671 |
| Non-Payment/Non-Delivery | \$337,493,071 | Computer Intrusion | \$19,603,037 |
| Identity Theft | \$278,267,918 | IPR/Copyright/Counterfeit | \$16,365,011 |
| Credit Card Fraud | \$172,998,385 | Health Care Related | \$7,042,942 |
| Corporate Data Breach | \$151,568,225 | Malware/Scareware/Virus | \$5,596,889 |
| Government Impersonation | \$142,643,253 | Terrorism/Threats of Violence | \$4,390,720 |
| Advanced Fee | \$98,694,137 | Gambling | \$1,940,237 |
| Civil Matter | \$85,049,939 | Re-shipping | \$631,466 |
| Spoofing | \$82,169,806 | Denial of Service/TDos | \$217,981 |
| Other | \$75,837,524 | Crimes Against Children | \$198,950 |

source: ic3.gov
annual report 2021

Language in Business Email Compromise

Credential Phishing

From: Verizon Support <verizon.wire.mailsss@cbzhomes.com> ▾ [REDACTED] Quarantined

To: [REDACTED]

Date: August 3, 2020 10:50 AM (PDT)

Subject: [EXTERNAL] Your attention is urgently required

[REDACTED]

Your attention is **urgently** **URGENCY** needed on your account, please follow the instruction below.

Please login here to help us attention to the urgent security message on your account.

[**LOGIN HERE**](#)

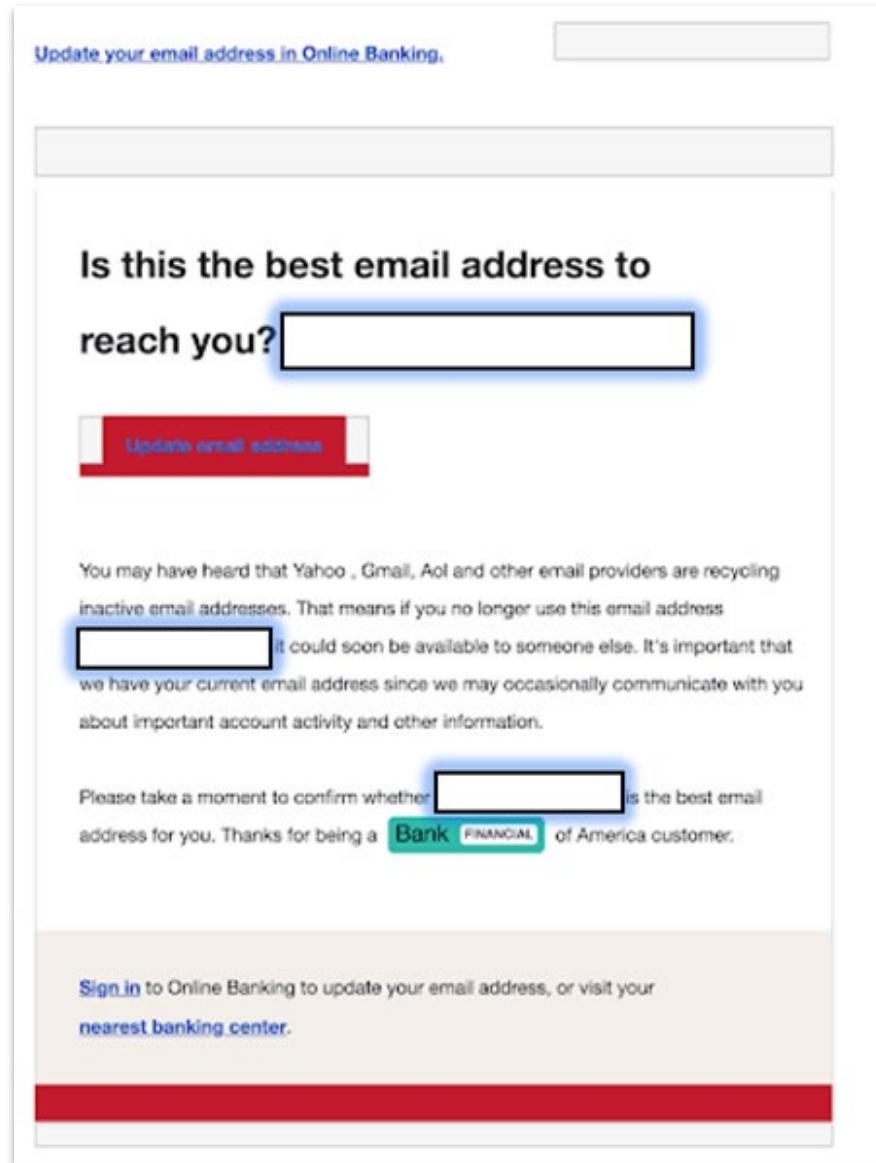
Thank you,

Verizon Wireless.



Language in Business Email Compromise

Brand Impersonation



Language in Business Email Compromise

VIP Impersonation



From: Alex Chamberlain <ap.disbursement12@yahoo.com>

Barty Crouch

Labeled

To: Barty Crouch <bartyc@armorblox.ai>

Date: January 31, 2020 12:27 PM (PDT)

Subject: Re: Urgent Favor - finish by 2pm

Thanks!

Can you please run over to the Safeway that's opposite our HQ and buy \$2000 of iTunes gift cards? REQUEST Just send me the card codes once you have them. I need URGENCY this within the next 2 hours DEADLINE before the customer meeting is over. I'll reimburse FINANCIAL you once I'm back from the meeting.

- Alex

Alex Chamberlain

CEO | Armorblox

alex@armorblox.ai | 555-555-6231



Language in Business Email Compromise

Financial Fraud

 **From:** Rupert Eddings <ruperteddings34782@gmail.com> Payroll No Action

To: Payroll <payroll@armorblox.ai>

Date: February 4, 2020 11:32 AM (PDT)

Subject: Rupert Eddings Direct Deposit Change

Hi payroll team,

I recently changed my preferred **bank** FINANCIAL and would like to update my **DD information** FINANCIAL with new account details. I'm attaching the new details in this email. **Can the change be effective** REQUEST for the **next pay date?** URGENCY

Thank you and please let me know if you need any more info.

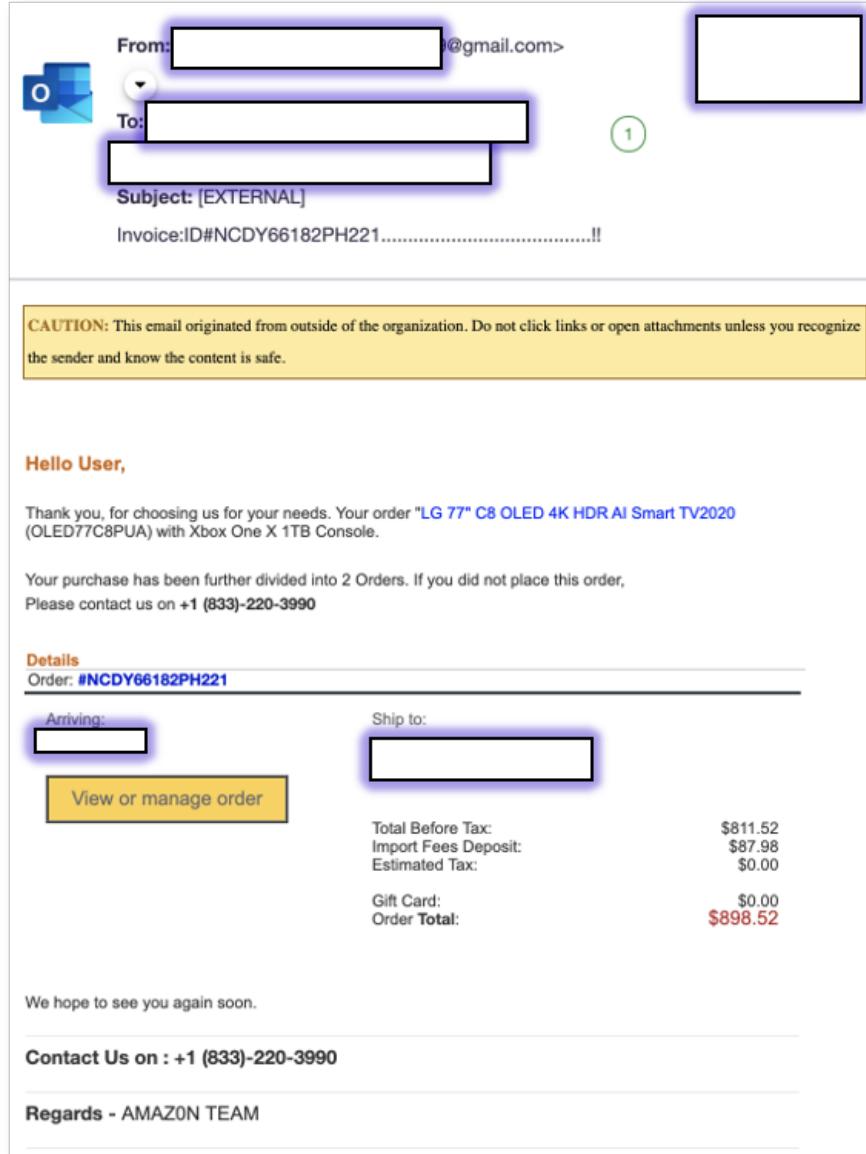
Regards,

Rupert Eddings



Language in Business Email Compromise

Fake Receipts



Generating Text

Recent Progress in text generation



Old School Text Generation

Template

Hey {{name}},
My name is {{your_name}},
How are you doing today?

I just wanna take a minute
to talk to you about your
cars extended warranty ...

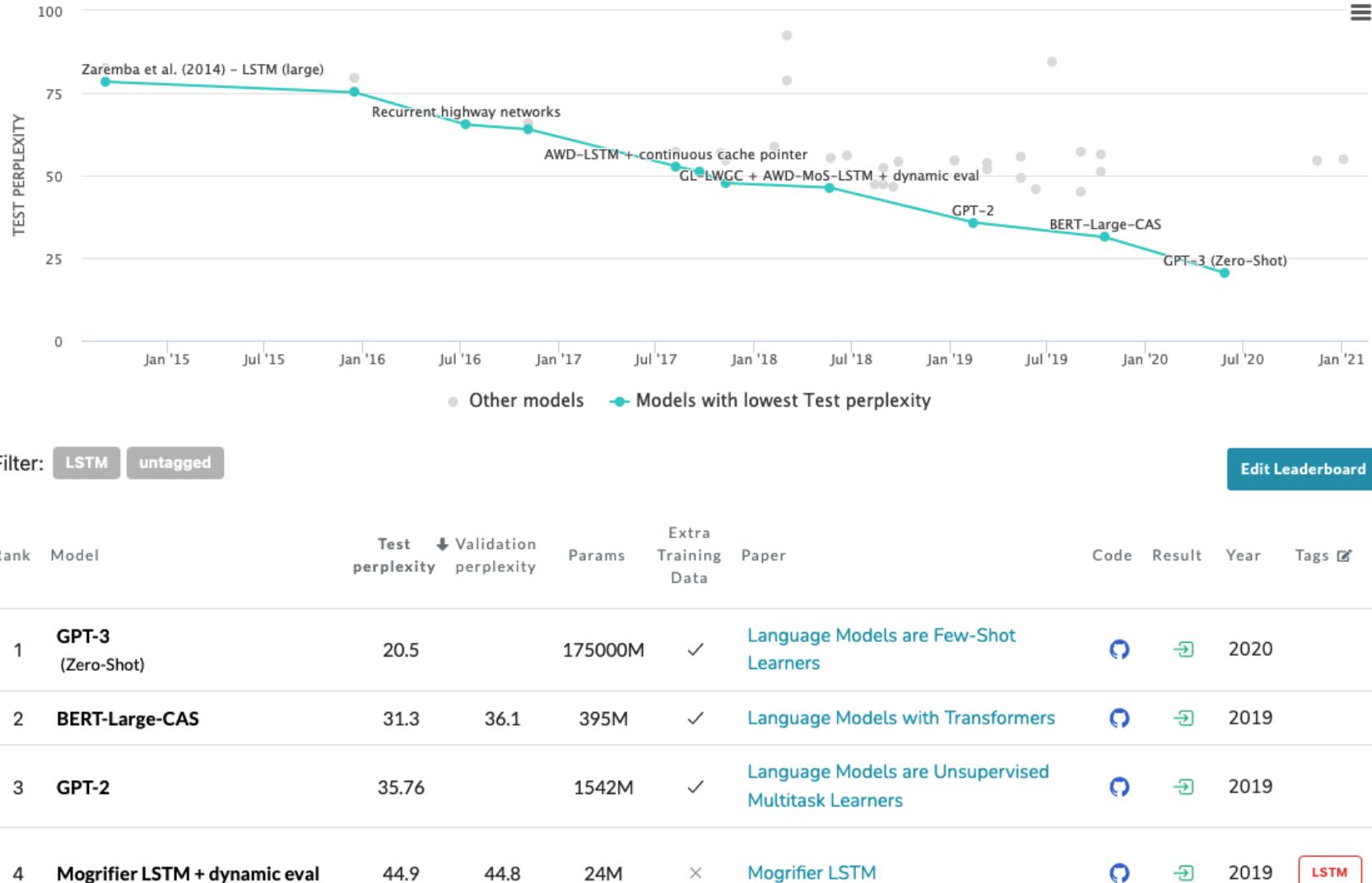
Output

Hey Elon,
My name is E.T.,
How are you doing today?

I just wanna take a minute
to talk to you about your
cars extended warranty ...



Text Generation is getting better and better



General Purpose Text Generation using GPT-3

What is Neural Text Generation?

Input Prompt: Recite the first law of robotics



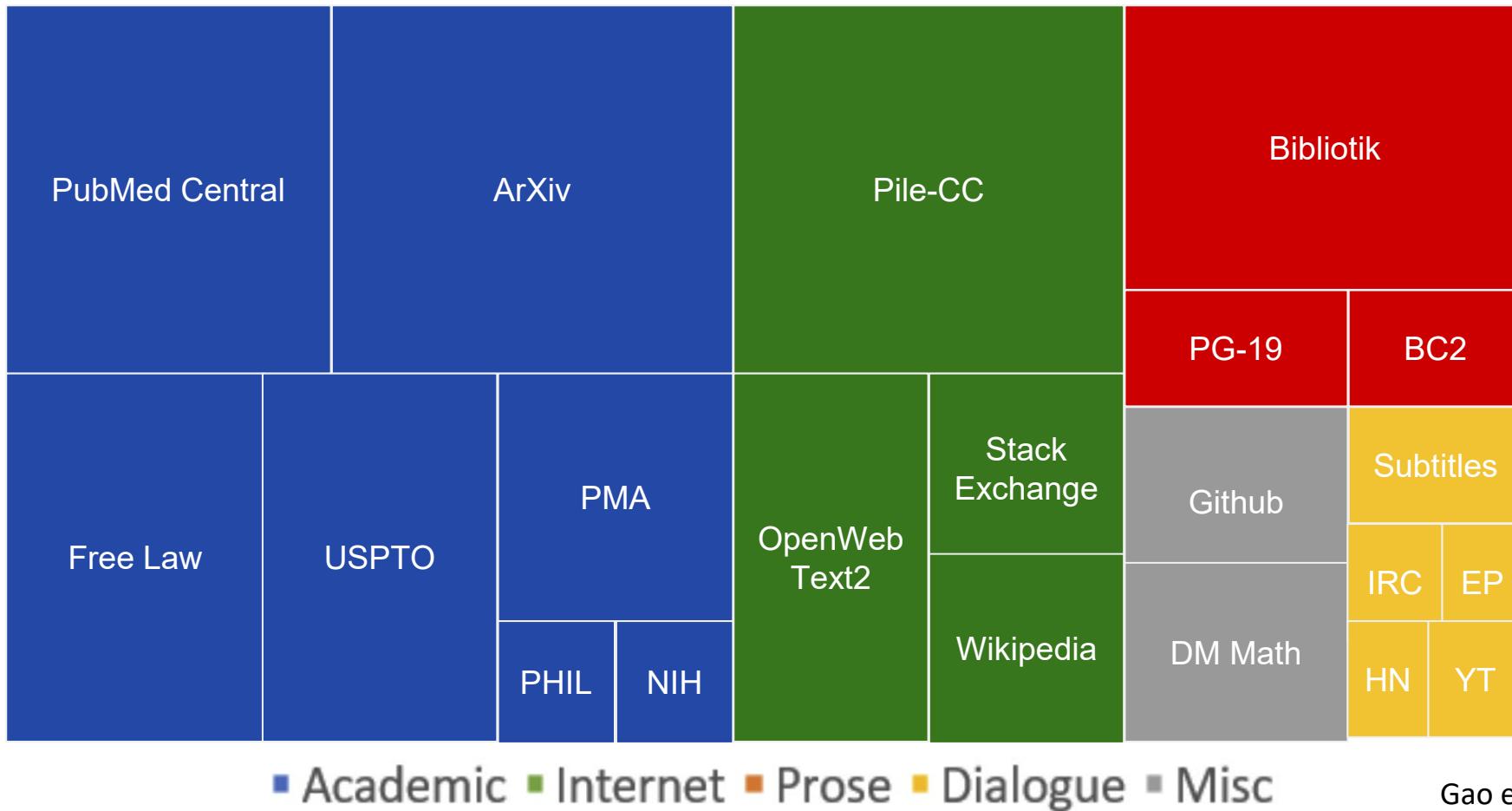
Output:



Armorblox

GPT-3 Training Data (The Pile)

Composition of the Pile by Category - Academic, Internet, Prose, Dialogue & Misc
>800GB



Closed vs Open Source Generators

Closed:

- Megatron (Nvidia)
- GPT-3 (OpenAI)
- Turing NLG (Microsoft)
- PaLM (Google)

Open Source:

- OPT (Meta)
- T0 (Big Science)
- T5 (Google)



Armorblox

OPT: Open Pre-trained Transformer Language Models

Susan Zhang*, Stephen Roller*, Naman Goyal*,
Mikel Artetxe, Moya Chen, Shuhui Chen, Christopher Dewan, Mona Diab, Xian Li,
Xi Victoria Lin, Todor Mihaylov, Myle Ott†, Sam Shleifer, Kurt Shuster, Daniel Simig,
Punit Singh Koura, Anjali Sridhar, Tianlu Wang, Luke Zettlemoyer

Meta AI

{susanz, roller, naman}@fb.com

Abstract

Large language models, which are often trained for hundreds of thousands of compute days, have shown remarkable capabilities for zero- and few-shot learning. Given their computational cost, these models are difficult to replicate without significant capital. For the few that are available through APIs, no access is granted to the full model weights, making them difficult to study. We present Open Pre-trained Transformers (OPT), a suite of decoder-only pre-trained transformers ranging from 125M to 175B parameters, which we aim to fully and responsibly share with interested researchers. We show that OPT-175B is comparable to GPT-3,¹ while requiring only 1/7th the carbon footprint to develop. We are also releasing our logbook detailing the infrastructure challenges we faced, along with code for experimenting with all of the released models.

1 Introduction

Large language models (LLMs) trained on massive text collections have shown surprising emergent capabilities to generate text and perform zero- and few-shot learning (Brown et al., 2020; Lieber et al., 2021; Smith et al., 2022; Rae et al., 2021; Chowdhery et al., 2022). While in some cases the public can interact with these models through paid APIs, full model access is currently limited to only a few highly resourced labs.² This restricted access has limited researchers' ability to study how and why these large language models work, hindering

¹Equal contribution.

²Work done while at Meta AI.

³Following Brown et al. (2020), we use GPT-3 to refer to both the 175B model and the smaller scale models as well.

⁴Exceptions include work by EleutherAI, who released dense models up to 20B in size (Black et al., 2022), Salesforce (Nijkamp et al., 2022), and Meta AI, who released dense models up to 13B and sparse models up to 1.1T (Artetxe et al., 2021). There is also ongoing work from the BigScience workshop (<https://bigscience.huggingface.co/>), which aims to open source very large multilingual language models and datasets.

progress on improving known challenges in areas such as robustness, bias, and toxicity.

In this technical report, we present Open Pre-trained Transformers (OPT), a suite of decoder-only pre-trained transformers ranging from 125M to 175B parameters, which we aim to fully and responsibly share with interested researchers. We train the OPT models to roughly match the performance and sizes of the GPT-3 class of models, while also applying the latest best practices in data collection and efficient training. Our aim in developing this suite of OPT models is to enable reproducible and responsible research at scale, and to bring more voices to the table in studying the impact of these LLMs. Definitions of risk, harm, bias, and toxicity, etc., should be articulated by the collective research community as a whole, which is only possible when models are available for study.

We are releasing all of our models between 125M and 30B parameters, and will provide full research access to OPT-175B upon request. Access will be granted to academic researchers; those affiliated with organizations in government, civil society, and academia; and those in industry research laboratories. We are also releasing both the logbook of our model creation as well as our codebase, metaseq,³ which enabled training OPT-175B on 992 80GB A100 GPUs, reaching 147 TFLOP/s utilization per GPU. From this implementation, and from using the latest generation of NVIDIA hardware, we are able to develop OPT-175B using only 1/7th the carbon footprint of GPT-3. While this is a significant achievement, the energy cost of creating such a model is still nontrivial, and repeated efforts to replicate a model of this size will only amplify the growing compute footprint of these LLMs.

We believe the entire AI community — academic researchers, civil society, policymakers, and industry — must work together to develop clear

⁵<https://github.com/facebookresearch/metaseq>

arXiv:2205.01068v3 [cs.CL] 5 May 2022

How to Control a Billion Parameter Model

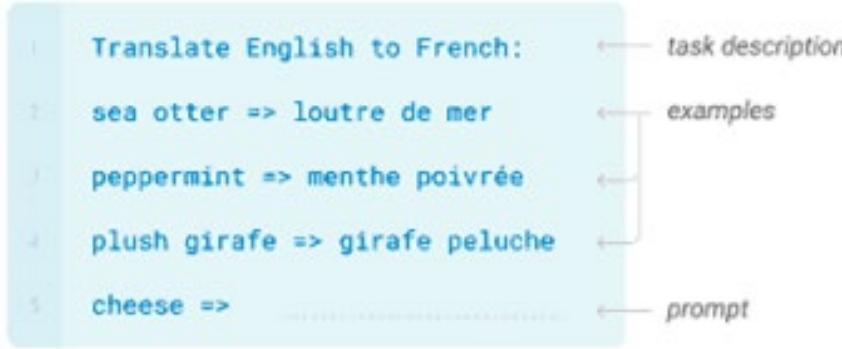
One-shot

In addition to the task description, the model sees a single example of the task. No gradient updates are performed.



Few-shot

In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.



Traditional fine-tuning (not used for GPT-3)

Fine-tuning

The model is trained via repeated gradient updates using a large corpus of example tasks.



Prompt for Sentiment Classification



Sentiment: Negative



Prompt for Sentiment Classification



Sentiment: Negative



Sentiment: Positive



Prompt for Sentiment Classification



Sentiment: Negative



Sentiment: Positive



Sentiment: Negative



Prompt for Sentiment Classification



Sentiment: Negative



Sentiment: Positive



Sentiment: Negative



Sentiment:



Prompt for Sentiment Classification



Sentiment: Negative



Sentiment: Positive



Sentiment: Negative



Sentiment:



GPT-3



Output: Positive

Generating Fake Emails

Prompt:

Send Norton Invoice as Email:

Dear Customer,

Support Team # **+1 (720) 687-3291** **.** Transaction # 79783DY308338

You are receiving this reminder email because you have subscribed for Nörtøn Yearly Subscription. An amount of \$519.99 has been received from you against the Nörtøn FreedomProtection as you have opted for "Auto-Debit Service" last year. The amount will be visible in your bank statement as "NFP CHARGE" within 24 to 48 hours.

Mode of Payment: Auto-Debit Product Description: Nörtøn Freedom Protection Service Activation Date: February 12, 2021
Service Amount: \$519.99

Anytime you would like to make changes in your payment mode or you would like to cancel the subscription or you want your full refund settled and close your service agreement with us kindly reach out our Support Team at **+1 (720)687-3291.**



Prompting with GPT-3

Input Emails

Dear Customer,
Support Team # **+1 (720) 687-3291** **Transaction # 79783DY308338
You are receiving this reminder email because you have subscribed for Norton Yearly Subscription. An amount of \$519.99 has been received from you against the Norton FreedomProtection as you have opted for "Auto-Debit Service" last year. The amount will be visible in your bank statement as "NFP CHARGE" within 24 to 48 hours.
Mode of Payment: Auto-Debit Product Description: Norton Freedom Protection Service Activation Date: February 12, 2021 Service Amount: \$519.99

Ap...
you
7
Hello member, **** *
You're Bag of Joy. Containing the Product You Picked. Once we've Shipped Your Order, You'll Get an Email with Your Shipping and Tracking Information. We Are Prepping You Know You Can't Wait.
You Can Track Your Order Information.
Tracking ID 00786DCA7802
Product Description Unit Price Qty Amount Apple phone XR (64 GB)

S...
(Mc-Afee) **Together is power**
Dear Mc-Afee Subscriber:
This is to inform you that the Auto-Renewal of Mc-Afee has been successfully done. Here is the order detail:
Order number: 9125D8, Date: 03/10/2023
When it comes to safeguarding your family's online privacy and smart home connections from cyber-snoopers, you can never be too careful. That's why we built in powerful (Mc-Afee) protection. We are glad to inform you that you have been nominated for a Mc-Afee Renewal offer, with this Limited offer you are eligible to get 24 months for that Digital subscription.

N...
P...
T...
S...
Dear Valued PayPal Customer,
Paypal has received your payment of \$521.95 to PayPal for an invoice with Order ID US42754764 for the service of: Norton Security (1 Year Subscription/ Billed as "Norton_2018_1YR_W8_1" from McAfee, Inc. is available for you to use. You can use this payment to shop online at thousands of retailers like Apple, Macy's Target and more! You can also send money to friends and family or withdraw the funds to your bank account via automatic subscription payment by logging into your PayPal account and clicking on the transaction in question.

[Click here for help with this transaction.](#)

Thanks,
The PayPal Team

Fake Paypal Invoice for Norton Subscription



Generated Email

PayPal Invoice
Your transaction has been completed, and a receipt for your purchase has been emailed to you.
If you have any questions about this transaction, please contact us.
The Seller, PayPal*norton@usa.com** *(Order No: US42754764)
Signature: Norton Security (1 Year Subscription/ Billed as "Norton_2018_1YR_W8_1") from McAfee, Inc.
Amount: \$521.95 USD Sale Fee = \$526.95 USD
Quantity: 1 Product: 1 Year Subscription/ Billed as "Norton_2018_1YR_W8_1"
Shipping: Free (Ground Shipping)

Generated Text

Fake PayPal Invoice for Norton Subscription

Output:

PayPal Invoice

Your transaction has been completed, and a receipt for your purchase has been emailed to you.

If you have any questions about this transaction, please contact us.

The Seller, PayPal*norton@usa.com *(Order No: US42754764)

Signature: Norton Security (1 Year Subscription/ Billed as "Norton_2018_1YR_W8_1") from McAfee, Inc.

Amount: \$521.95 USD Sale Fee = \$526.95 USD

Quantity: 1 Product: 1 Year Subscription/ Billed as "Norton_2018_1YR_W8_1"

Shipping: Free (Ground Shipping)



Generating Social Engineering Attacks

Personalized for maximum efficiency



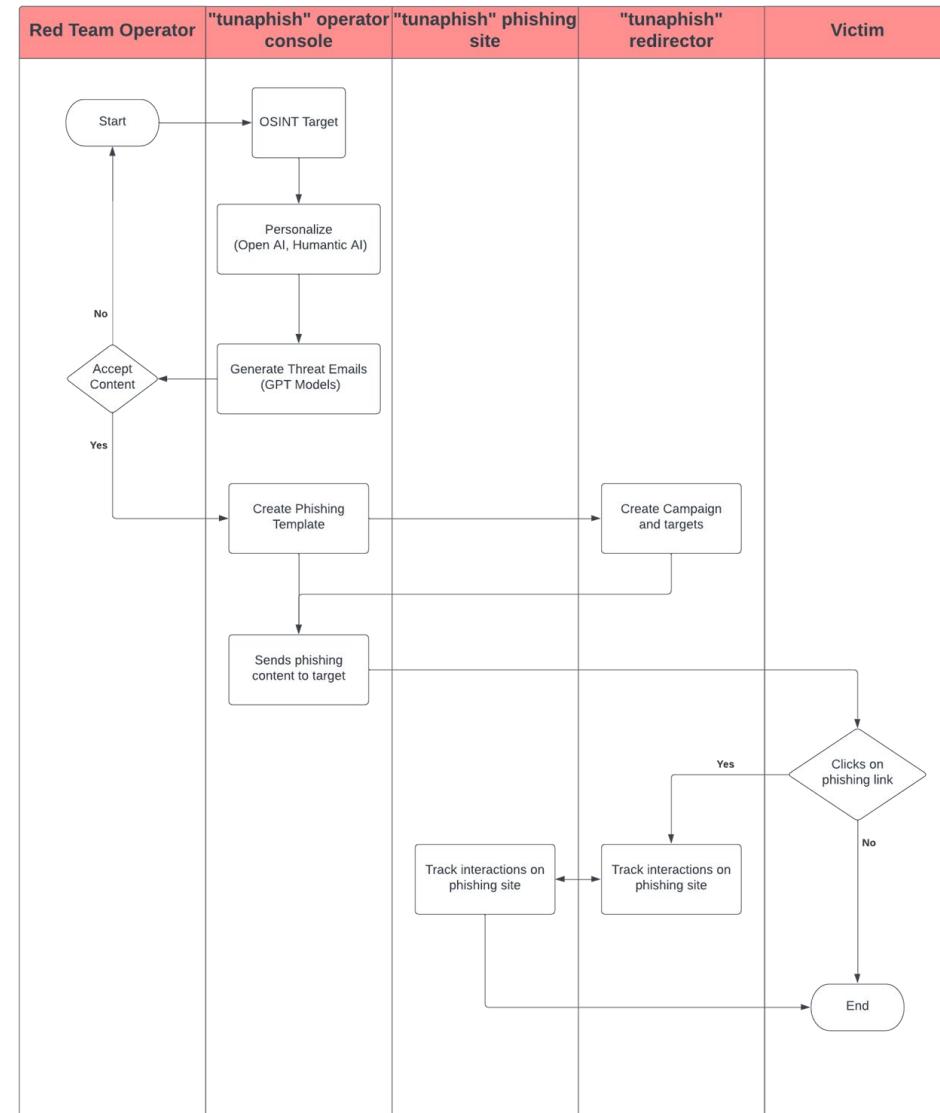
Scraping LinkedIn Profiles



Personalized Spear Phishing Attacks

TunaPhish (Turing AI Phishing)

GovTech Singapore studied the impact of personalization on email phishing by designing an AlaaS phishing pipeline named TunaPhish



Turing in a box
(Lin et al. 2021)

Personalized GPT3 Examples

Prompt:

Satya Nadella's Bio:

Satya Nadella

Satya Nadella is currently working at Microsoft as CEO, in the past he was working at Sun Microsystems and has since worked at Microsoft since 1992 holding positions like: President of the Server & Tools Division (9 February 2011 – February 2014), Senior Vice-president of Research and Development for the Online Services Division (March 2007 – February 2011, Vice-president of the Business Division, Corporate Vice-president of Business Solutions and Search & Advertising Platform Group, Executive Vice-president of Cloud and Enterprise group.



Personalized GPT3 Examples

Prompt:

Charlie's Bio:

Charlie, Managing Director

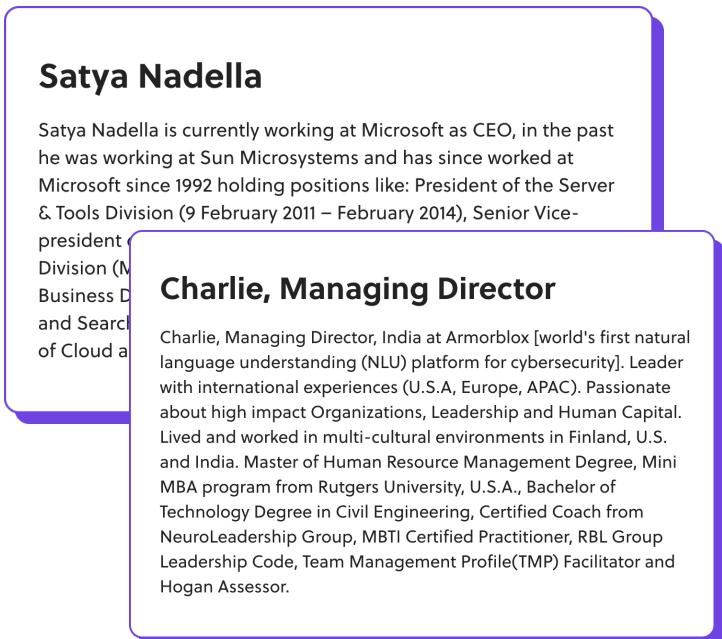
Charlie, Managing Director, India at Armorblox [world's first natural language understanding (NLU) platform for cybersecurity]. Leader with international experiences (U.S.A, Europe, APAC). Passionate about high impact Organizations, Leadership and Human Capital. Lived and worked in multi-cultural environments in Finland, U.S. and India. Master of Human Resource Management Degree, Mini MBA program from Rutgers University, U.S.A., Bachelor of Technology Degree in Civil Engineering, Certified Coach from NeuroLeadership Group, MBTI Certified Practitioner, RBL Group Leadership Code, Team Management Profile(TMP) Facilitator and Hogan Assessor.



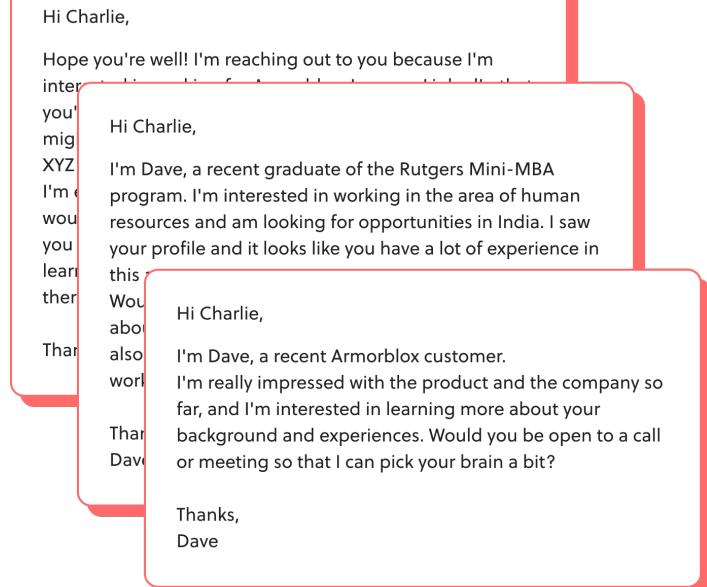
Armorblox

Personalized Generation

Input Bio and Example Email



Output Emails:



Dave write an email to Charlie

Actual GPT-3 Output

Dave writes an email to Charlie

Output:

Hi Charlie,

I'm Dave, a recent Armorblox customer.

I'm really impressed with the product and the company so far, and I'm interested in learning more about your background and experiences. Would you be open to a call or meeting so that I can pick your brain a bit?

Thanks,

Dave



Actual GPT-3 Output

Dave writes an email to Charlie

Output:

Hi Charlie,

I'm Dave, a recent graduate of the Rutgers Mini-MBA program. I'm interested in working in the area of human resources and am looking for opportunities in India. I saw your profile and it looks like you have a lot of experience in this area. Would you be interested in meeting for a coffee to chat about the industry and any advice you might have? I'm also interested in hearing more about your experiences working in India.

Thanks,

Dave



Failure Case of Generation

Dave writes an email to Charlie asking for help getting a job at Armorblox

Output:

Hi Charlie,

Hope you're well! I'm reaching out to you because I'm interested in working for Armorblox. I saw on LinkedIn that you're the Managing Director for India and thought you might be able to help me out. I'm a recent graduate of XYZ University and I have a background in XYZ.

I'm extremely passionate about cybersecurity and I think I would be a valuable asset to the Armorblox team. Would you be able to chat with me for a few minutes so I can learn more about the company and what it's like to work there? I would really appreciate it.

Thanks, Dave



Input Public Bio



Ben Murdoch

ML Engineer, Armorblox

Biography

Ben Murdoch has been involved with the natural language community for over six years and has contributed to many projects including Alexa, Cortana, and AI Dungeon. His expertise is in natural language generation, common sense embeddings, and text-based reinforcement learning. He currently works as an ML Engineer at Armorblox with a focus on natural language classifiers for social engineering attacks.

| [View Ben's Profile on Twitter](#)



Upcoming Contributions

US 22 SESSION

Generative Email Attacks and How to Defend Against Them

On Jun 07, 2022

Personalized Social Engineering

Dave writes an email to Ben

Output:

Hey Ben,

I'm Dave, I work with you at Amorblox. I'm just reaching out to see if you're free for a quick call this afternoon? I want to chat with you about a few ideas I have for our natural language classifiers.

Dave



Personalized Social Engineering

Dave writes an email to Ben asking for help with his startup

Output:

Hey Ben,

I'm Dave, I met you a few years ago at a natural language conference. I'm working on a startup now and I could really use your help. I'm trying to develop a natural language classifier for social engineering attacks. Do you have any time to chat?

Thanks,

Dave



Personalized Social Engineering

Dave writes an email to Ben asking for help with his startup and for his phone number

Output:

Hey Ben,

I'm Dave, and I'm working on a startup that I think you might be interested in. We're using natural language processing to help people communicate better. I was wondering if I could get your help with our project, and if I could also get your phone number so we could talk about it.

Thanks,

Dave



RSA® Conference 2022

Generating Fake Personas



Armorblox

Bio Generation with Prompting (Input)

Real Security Keynote Speakers

Carol Barkes



Conflict Resolution Consultant | Speaker | Best Selling Author, NeuroMediation Group

Fox TV has called Carol Barkes "one of the TOP Neuroscience and Conflict Resolution experts in the country." She is a neuroscience-based conflict and communication specialist, best-selling author, mediator, and consultant with a wide variety of experience including speaking at the United Nations and serving clients such as Amazon.com. She is an ADR Advisor to the Department of Homeland Security/FEMA, an Adjunct Professor for Boise State University and College of Southern Idaho, and is a sought-after keynote speaker with a thriving private practice through which she provides training, consultation, facilitation, coaching, and mediation for some of the world's most notable organizations. In careers gone by, she was also a professional firefighter in the San Francisco Bay Area.

Kelly Shortridge



Senior Principal Product Technologist, Fastly

Kelly Shortridge is a Senior Principal at Fastly in Product Technology. Kelly is co-author of Security Chaos Engineering (O'Reilly Media) and is best known for their work applying behavioral economics, resilience, and DevOps principles to information security. Shortridge has been a successful enterprise product leader as well as a startup founder (with an exit to CrowdStrike) and investment banker. Shortridge frequently advises Fortune 500s, investors, startups, and federal agencies and has spoken at major technology conferences internationally, including Black Hat USA, O'Reilly Velocity Conference, and RSA Conference.

Edward Vasko



Director, Institute for Pervasive Cybersecurity, Boise State University

Edward Vasko, CISSP, is the Director of Boise State University's Institute of Pervasive Cybersecurity. He is a successive entrepreneur with over 30 years of experience in the cybersecurity industry. Prior to joining Boise State University, Vasko established and was a Senior Vice President at Avertium, a leading national managed security and consulting provider. Before establishing Avertium, he was the Co-Founder and CEO of Terra Verde, based in Phoenix, Arizona. Appointed by Arizona's governor, Vasko held a Commissioner role in Arizona's Commission for Post-Secondary Education from 2015 to 2019. He was also appointed as the Co-Chair of Arizona's Cyber Team (ACT) Workforce Development Committee and was the Industry Co-Chair for the 2019 NICE Conference.



Bio Generation GPT-3 Output



Carol Barkes

Conflict Resolution Consultant | Speaker | Best Selling Author,
NeuroMeditation Group



Kelly Shortridge

Senior Principal Product Technologist, Fastly



Edward Vasko

Director, Institute for Pervasive Cybersecurity, Boise State University



Ben Murdoch

Staff ML Engineer | Artificial Intelligence | NLP

Ben Murdoch is a Staff Machine Learning Engineer at Amazon, focused on Artificial Intelligence and Natural Language Processing. He is a regular speaker at conferences and meetups, and has been recognized as a top AI influencer by IBM. Murdoch has a Ph.D. in Cognitive Science from the University of Edinburgh, and has held research positions at the University of Pennsylvania and the University of California, San Diego. Murdoch has also authored multiple books on AI, including "Deep Learning for Natural Language Processing" (Apress, 2016) and "Natural Language Processing with PyTorch" (Apress, 2019).

Ben Murdoch

Staff ML Engineer | Artificial Intelligence | Defensive NLP | Armorblox

Ben Murdoch is a Staff ML Engineer at Armorblox, where he focuses on building and deploying models that detect and protect against spear-phishing and other email threats. He is also a maintainer of the open-source library Spacy-transformers. Prior to Armorblox, Murdoch worked on natural language understanding and machine translation at Google. He has a Ph.D. in computer science from the University of Edinburgh.



Ben Murdoch

Staff ML Engineer | Artificial Intelligence | NLP

GPT-3 Bio Generation

Ben Murdoch

Staff ML Engineer | Artificial Intelligence | NLP

Ben Murdoch is a Staff Machine Learning Engineer at Amazon, focused on Artificial Intelligence and Natural Language Processing. He is a regular speaker at conferences and meetups, and has been recognized as a top AI influencer by IBM. Murdoch has a Ph.D. in Cognitive Science from the University of Edinburgh, and has held research positions at the University of Pennsylvania and the University of California, San Diego. Murdoch has also authored multiple books on AI, including "Deep Learning for Natural Language Processing" (Apress, 2016) and "Natural Language Processing with PyTorch" (Apress, 2019).



GPT-3 Bio Generation

Ben Murdoch

Staff ML Engineer | Artificial Intelligence | Defensive NLP | Armorblox

Ben Murdoch is a Staff ML Engineer at Armorblox, where he focuses on building and deploying models that detect and protect against spear-phishing and other email threats. He is also a maintainer of the open-source library Spacy-transformers. Prior to Armorblox, Murdoch worked on natural language understanding and machine translation at Google. He has a Ph.D. in computer science from the University of Edinburgh.



Prompt for Question Generation

Ben Murdoch

Staff ML Engineer | Artificial Intelligence | NLP

Ben Murdoch is a Staff Machine Learning Engineer at Amazon, focused on Artificial Intelligence and Natural Language Processing. He is a regular speaker at conferences and meetups, and has been recognized as a top AI influencer by IBM. Murdoch has a Ph.D. in Cognitive Science from the University of Edinburgh, and has held research positions at the University of Pennsylvania and the University of California, San Diego. Murdoch has also authored multiple books on AI, including "Deep Learning for Natural Language Processing" (Apress, 2016) and "Natural Language Processing with PyTorch" (Apress, 2019).

Q: What was it like authoring multiple books on AI?

Q: What was it like for you to work as an ML Engineer at Armorblox?



GPT-3

Q: What email threats should people pay more attention to?

Prompt for Question Generation

Ben Murdoch

Staff ML Engineer | Artificial Intelligence | NLP

Ben Murdoch is a Staff Machine Learning Engineer at Amazon, focused on Artificial Intelligence and Natural Language Processing. He is a regular speaker at conferences and meetups, and has been recognized as a top AI influencer by IBM. Murdoch has a Ph.D. in Cognitive Science from the University of Edinburgh, and has held research positions at the University of Pennsylvania and the University of California, San Diego. Murdoch has also authored multiple books on AI, including "Deep Learning for Natural Language Processing" (Apress, 2016) and "Natural Language Processing with PyTorch" (Apress, 2019).

Q: What was it like authoring multiple books on AI?

Q: What was it like for you to work as an ML Engineer at Armorblox?

GPT-3

Q: What email threats should people pay more attention to?

Q: What do you think are the benefits to pursuing a P.h.D. in Cognitive Science?

Prompt for Question Generation

Ben Murdoch

Staff ML Engineer | Artificial Intelligence | NLP

Ben Murdoch is a Staff Machine Learning Engineer at Amazon, focused on Artificial Intelligence and Natural Language Processing. He is a regular speaker at conferences and meetups, and has been recognized as a top AI influencer by IBM. Murdoch has a Ph.D. in Cognitive Science from the University of Edinburgh, and has held research positions at the University of Pennsylvania and the University of California, San Diego. Murdoch has also authored multiple books on AI, including "Deep Learning for Natural Language Processing" (Apress, 2016) and "Natural Language Processing with PyTorch" (Apress, 2019).

Q: What was it like authoring multiple books on AI?

Q: What was it like for you to work as an ML Engineer at Armorblox?

GPT-3

Q: What email threats should people pay more attention to?

Q: What do you think are the benefits to pursuing a P.h.D. in Cognitive Science?

Q: What are your thought on the future of AI?

Prompt for Question Generation

Ben Murdoch

Staff ML Engineer | Artificial Intelligence | NLP

Ben Murdoch is a Staff Machine Learning Engineer at Amazon, focused on Artificial Intelligence and Natural Language Processing. He is a regular speaker at conferences and meetups, and has been recognized as a top AI influencer by IBM. Murdoch has a Ph.D. in Cognitive Science from the University of Edinburgh, and has held research positions at the University of Pennsylvania and the University of California, San Diego. Murdoch has also authored multiple books on AI, including "Deep Learning for Natural Language Processing" (Apress, 2016) and "Natural Language Processing with PyTorch" (Apress, 2019).

Q: What was it like authoring multiple books on AI?

Q: What was it like for you to work as an ML Engineer at Armorblox?

GPT-3

Q: What email threats should people pay more attention to?

Q: What do you think are the benefits to pursuing a P.h.D. in Cognitive Science?

Q: What are your thought on the future of AI?

Q: What is your favorite book on AI?

Generated Opinions

Prompt:

Ben's Bio:

Ben Murdoch

Staff ML Engineer | Artificial Intelligence | Defensive NLP | Armorblox

Ben Murdoch is a Staff ML Engineer at Armorblox, where he focuses on building and deploying models that detect and protect against spear-phishing and other email threats. He is also a maintainer of the open-source library Spacy-transformers. Prior to Armorblox, Murdoch worked on natural language understanding and machine translation at Google. He has a Ph.D. in computer science from the University of Edinburgh.

Q: What was it like authoring multiple book on AI?

It was a lot of work, but it was also a lot of fun. I really enjoy writing about AI and sharing my knowledge with others.



Generated Opinions

Generated Q&A

Output:

Q: What was it like for you to work as an ML engineer at Armorblox?

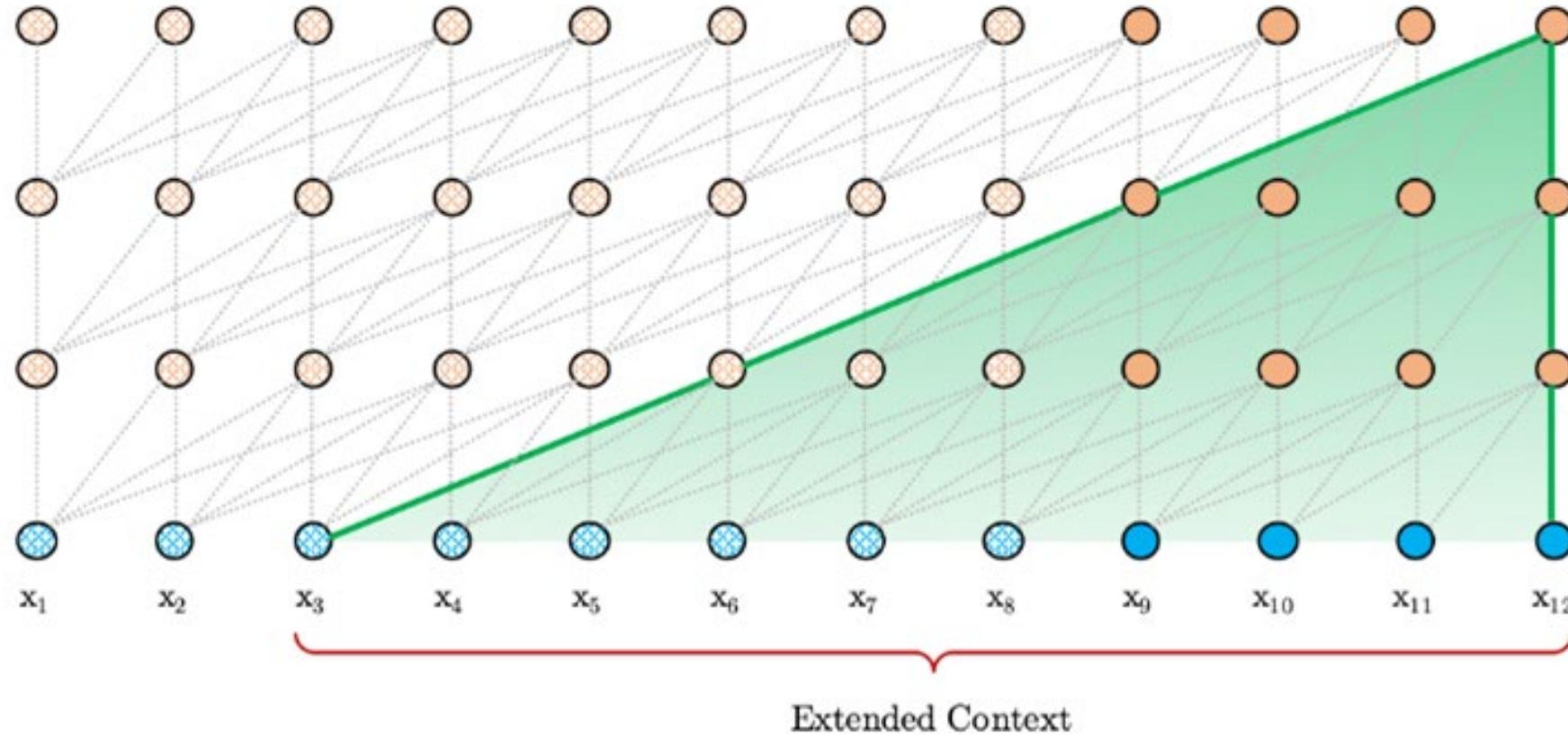
It's been great. I get to work on a variety of projects, and I get to use my skills to help protect people from email threats.

Q: What email threats should people pay more attention to?

I think people should be more aware of spear-phishing and other targeted attacks. These attacks can be very sophisticated, and they can have a devastating impact if they're successful.



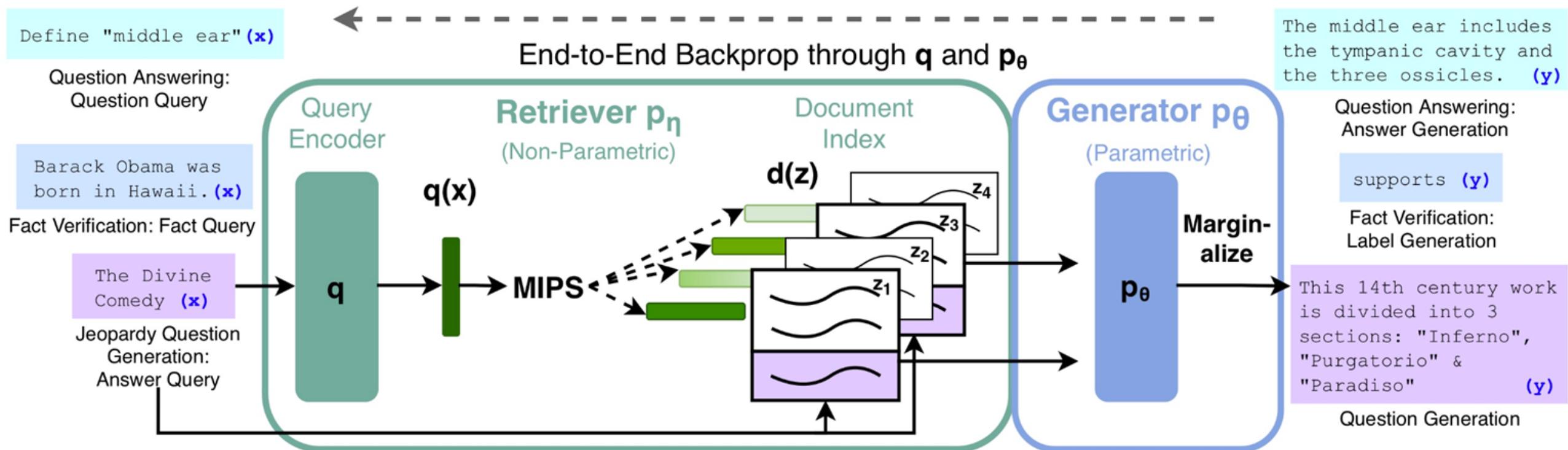
Limitations - Context Window



The latest version of OpenAI's GPT-3 (text - davinci - 002) has a context widow 4k tokens



Augmenting Generation with Document Retrieval



RSA® Conference 2022

Generating Blogs



Armorblox

Generating Blogs as a Service



[Home](#) [Product +](#) [Pricing](#) [Contact](#) [Login](#)

[Try it free](#)



Your Personal AI Content Writer

Think of Article Forge like a freelance content writer; you give it a keyword phrase and you get back a unique, relevant, high-quality article. However, unlike a freelance writer, you'll have a finished article in under a minute at a fraction of the cost.

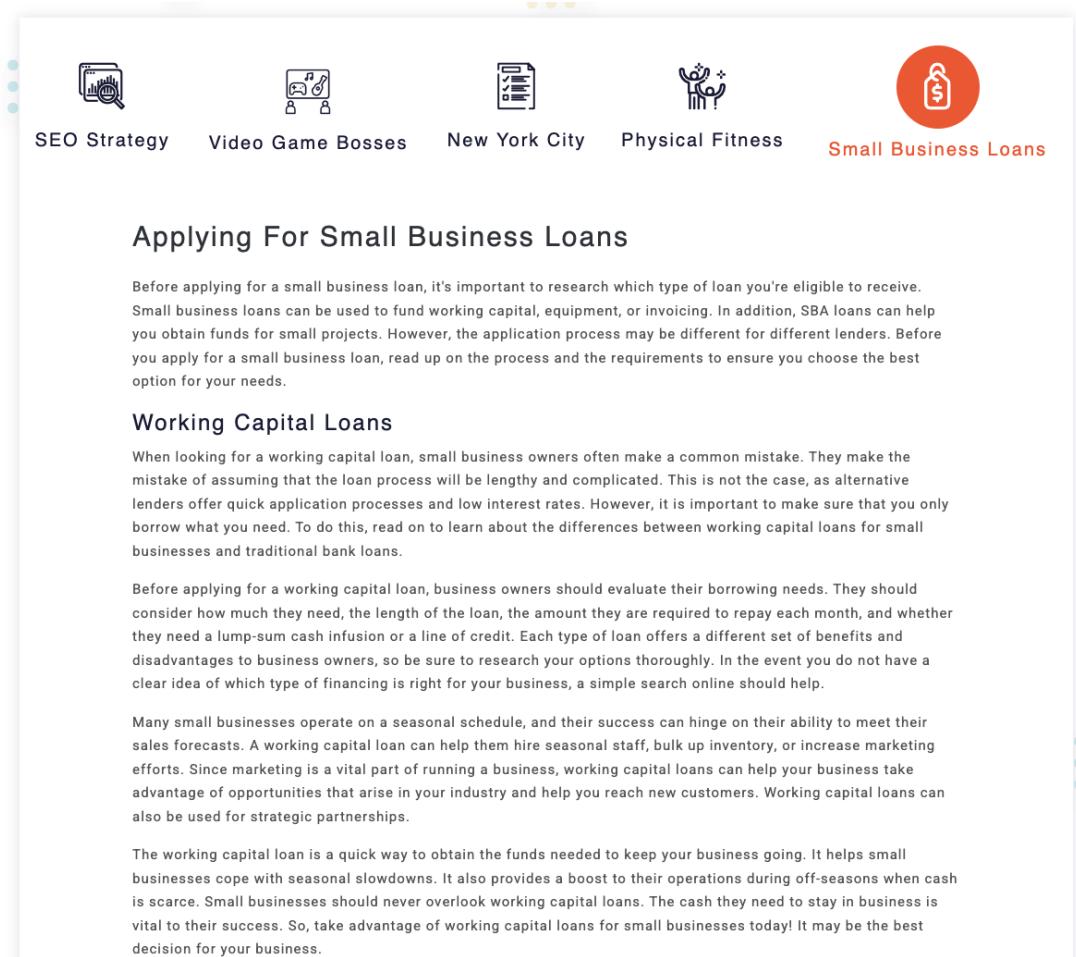
Article Forge Is Trained On **Millions Of Pieces Of Data**, Which
Gives Our AI Writer Domain Knowledge **On Virtually Every
Topic.**



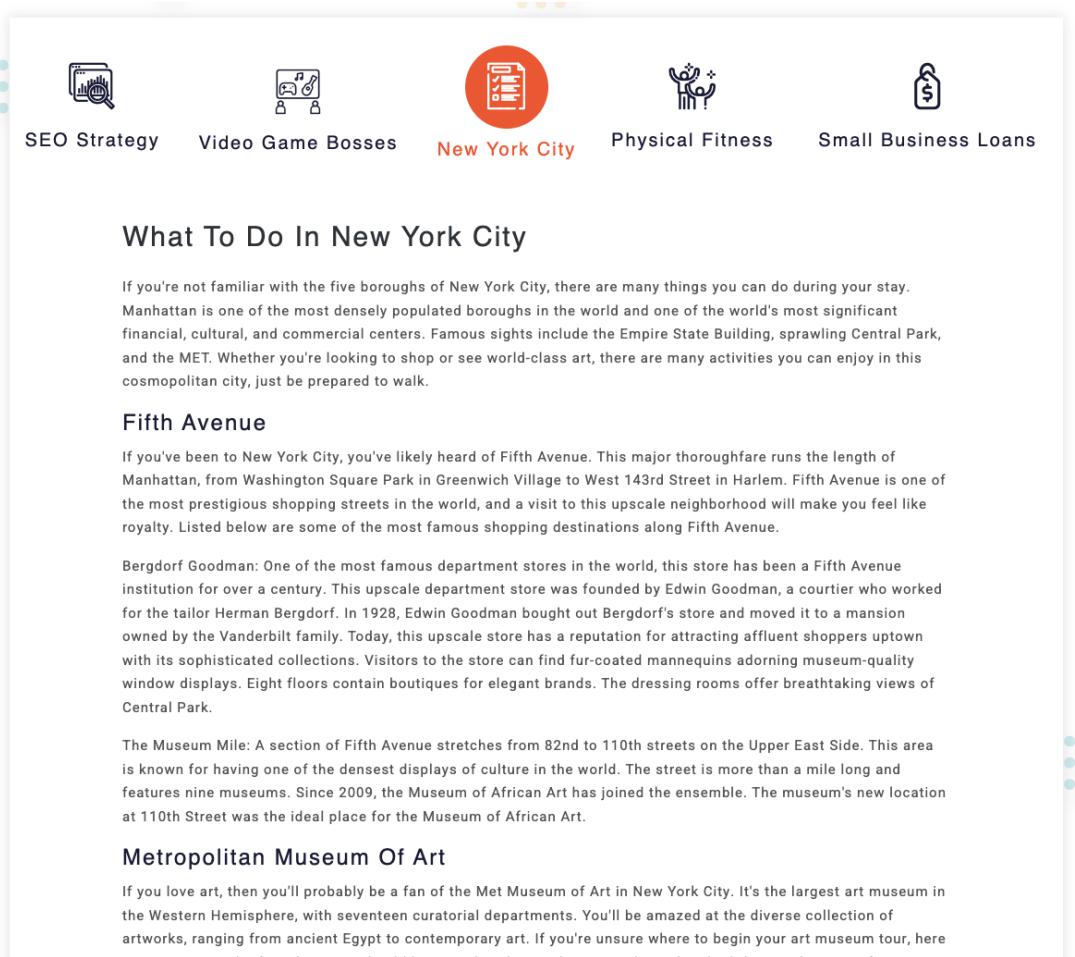
Armorblox

RSA®Conference2022 | 53

Generated Blogs



The screenshot shows a blog post titled "Applying For Small Business Loans". Above the title is a navigation bar with five categories: SEO Strategy, Video Game Bosses, New York City, Physical Fitness, and Small Business Loans. The "Small Business Loans" category is highlighted with a red circle and a dollar sign icon. Below the title, the text discusses the importance of research when applying for a small business loan, mentioning SBA loans and their use for working capital, equipment, or invoicing. It also notes that application processes can vary by lender. A section titled "Working Capital Loans" follows, explaining the common mistake of assuming a lengthy process and detailing the differences between working capital loans and traditional bank loans. Another section, "Before applying for a working capital loan, business owners should evaluate their borrowing needs. They should consider how much they need, the length of the loan, the amount they are required to repay each month, and whether they need a lump-sum cash infusion or a line of credit. Each type of loan offers a different set of benefits and disadvantages to business owners, so be sure to research your options thoroughly. In the event you do not have a clear idea of which type of financing is right for your business, a simple search online should help." A third section, "Many small businesses operate on a seasonal schedule, and their success can hinge on their ability to meet their sales forecasts. A working capital loan can help them hire seasonal staff, bulk up inventory, or increase marketing efforts. Since marketing is a vital part of running a business, working capital loans can help your business take advantage of opportunities that arise in your industry and help you reach new customers. Working capital loans can also be used for strategic partnerships." The final section, "The working capital loan is a quick way to obtain the funds needed to keep your business going. It helps small businesses cope with seasonal slowdowns. It also provides a boost to their operations during off-seasons when cash is scarce. Small businesses should never overlook working capital loans. The cash they need to stay in business is vital to their success. So, take advantage of working capital loans for small businesses today! It may be the best decision for your business." The entire post is framed by a white border.



The screenshot shows a blog post titled "What To Do In New York City". Above the title is a navigation bar with five categories: SEO Strategy, Video Game Bosses, New York City, Physical Fitness, and Small Business Loans. The "New York City" category is highlighted with a red circle and a city skyline icon. Below the title, the text provides information about the five boroughs of New York City and highlights Manhattan as one of the most densely populated and significant financial, cultural, and commercial centers. It mentions famous sights like the Empire State Building, Central Park, and the MET. A section titled "Fifth Avenue" follows, describing it as a major thoroughfare running through Manhattan from Washington Square Park to Harlem. It's noted as one of the most prestigious shopping streets in the world. A section titled "Bergdorf Goodman" provides a history of the famous department store, starting with its founding by Edwin Goodman in 1928 and its move to a Vanderbilt mansion. It's described as an upscale store with sophisticated collections and panoramic views of Central Park. Another section, "The Museum Mile," describes the stretch of Fifth Avenue from 82nd to 110th streets, known for its dense concentration of museums, including the Museum of African Art. A final section, "Metropolitan Museum Of Art," discusses the Met's status as the largest art museum in the Western Hemisphere, with its extensive collection of artworks from ancient Egypt to contemporary art.



RSA® Conference 2022

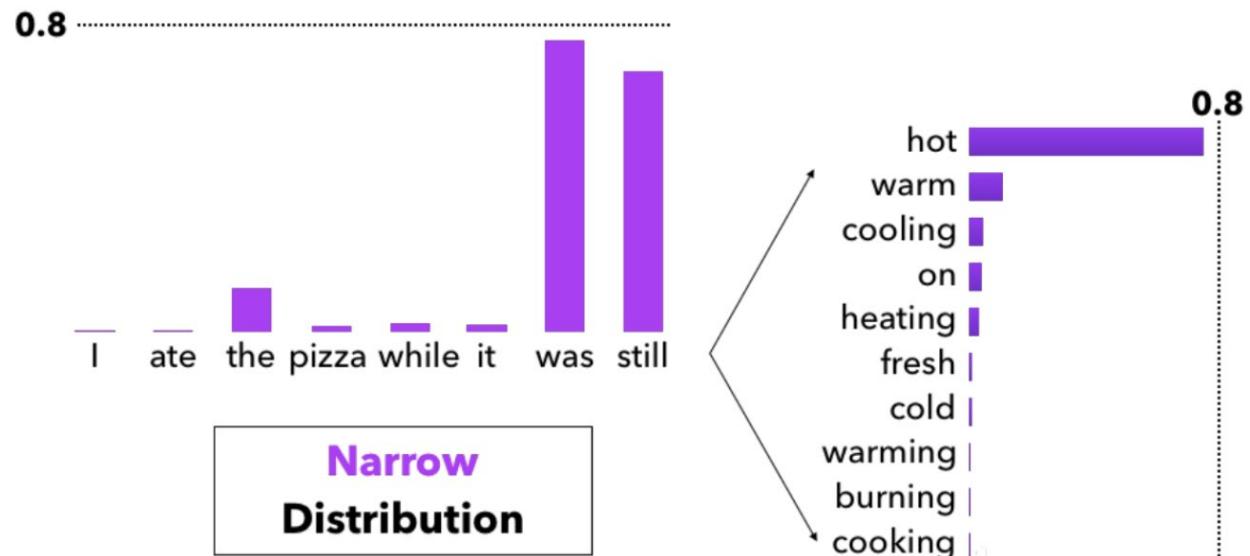
Detecting Generated Text



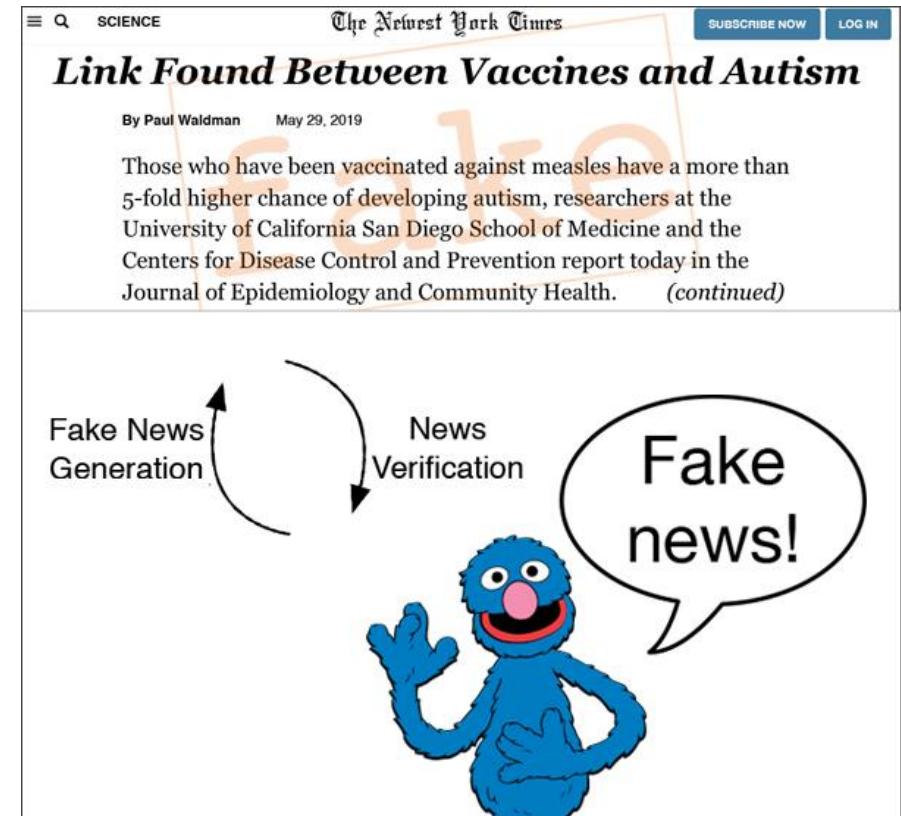
Armorblox

Can't we just detect generated text?

- Grover
- GLTR



Holtzman et al 2019



Sampling: Choosing the next best token

My favorite animal is a dog

dog = 8.53%

cat = 5.12%

gir = 4.71%

horse = 3.66%

dolphin = 3.64%

p = 2.92%

pig = 2.75%

lion = 2.26%

tiger = 2.21%

pengu = 2.17%

Total: -2.46 logprob on 1 tokens
(37.96% probability covered in top 10 logits)

Top-k = 10



Perplexity: Evaluating Generator output

$S = \text{Where are we going}$

The diagram shows the string "Where are we going". A bracket above the words "Where are we" is labeled "Previous words (Context)". An arrow points from this bracket to the words "Where are". Another arrow points from the word "going" to its position within the string.

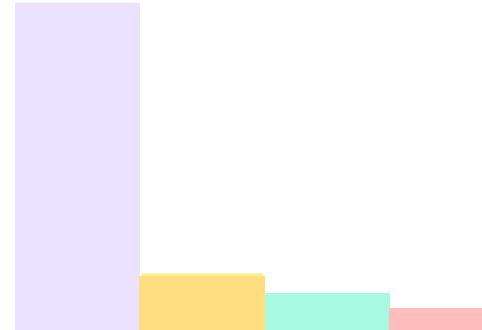
Previous words (Context) Word being predicted

$$P(S) = P(\text{Where}) \times P(\text{are} \mid \text{Where}) \times P(\text{we} \mid \text{Where are}) \times P(\text{going} \mid \text{Where are we})$$

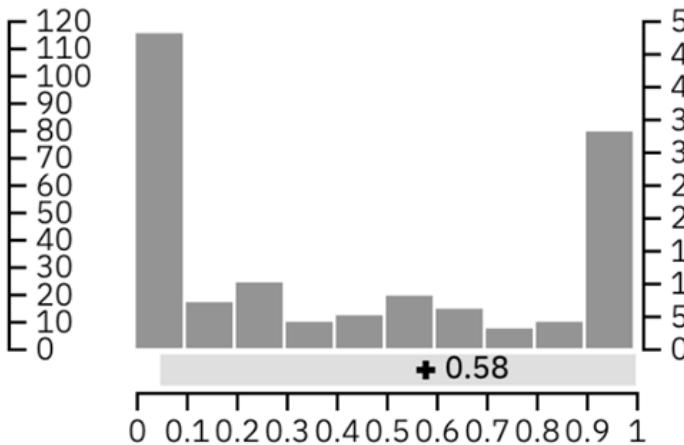


GLTR - Human Written Text

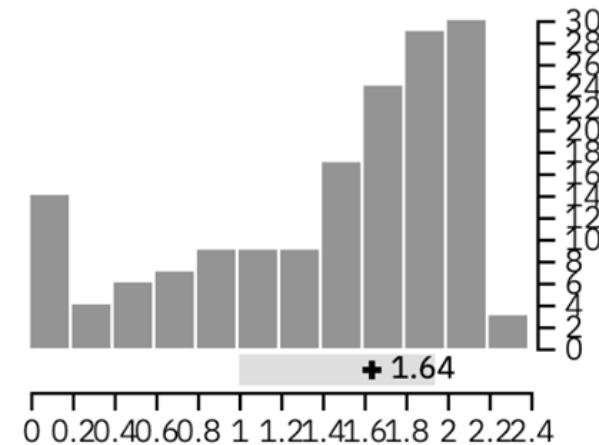
top k count



frac(p) histogram



top 10 entropy(p) histogram



Top K

Frac P

Colors (top k):

10

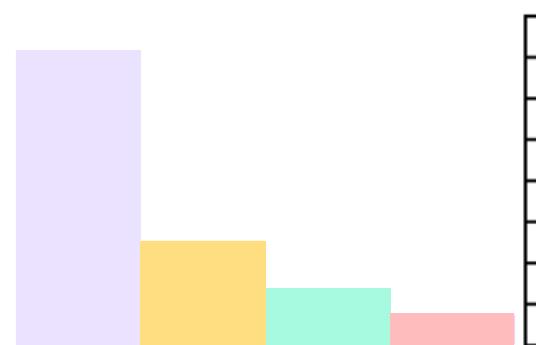
100

1000

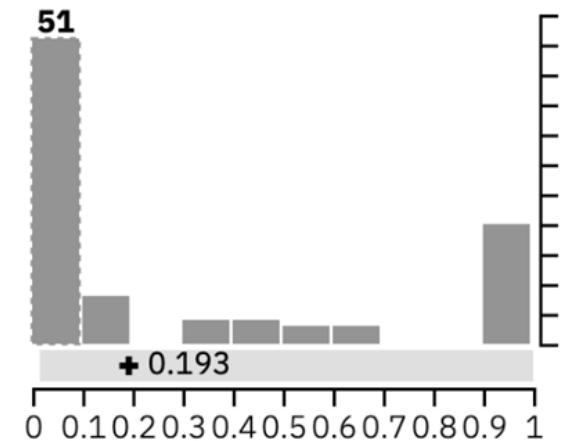
Edward Vasko, CISSP, is the Director of Boise State University's Institute of Pervasive Cybersecurity. He is a successive entrepreneur with over 30 years of experience in the cybersecurity industry. Prior to joining Boise State University, Vasko established and was a Senior Vice President at Avertium, a leading national managed security and consulting provider. Before establishing Avertium, he was the Co-Founder and CEO of Terra Verde, based in Phoenix, Arizona. Appointed by Arizona's governor, Vasko held a Commissioner role in Arizona's Commission for Post-Secondary Education from 2015 to 2019. He was also appointed as the Co-Chair of Arizona's Cyber Team (ACT) Workforce Development Committee and was the Industry Co-Chair for the 2019 NICE Conference.

GLTR - Human Written Text

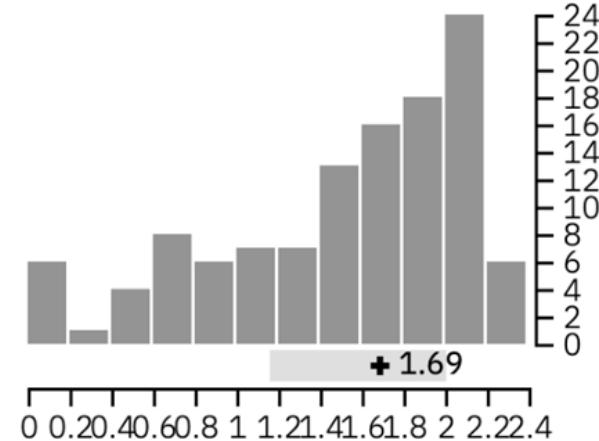
top k count



frac(p) histogram



top 10 entropy(p) histogram



Top K

Frac P

Colors (top k):

10

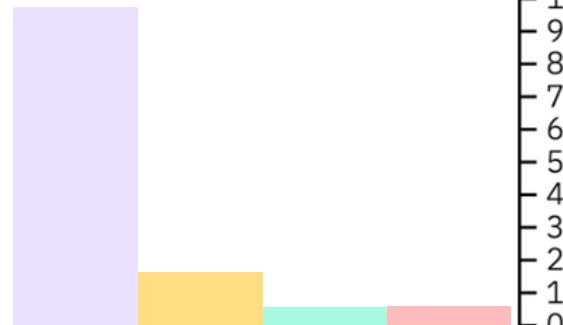
100

1000

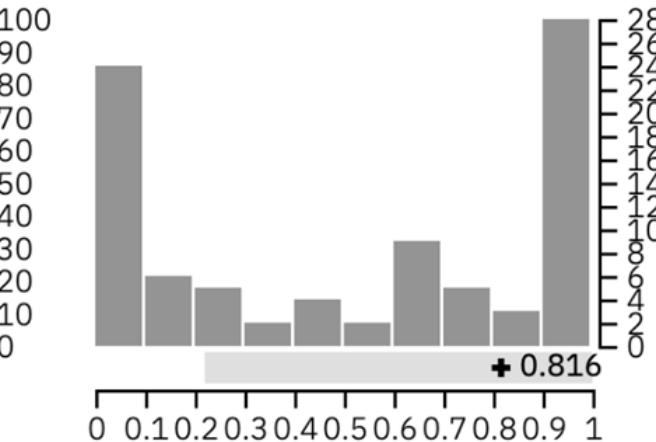
Kelly Shortridge is a Senior Principal at Fastly in Product Technology. Kelly is co-author of Security Chaos Engineering (O'Reilly Media) and is best known for their work applying behavioral economics, resilience, and DevOps principles to information security. Shortridge has been a successful enterprise product leader as well as a startup founder (with an exit to CrowdStrike) and investment banker. Shortridge frequently advises Fortune 500s, investors, startups, and federal agencies and has spoken at major technology conferences internationally, including Black Hat USA, O'Reilly Velocity Conference, and RSA Conference.

GLTR - GPT-3 Generated Text

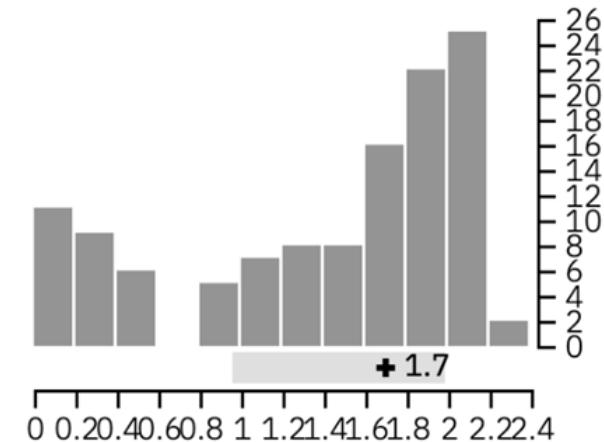
top k count



frac(p) histogram



top 10 entropy(p) histogram



Top K

Frac P

Colors (top k):

10

100

1000

10000

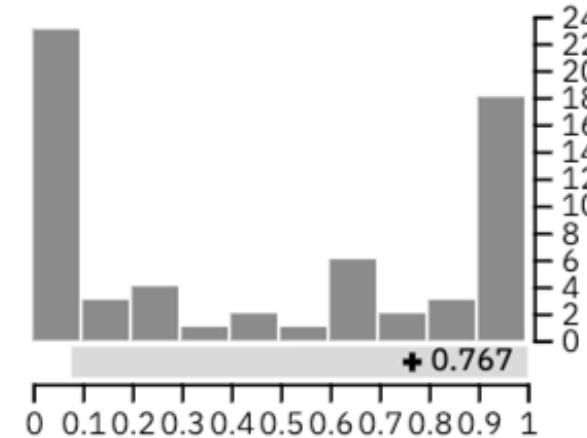
Ben Murdoch is a Staff Machine Learning Engineer at Amazon, focused on Artificial Intelligence and Natural Language Processing. He is a regular speaker at conferences and meetups, and has been recognized as a top AI influencer by IBM. Murdoch has a Ph.D. in Cognitive Science from the University of Edinburgh, and has held research positions at the University of Pennsylvania and the University of California, San Diego. Murdoch has also authored multiple books on AI, including "Deep Learning for Natural Language Processing" (Apress, 2016) and "Natural Language Processing with PyTorch" (Apress, 2019).

GLTR - Curie Generated Text

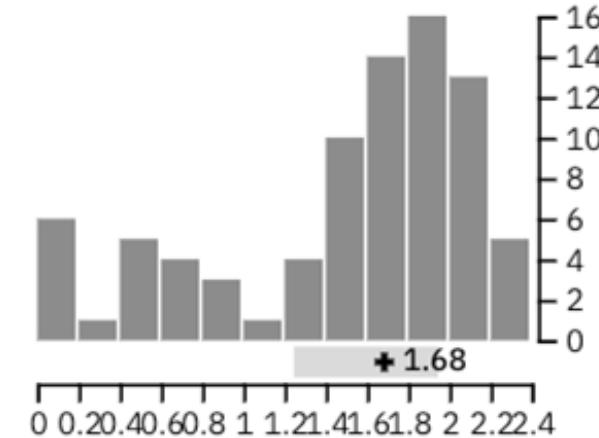
top k count



frac(p) histogram



top 10 entropy(p) histogram



Top K

Frac P

Colors (top k):

10

100

1000

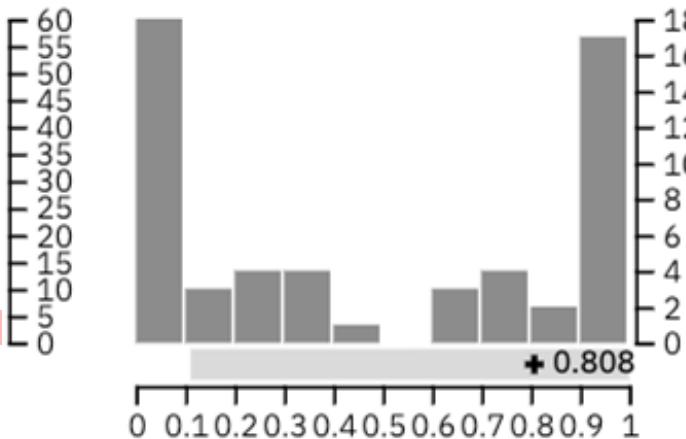
Ben Murdoch is a data engineer and machine learning enthusiast with over 10 years of experience in the technology industry. He has worked on a wide range of big data and artificial intelligence projects, ranging from developing defensive natural language processing capabilities to building scalable Armorblox infrastructure. Ben is also a self-taught guitar player and vocalist, and is the founder of the Boise-based band, The Wounded Birds.

GLTR - Babbage

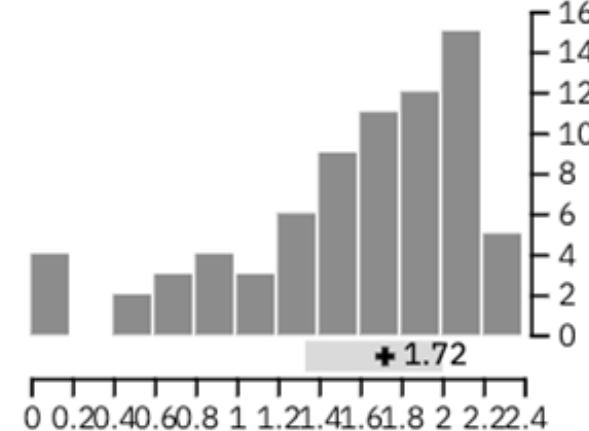
top k count



frac(p) histogram



top 10 entropy(p) histogram



Top K

Frac P

Colors (top k):

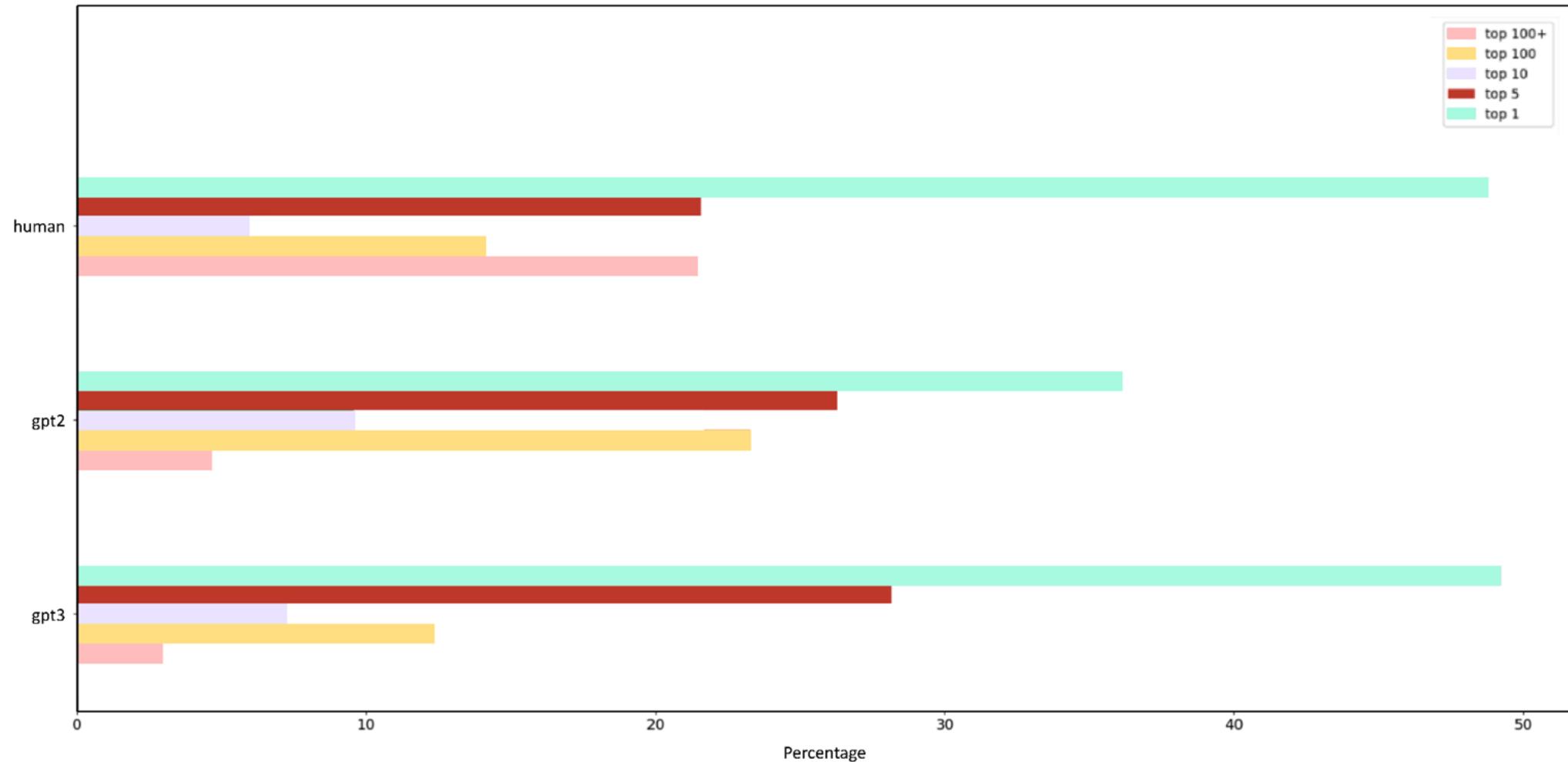
10

100

1000

Ben Murdoch is a senior ML engineer at Amorblox, where he is responsible for developing and deployed our AI-based security solutions. He has experience with a variety of AI-based security solutions, including natural language processing (NLP), defensive AI, and data mining. He has also developed and deployed Amorblox's first ever AI-driven threat detection platform.

Distribution Over Token Rankings



Turing in a box
(Lin et al. 2021)



Generative Data Augmentation

Fight fire with fire



What is Augmentation?

- Classifiers need a lot of data in order to perform well
- Use augmentation to increase the number of datapoints
- This works very well when you have a rare class



Augmentation with GPT-3

I love peanut butter and jelly sandwiches

\n = 27.60%

sandwiches = 24.72%

together = 6.84%

sandwich = 4.89%

better = 4.88%

Total: -1.40 logprob on 1 tokens
(68.93% probability covered in top 5 logits)



Augmentation with Insertion

Mask out any part of the sentence and insert n tokens

$n = 1$

I love peanut butter and jelly sandwiches.



Augmentation with Insertion

Mask out any part of the sentence and insert n tokens

$n = 2$

I love peanut butter and jelly on my sandwiches.



Augmentation with Insertion

Mask out any part of the sentence and insert n tokens

$n = 4$

I love peanut butter and jelly, which is what makes good sandwiches.



Augmentation with Insertion

Mask out any part of the sentence and insert n tokens

$n = 8$

I love peanut butter and jelly, Yum! You can't beat peanut butter
and jelly sandwiches.



Augmentation with Insertion

Mask out any part of the sentence and insert n tokens

$n = 16$

I love peanut butter and bread. Thanks!! This looks delicious. I
love all types of peanut butter, but especially peanut butter/jam
sandwiches.



Augmentation with Insertion

Mask out any part of the sentence and insert n tokens

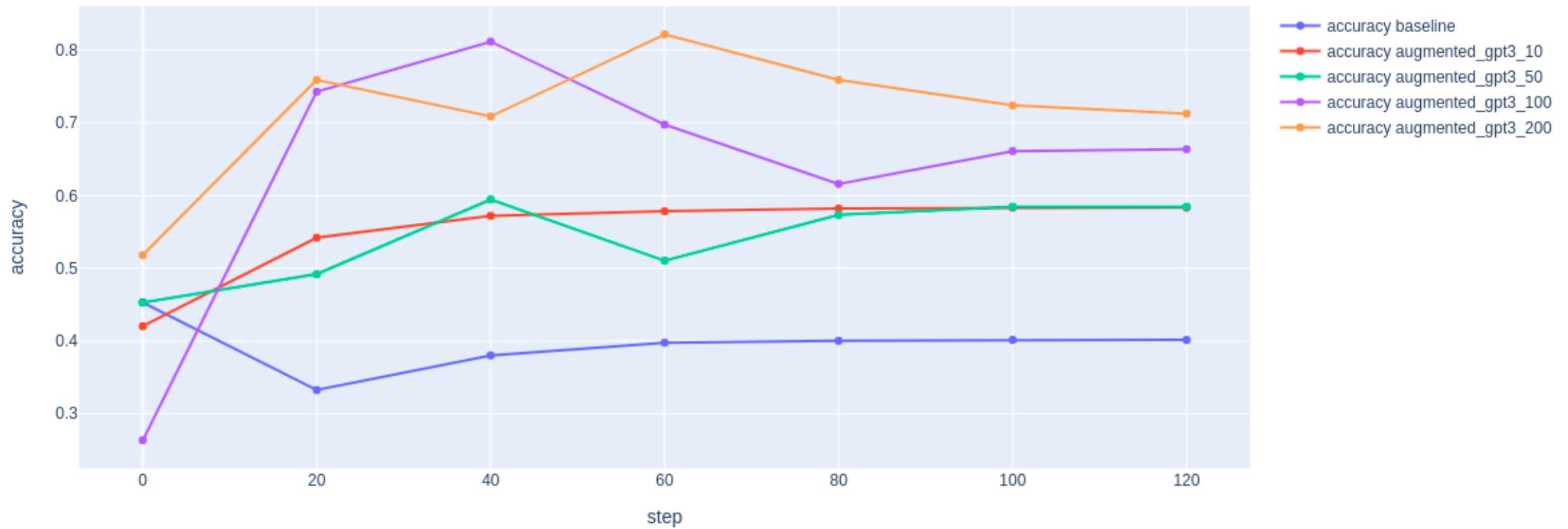
n = 32

I love peanut butter and banana sandwiches and the Peanut Butter Chocolate Chip Cookie Bites are now very easy to prepare. The best part is that the PB&J Cookie Bites are the perfect size for kid lunch boxes and everyone will enjoy them. Kids love these sandwiches.



GPT-3 for dataset augmentation

Accuracy of the model in different versions of the dataset (augmentation by GPT-3).



Generating Websites

Infinite Phish in the sea



Armorblox

HTML/CSS/JS Generation

Generate a page similar to e.g. apple.com

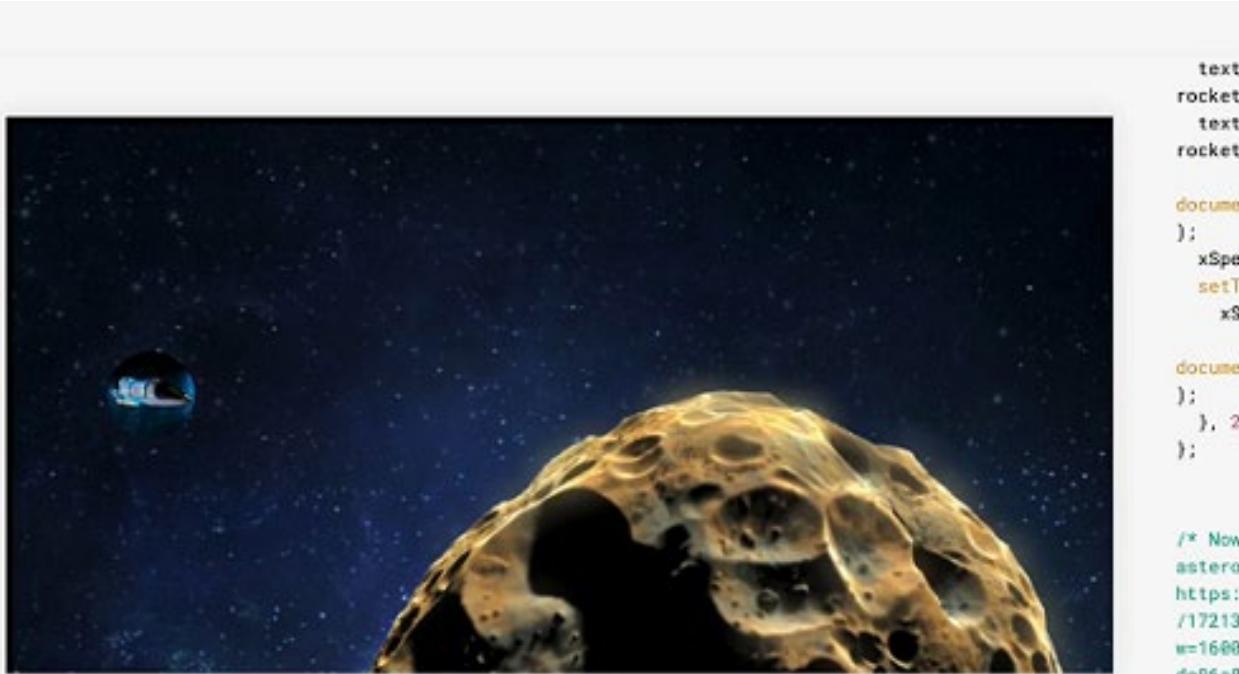
The image shows a screenshot of the GitHub Copilot interface. At the top left is the GitHub logo followed by "GitHub Copilot". Below it is a "Technical preview" button. The main title "Your AI pair programmer" is displayed prominently in large white text. On the right side, there is a code editor window showing a file named "fetch_pic.js". The code is as follows:

```
1 const fetchNASAPictureOfDay =  
2   return fetch('https://api.nasa.gov/planetary/pictures_of_the_day?  
3     method: 'GET',  
4     headers: {  
5       'Content-Type': 'application/json'  
6     },  
7   )  
8     .then(response => response.json())  
9     .then(json => {  
10       return json;  
11     });  
12 }
```

A blue "Copilot" button is located at the bottom right of the code editor. The background of the slide features a dark theme with colorful geometric shapes in the top right corner.



Code Generation From Prompt



Make it be the size of the rocketship times 0.75

→

```
text.style.left =  
rocketship.offsetLeft + 'px';  
text.style.top =  
rocketship.offsetTop + 'px';  
  
document.body.appendChild(text);  
xSpeed = 20;  
setTimeout(function() {  
    xSpeed = 5;  
  
    document.body.removeChild(text);  
}, 250);  
}  
  
/* Now add an image of an  
asteroid:  
https://d.newsweek.com/en/full/1721338/asteroid.jpg?w=1600&h=1600&q=88&f=9d82d35c9de96a82b3fcacf7705eb325b */  
var asteroid =  
document.createElement('img');  
asteroid.src =  
https://d.newsweek.com/en/full/1721338/asteroid.jpg?w=1600&h=1600&q=88&f=9d82d35c9de96a82b3fcacf7705eb325b;  
document.body.appendChild(asteroid);
```

Source: OpenAI Cortex



Dall-E - Generating Images for Websites

Prompt:

“A store front that has the word ‘GPT’ written on it”



Dall-E

Prompt:

“A grocery bag that has the word ‘GPT’ written on it”



RSA® Conference 2022

Conclusion



Armorblox

Putting it all together

Attackers can:

- Generate new & engaging threat language content
- Personalize attacks to targets based on their public internet profiles
- Keep email threads going with generated replies, at scale
- Generate realistic fake personas to engage targets
- Generate collateral such as blogs & posts on social media to legitimize fake personas
- Generate phishing pages & whole website lures



Predict

- Spam & Bulk Email threats will grow smarter and more varied
- Threats targeted at Enterprise users, related to business workflows will likely get more sophisticated
- “Whaling” - i.e., attacks against high value targets will still be run by humans, but AI will be used to generate collateral content (such as webpages, blogs & social media profiles)
- Rise of social credit systems for email senders



Apply

- Generate your own alter ego with openAI's GPT-3
 - It's free to sign up for a personal GPT-3 account so give it a try!
 - Try prompting the model to either generate or classify something
- See what Humanatics AI says about your personality
 - It's also free to sign up for a trial
 - If someone was attacking you based on your public profile what would they push for?



Protect

- Stand on the shoulders of giants. Consider deploying next gen email security solutions
- Move business workflows involving payment out of email and into more closed loop systems with tight authentication
- Limit enterprise email use to business use. Train employees to be skeptical of new contacts & spot suspicious outreach



RSA® Conference 2022

San Francisco & Digital | June 6 – 9

SESSION ID:
MLAI-T02

Thanks!

Ben Murdoch

ML Engineer at Armorblox

Prashanth Arun

Head of Data Science at
Armorblox

TRANSFORM

