

# The Future of Software Engineering in the light of LLMs

Erik Johannes Husom 24th May 2023 SCT – Department gathering

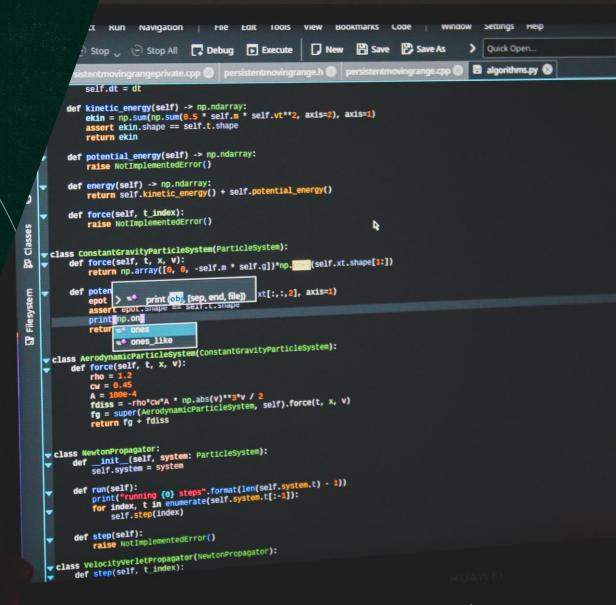


Photo by <u>Árpád Czapp</u> on <u>Unsplash</u>

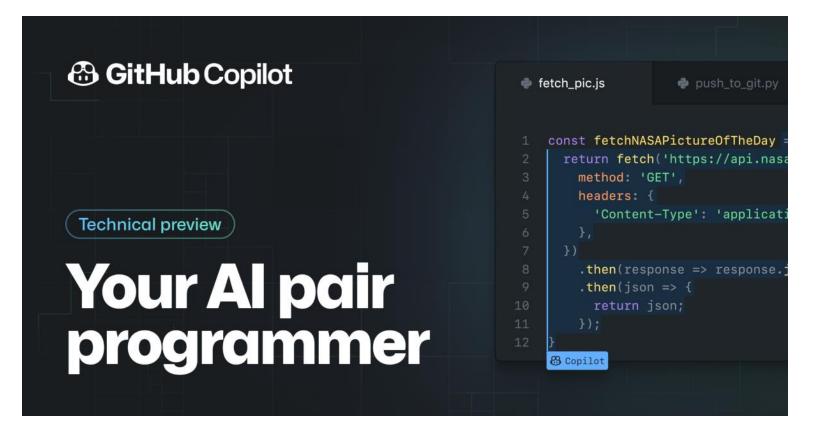






- Part 1: How we use LLMs to engineer software
- Part 2: How we engineer software to use LLMs







- Auto-completion
- Code generation based on natural language
- Building software through "pair-programming"

🔁 GitHub Copilot	fetch_pic.js
	1 const fetchNASAPictureOfTheDay
	2 return fetch('https://api.nas
	3 method: 'GET',
	4 headers: { 5 'Content-Type': 'applicat
(Technical preview)	5 'Content-Type': 'applicat
	7 })
	8 .then(response => response.
Your Al pair	9 .then(json => {
	10 return json;
	11 });
programmer	12 }
	Copilot





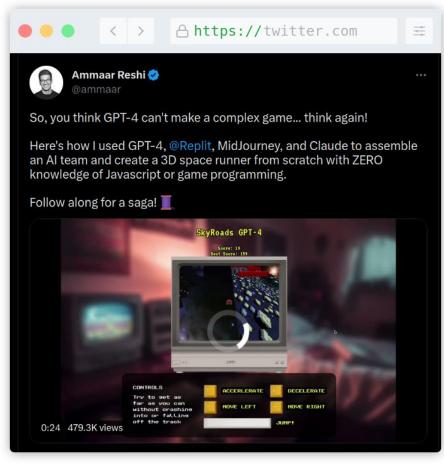












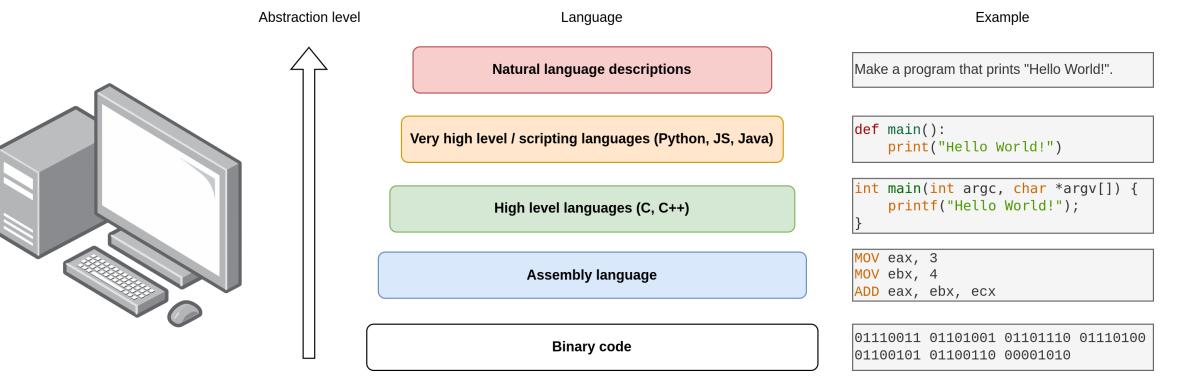
Source: twitter.com/ammaar

# Developer debugging - 6 hours Days after OpenAI Days after OpenAI ChatGPT generates Codes - 5 min Developer debugging - 24 hours

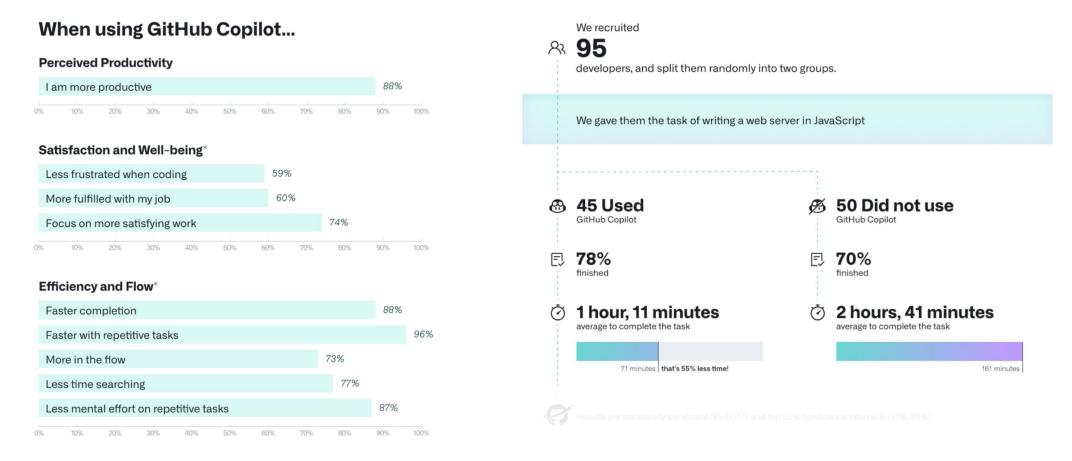


- Generative AI "disrupts" several professions, but affects programmers differently than artists
- Computer code is only valued based on its functional value





# Impact of AI on programmers: Measuring effects of GitHub Copilot



# Source: <u>GitHub blog – Quantifying GitHub Copilot's impact on developer productivity and happiness</u>



# Impact of AI on programmers

- Useful code generation
- Faster programming
- "Double-edged sword"
- Asset for experienced developers
- Liability for novice developers



GitHub Copilot AI pair programmer: Asset or Liability?

Arghavan Moradi Dakhel\*, Vahid Majdinasab\*, Amin Nikanjam, Foutse Khomh, Michel C. Desmarais Polytechnique Montreal, Montreal, Canada

Zhen Ming (Jack) Jiang York University, Toronto, Canada

#### Abstract

C 1. Introduction

ep

Automatic program synthesis is a long-lasting dream in software engineering. Recently, a promising Deep Learning (DL) based solution, called Copilot, has been proposed by OpenAI and Microsoft as an industrial product. Although some studies evaluate the correctness of Copilot solutions and report its issues, more empirical evaluations are necessary to understand how developers can benefit from it effectively. In this paper, we study the capabilities of Copilot in two different programming tasks: (i) generating (and reproducing) correct and efficient solutions for fundamental algorithmic problems, and (ii) comparing Copilot's proposed solutions with those of human programmers on a set of programming tasks. For the former, we assess the performance and functionality of Copilot in solving selected fundamental problems + in computer science, like sorting and implementing data structures. In the latter, a dataset of programming problem <sup>a</sup> with human-provided solutions is used. The results show that Copilot is capable of providing solutions for almost all fundamental algorithmic problems, however, some solutions are buggy and non-reproducible. Moreover, Copilot has some difficulties in combining multiple methods to generate a solution. Comparing Copilot to humans, our results show that the correct ratio of humans' solutions is greater than Copilot's suggestions, while the buggy solutions generated by Copilot require less effort to be repaired. Based on our findings, if Copilot is used by expert developers in software projects, it can become an asset since its suggestions could be comparable to humans' contributions in terms of quality projects, it can become an asset since its sugge "However, Conilot can become a liability if it is used by novice developers who may fail to filter its burgy or non-optima -J solutions due to a lack of expertise

Keywords: Code Completion, Language Model, GitHub Copilot, Testing

Recent breakthroughs in Deep Learning (DL), in par-ticular the Transformer architecture, have revived the Software Engineering (SE) decades-long dream of automating 4 code generation that can speed up programming activities. Program generation aims to deliver a program that meets a user's intentions in the form of input-output examples, natural language descriptions, or partial programs 2 12 33 25

former architecture recently achie in automatic program synthesis 6 8 9 20. One such model is Codex [8]; a GPT-3 [6] based language mode with up to 12 billion parameters which has been pretrained on 159 GB of code samples from 54 million GitHub reposi tories. Codex shows a good performance in solving a set of hand-written programming problems (i.e., not in the train

formal models [15, 27] to Evolutionary Algorithms [48] and machine-learned translation [12]

Novel Large Language Models (LLMs) with the trans

## Dakhel et al. (2023)



ployed in the real-world, how do they change labor productivity? While there is a growing literature studying perceptions of AI tools, how people use them, and their implications for

Peng et al. (2023)

<sup>1</sup> School of Computer Science, <sup>3</sup> School of Computing and Communications, I	, Aakash Ahmad <sup>1</sup> , Muhammad Waseem <sup>1</sup> Wahan University, Wahan, China ancaster University Leipzig, Leipzig, Germany n) wha.edu.cn, ahmad.aakash@gmail.com
A share-with the advances in Mandow I cardine, there is a proving interve code, additional advances of the share-with the strength strength of the strength of the strength strength of the st	Ables with the reserves of A minister perpendition the end of the second secon
I. INTEODUCTION	compared the effectiveness of programming with Copilot versus human programming, and found that the generated code
Large Language Models (LLMs) and Mathine Larning term ancountyping uncert cole are becoming more and more frammer proceeding (NLF) capabilities movalary (II), and Mathine Larning proceeding (NLF) capabilities movalary (II), and Mathine Larning queries have been valued on the being endpose of acceleration (II). The second second second cole and the second second second second second large anomator of each in antievant is provide to inse integration with second second second second large anomator of each in antievant is provide a more large granum of each is an antievant to provide removalities from S2D. Goldba Capital Kin second by Capability (II) and provide the second second second second second second sectors.	In Capita is inferred than limits meaturemine such. Strutture era di generosci du cognitivo quali era generatica clus D'experti in sums aboves. Capita i a sum sum sub-sus a capita i a portining toxi (Mali er di generati clus du D'experiment) and the sum

● ● ● < > 合 https://arxiv.org

Practices and Challenges of Using GitHub Copilot:

An Empirical Study



• < > Aht	tps://arxiv.o	rg =		
Assessing the Quality o	f GitHub Copile	ot's Code Generation		
Burak Yetiştiren burakyetiştiren@hotmail.com Bilkent University Ankara, Turkey	Işak Özsoy ozsoyisik@gmail.com Bilkent University Ankara, Turkey	Eray Tüzün eraytuzun@cs.bilkent.edu.tr Bilkent University Ankara, Turkey		
<b>FIRECT</b> Construct of Catability user code generation to the control for the first well established instances and an end of the first well established instances and an end of the second to a length of the second	A. Giffshan Model <sup>2</sup> Cathlub Visual Studios Cashey mundrer of and M minitor, and minitor, and mini	Model." Galifable Capitales caples cannot be an entert some start frame the transmission of the transmiss		
CONCEPTS are and its engineering → Source code gens VORDS Copilat, code generation, code completion, AI pai prioras trady offerense Tormat: efferense Tormat:	ration chufe brief and b service. We agree GitHub Copilot huncements a se Cepilot is capabl To systematical an experimenta validity, corrects	about Gibbah Capital we tranched upon only in- burritit is fould-as in terms of the evaluation of the work of the general concerns of the opinism about and ind it worthwhile to evaluate the possible en- roy like Gibbah Capitol can offer. Charly, Gibbah et al. (1990) and the state of the state of the etal of the state of the state of the state of the etal of the state		
ab Cogisti's Code Generations. In Proceedings of ACM worlds. ACM, New York, NY, USA, 11 pages, Impaction is naturation ATRODUCTION Cognited <sup>1</sup> is a code generation tool that utilizes a logies, including a compatible IDE, and the Open without.	rg (0.1143) RQ1 What is the RQ1.1 How v RQ1.2 How v RQ1.3 How v RQ2 What is th ratio code quality?	equility of the code generated by Github Copilor? valid are Github Copilor's code suggestions? strent are Github Copilor's code suggestions? efficient are Github Copilor's code suggestions? effect of using the docstings on the generated effect of using appropriate function names on the		
protections on to make digital or hard copies of all or part of this work for an or systemic distinct for peopled that sopies are not could or communic advantage architecture bare the notice and the best of the source of the software to be source and the source of the source. A software to be source and the source of the software are presentation to be presentation to be source and the presist art presentation from presentation and goes a possible presist art presentations. These presentations are goes a possible presist of the source of the software presentation and goes a possible presist of the source of the software presentation are goes a possible presist of the source of the software presentation are goes a possible presist of the source of the software presentation are goes a possible presist of the source of the software presentation are goes a possible presist of the source of the software presentation are goes a possible presist of the software presentation and the software possible presist of the software possible presentation and the source of the software possible possibl	e personal or or dateband e ful citation million ACM or opablish, an and or a	quanty partitions, we first elaborate on our experimental 2. In Section 3, we present the results we gathered In Section 4, we share and evaluate our results, in com factors that might influence the validity of our side an overview of other works that study (Selful) on 6. Lastly, we conclude our study in Section 7.		

Yetistiren et al. (2022)



- Knowing which code to accept from the AI tool
- Knowing what prompts to use

Why should I hire a software engineer if I can just copy and paste code from Stack Overflow?



## Jessica Su, CS PhD student at Stanford

2+

Answered Dec 28 · Upvoted by Rupak Hattikudur, Software Engineer @ L&T Infotech and Terry Lambert, Senior Software Engineer: Novell, Artisoft, IBM, Array Netw...

It's still worth the money. The breakdown is

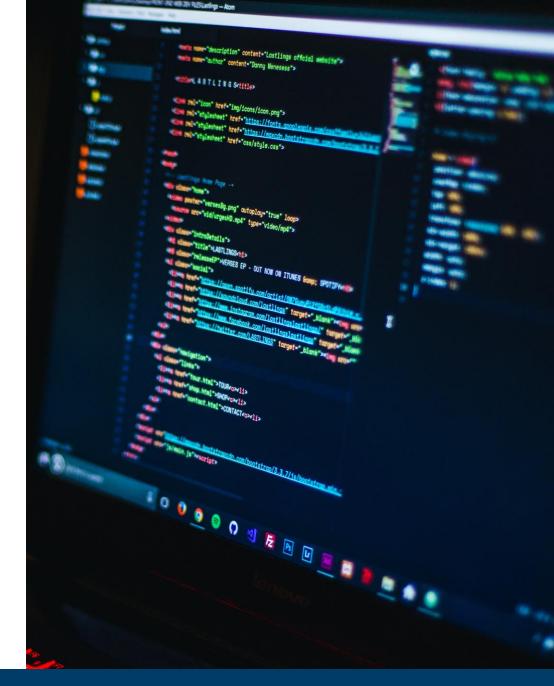
- Copying code from StackOverflow:
   \$1
- Knowing which code to copy from StackOverflow: \$100000/year



# The future of software engineering using LLMs

- LLMs for programming -> Faster software engineering
- Software engineers are needed for
  - Design choices
  - Specifications of requirements and needs

# **Part 2**: How do we engineer software to use LLMs?





- GPT-4 shows impressive capabilities
- Paper is both praised and criticized
- Does the model "understand"?

#### Sparks of Artificial General Intelligence: Early experiments with GPT-4

Sébastien Bubeck Varun Chandrasekaran Ronen Eldan Johannes Gehrke Eric Horvitz Ece Kamar Peter Lee Yin Tat Lee Yuanzhi Li Scott Lundberg Harsha Nori Hamid Palangi Marco Tulio Ribeiro Yi Zhang

Microsoft Research

#### Abstract

Artificial intelligence (AI) researchers have been developing and refining large language models (LLMs) that exhibit remarkable capabilities across a variety of domains and tasks, challenging our understanding of learning and cognition. The latest model developed by OpenAI, GPT-4 [Ope23], was trained using an unprecedented scale of compute and data. In this paper, we report on our investigation of an early version of GPT-4, when it was still in active development by OpenAI. We contend that (this early version of) GPT-4 is part of a new cohort of LLMs (along with ChatGPT and Google's PaLM for example) that exhibit more general intelligence than previous AI models. We discuss the rising capabilities and implications of these models. We demonstrate that, beyond its mastery of language, GPT-4 can solve novel and difficult tasks that span mathematics, coding, vision, medicine, law, psychology and more, without needing any special prompting. Moreover, in all of these tasks, GPT-4's performance is strikingly close to human-level performance, and often vastly surpasses prior models such as ChatGPT. Given the breadth and depth of GPT-4's capabilities, we believe that it could reasonably be viewed as an early (vet still incomplete) version of an artificial general intelligence (AGI) system. In our exploration of GPT-4, we put special emphasis on discovering its limitations, and we discuss the challenges ahead for advancing towards deeper and more comprehensive versions of AGI, including the possible need for pursuing a new paradigm that moves beyond next-word prediction. We conclude with reflections on societal influences of the recent technological leap and future research directions.

#### Contents

2023

Apr

13

CL]

[CS.

5

arXiv:2303.12712v

1	Intr 1.1 1.2	roduction       4         Our approach to studying GPT-4's intelligence       6         Organization of our demonstration       8	
2	Mu	ltimodal and interdisciplinary composition 13	
	2.1	Integrative ability	
	2.2	Vision	
		2.2.1 Image generation beyond memorization	
		2.2.2 Image generation following detailed instructions (à la Dall-E)	
		2.2.3 Possible application in sketch generation	
	2.3	Music	
3		ding 21	
	3.1	From instructions to code	
		3.1.1 Coding challenges	
		3.1.2 Real world scenarios	
	3.2	Understanding existing code	

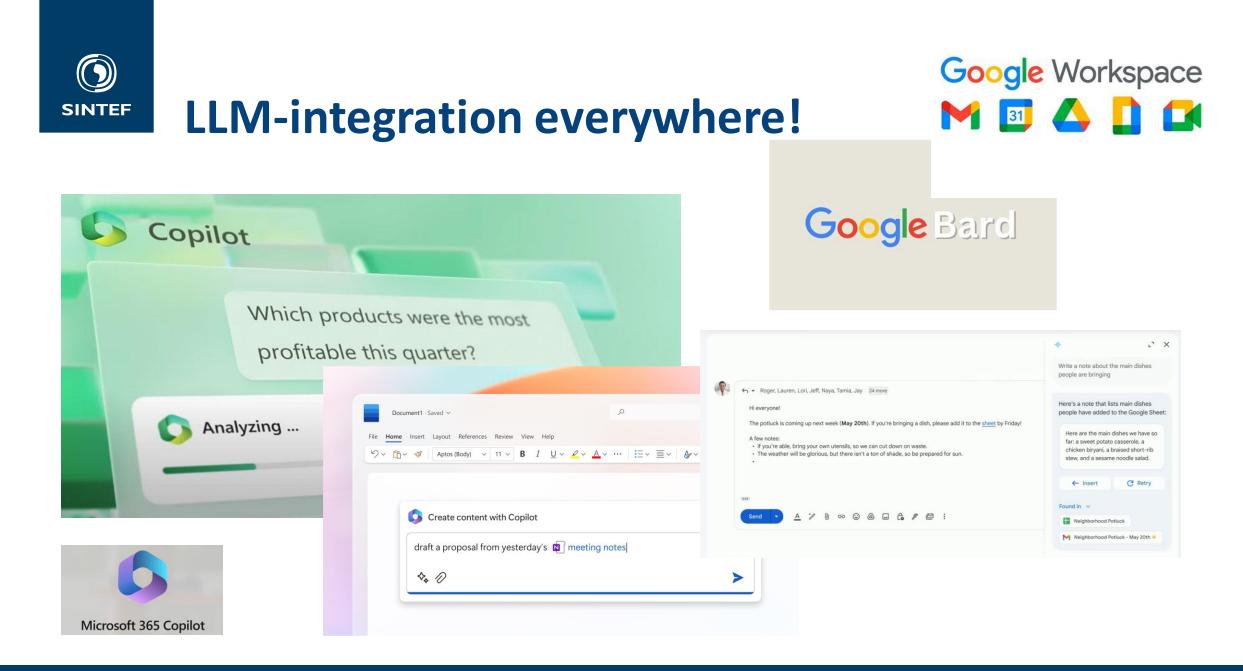
Bubeck et al. (2023)



To a person with a hammer, everything looks like a nail.

- To a person with an LLM, everything looks like a language problem.
- We should search for solutions to problems, not search for problems on which to apply our solution





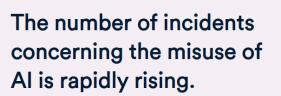






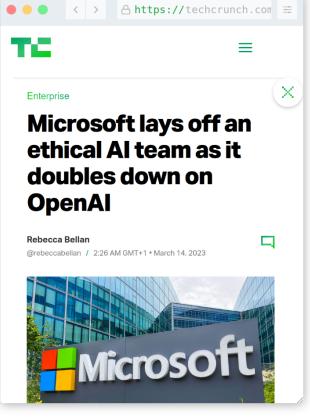
# **Ethical issues**

# THE AI INDEX REPORT Measuring trends in Artificial Intelligence



According to the AIAAIC database, which tracks incidents related to the ethical misuse of AI, the number of AI incidents and controversies has increased 26 times since 2012. Some notable incidents in 2022 included a deepfake video of Ukrainian President Volodymyr Zelenskyy surrendering and U.S. prisons using callmonitoring technology on their inmates. This growth is evidence of both greater use of AI technologies and awareness of misuse possibilities.

Source: Al Index Report 2023



Source: TechCrunch.com



Intelligence Index

Stanford University Human-Centered Artificial Intelligence



# **Ethical issues**

# THE ALINDEX REPORT Measuring trends in Artificial Intelligence





Artificial Intelligence Index **Stanford University** Human-Centered Artificial Intelligence

# The number of incidents concerning the misuse of Al is rapidly rising.

According to the AIAAIC database, which tracks incidents related to the ethical misuse of AI, the number of AI incidents and controversies has increased 26 times since 2012. Some notable incidents in 2022 included a deepfake video of Ukrainian President Volodymyr Zelenskyy surrendering and U.S. prisons using callmonitoring technology on their inmates. This growth is evidence of both greater use of AI technologies and awareness of misuse possibilities.

# Al is both helping and harming the environment.

New research suggests that AI systems can have serious environmental impacts. According to Luccioni et al., 2022, BLOOM's training run emitted 25 times more carbon than a single air traveler on a one-way trip from New York to San Francisco. Still, new reinforcement learning models like BCOOLER show that AI systems can be used to optimize energy usage. The demand for Alrelated professional skills is increasing across virtually every American industrial sector.

Across every sector in the United States for which there is data (with the exception of agriculture, forestry, fishery and hunting), the number of Al-related job postings has increased on average from 1.7% in 2021 to 1.9% in 2022. Employers in the United States are increasingly looking for workers with Alrelated skills.



# The dark side of LLMs

ΞQ The Washington Post Sign in

Help Desk Tech in Your Life Future of Work

#### WHAT'S NEW

## Snapchat tried to make a safe AI. It chats with me about booze and sex.

Our tech columnist finds Snapchat can't control its new My Al chatbot friend. Tech companies shouldn't treat users as test subjects - especially young ones.



Perspective by Geoffrey A. Fowler Columnist | + Follow

March 14, 2023 at 9:00 a.m. EDT



The Washington Post



**BUSINESS • TECHNOLOGY** 

Exclusive: OpenAI Used

Kenyan Workers on Less

Than \$2 Per Hour to

Make ChatGPT Less

Toxic

This image was generated by OpenAI's image-

generation software, Dall-E 2. The prompt was: "A

front of computer screens in a printmaking style."

seemingly endless view of African workers at desks in

Time

The New Hork Times =

A.I. and Test A.I.'s Literary Skills Spot the A.I. Chatbots

## Lawsuit Takes Aim at the Way A.I. Is Built

A programmer is suing Microsoft, GitHub and OpenAI over artificial intelligence technology that generates its own computer code.

119 Give this article 



Tom Smith, a veteran programmer, shows how Codex can instantly generate computer code from a request in plain English.

The New York Times

### On the Impossible Safety of Large AI Models

El-Mahdi El-Mhamdi<sup>1,2</sup>, Sadegh Farhadkhani<sup>3</sup>, Rachid Guerraoui<sup>3</sup>, Nirupam Gupta<sup>3</sup>, Lê-Nguyên Hoang<sup>2,4</sup>, Rafaël Pinot<sup>3</sup>, Sébastien Rouault<sup>2</sup>, and John Stephan<sup>3</sup>

> <sup>1</sup>École Polytechnique <sup>2</sup>Calicarpa <sup>3</sup>EPFL <sup>4</sup>Tournesol Association

#### Abstract

Large AI Models (LAIMs), of which large language models are the most prominent recent example, showcase some impressive performance. However they have been empirically found to pose serious security issues. This paper systematizes our knowledge about the fundamental impossibility of building arbitrarily accurate and secure machine learning models. More precisely, we identify key challenging features of many of today's machine learning settings. Namely, high accuracy seems to require memorizing large training datasets, which are often user-generated and highly heterogeneous, with both sensitive information and fake users. We then survey statistical lower bounds that, we argue, constitute a compelling case against the possibility of designing high-accuracy LAIMs with strong security guarantees.

El-Mhamdi et al. (2023)

## "To scientists and journalists:

[...] The current academic focus on algorithmic performance, and its inattention to social impacts, are endangering our societies



# Intelligence and understanding

• •	●	
	Sparks of Artificial General Intelligence: Early experiments with GPT-4	
	Sébastien Bubeck Varun Chandrasekaran Ronen Eldan Johannes Gehrke Eric Horvitz Ece Kamar Peter Lee Yin Tat Lee Yuanzhi Li Scott Lundber Harsha Nori Hamid Palangi Marco Tulio Ribeiro Yi Zhang	rg
	Microsoft Research	
arXiv:2303.12712v5 [cs.CL] 13 Apr 2023	Abstract Artificial intelligence (AI) researchers have been developing and refining large language models (LLMs) that exhibit remarkable capabilities across a variety of domains and tasks, challenging our understanding of learning and cognition. The lastst model developed by OpenAI, GPT4 [Ope2], sus trained using an of GPT4, ether and a cognition on the active development by OpenAI. We contend that (this acrily version of OPT4 is part of a new color of LLMs (along with ChatGPT and Google's PLAI for example, that exhibit more general infeligence than previous AI models. We discuss the rising capabilities and implications of tasks that apart of a new color of LLMs (along with ChatGPT and Google's PLAI for example, that exhibit have general infeligence than previous AI models. We discuss the rising capabilities and implications of tasks that apan mathematics, coding, vision, medicine, law, poychology and mark, without needing any part of a new color of LLMs (along the tasks, GPT-4's performance's sufficiency close to busines)-level of GPT-4, which does not show that is of models by the or GPT-4 can solve novel and difficult of a neutrical general intelligence (AGO) reptem. In our explorations of GPT-4, we put opped by beying of microwing its limitations of AGI, including the possible need for partnering a new paradigm that mores bypound on discovering its limitations of AGI, including the possible need for paraming a new paradigm that mores bypound neutrowing prediction. We conclude with reflections on societal influences of the recent technological leap and future research directions.	
03.	Contents	
čiv:23	1 Introduction         4           1.1 Our approach to studying GPT-4's intelligence         6           1.2 Organization of our demonstration         8	
arX	2 Multimodal and interdisciplinary composition	

		nization of our demonstration	
2	Multimod	dal and interdisciplinary composition	1
	2.1 Integr	rative ability	. [
	2.2 Vision	n	. 🛛
	2.2.1	Image generation beyond memorization	
	2.2.2	Image generation following detailed instructions (à la Dall-E)	. 0
	2.2.3		
	2.3 Music		- D
3	Coding		2
	3.1 From	instructions to code	
	3.1.1	Coding challenges	. 0
	3.1.2	Real world scenarios	. [
	3.2 Under	rstanding existing code	. [2

#### A https://aclanthology.org

#### **Climbing towards NLU:** On Meaning, Form, and Understanding in the Age of Data

Emily M. Bender University of Washington Department of Linguistics ebender@uw.edu

Alexander Koller Saarland University Dept. of Language Science and Technology koller@coli.uni-saarland.de

#### Abstract

The success of the large neural language models on many NLP tasks is exciting. However, we find that these successes sometimes lead to hype in which these models are being described as "understanding" language or capturing "meaning". In this position paper, we argue that a system trained only on form has a priori no way to learn meaning. In keeping with the ACL 2020 theme of "Taking Stock of Where We've Been and Where We're Going", we argue that a clear understanding of the distinction between form and meaning will help guide the field towards better science around natural language understanding.

#### 1 Introduction

The current state of affairs in NLP is that the large neural language models (LMs), such as BERT (Devlin et al., 2019) or GPT-2 (Radford et al., 2019), are making great progress on a wide range of tasks, including those that are ostensibly meaningsensitive. This has led to claims, in both academic and popular publications, that such models "understand" or "comprehend" natural language or learn its "meaning". From our perspective, these are overclaims caused by a misunderstanding of the relationship between linguistic form and meaning. We argue that the language modeling task, be-

the structure and use of language and the ability to ground it in the world. While large neural LMs may well end up being important components of an eventual full-scale solution to human-analogous NLU, they are not nearly-there solutions to this grand challenge. We argue in this paper that genuine progress in our field - climbing the right hill, not just the hill on whose slope we currently sitdepends on maintaining clarity around big picture notions such as meaning and understanding in task design and reporting of experimental results. After briefly reviewing the ways in which large LMs are spoken about and summarizing the recent flowering of "BERTology" papers (§2), we offer a working definition for "meaning" (§3) and

a series of thought experiments illustrating the impossibility of learning meaning when it is not in the training signal  $(\S4,5)$ . We then consider the human language acquisition literature for insight into what information humans use to bootstrap language learning (§6) and the distributional semantics literature to discuss what is required to ground distributional models (§7). §8 presents reflections on how we look at progress and direct research effort in our field, and in §9, we address possible counterarguments to our main thesis.

#### 2 Large LMs: Hype and analysis

cause it only uses form as training data, cannot in Publications talking about the application of large principle lead to learning of meaning. We take the LMs to meaning-sensitive tasks tend to describe term language model to refer to any system trained the models with terminology that, if interpreted at

## A https://dl.acm.org **On the Dangers of Stochastic Parrots:** Can Language Models Be Too Big? 🂐

Emily M. Bender' ebender@uw.edu University of Washington Seattle WA USA

Angelina McMillan-Major avmm@uw.edu University of Washington Seattle, WA, USA

#### ABSTRACT

The past 3 years of work in NLP have been characterized by the development and deployment of ever larger language models, especially for English. BERT, its variants, GPT-2/3, and others, most recently Switch-C, have pushed the boundaries of the possible both through architectural innovations and through sheer size. Using these pretrained models and the methodology of fine-tuning them for specific tasks, researchers have extended the state of the art on a wide array of tasks as measured by leaderboards on specific benchmarks for English. In this paper, we take a step back and ask: How big is too big? What are the possible risks associated with this technology and what paths are available for mitigating those risks? We provide recommendations including weighing the environmental and financial costs first, investing resources into curating and carefully documenting datasets rather than ingesting everything on the web, carrying out pre-development exercises evaluating how the planned approach fits into research and development goals and supports stakeholder values, and encouraging research directions beyond ever larger language models.

#### CCS CONCEPTS

#### • Computing methodologies $\rightarrow$ Natural language processing. ACM Reference Format:

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? & . In Conference on Fairness, Accountability, and Trans parency (FAccT '21), March 3-10, 2021, Virtual Event, Canada. ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3442188.3445922

#### 1 INTRODUCTION

One of the biggest trends in natural language processing (NLP) has been the increasing size of language models (LMs) as measured by the number of parameters and size of training data. Since 2018 Toint first authors

Timnit Gebru\* timnit@blackinai.org Black in AI Palo Alto, CA, USA

Shmargaret Shmitchell shmargaret.shmitchell@gmail.com The Aether

alone, we have seen the emergence of BERT and its variants [39, 70, 74, 113, 146], GPT-2 [106], T-NLG [112], GPT-3 [25], and most recently Switch-C [43], with institutions seemingly competing to produce ever larger LMs. While investigating properties of LMs and how they change with size holds scientific interest, and large LMs have shown improvements on various tasks (§2), we ask whether enough thought has been put into the potential risks associated with developing them and strategies to mitigate these risks.

We first consider environmental risks. Echoing a line of recent work outlining the environmental and financial costs of deep learning systems [129], we encourage the research community to prioritize these impacts. One way this can be done is by reporting costs and evaluating works based on the amount of resources they consume [57]. As we outline in §3, increasing the environmental and financial costs of these models doubly punishes marginalized communities that are least likely to benefit from the progress achieved by large LMs and most likely to be harmed by negative environmental consequences of its resource consumption. At the scale we are discussing (outlined in §2), the first consideration should be the environmental cost.

Just as environmental impact scales with model size, so does the difficulty of understanding what is in the training data. In §4, we discuss how large datasets based on texts from the Internet overrepresent hegemonic viewpoints and encode biases potentially damaging to marginalized populations. In collecting ever larger datasets we risk incurring documentation debt. We recommend mitigating these risks by budgeting for curation and documentation at the start of a project and only creating datasets as large as can be sufficiently documented.

As argued by Bender and Koller [14], it is important to understand the limitations of LMs and put their success in context. This not only helps reduce hype which can mislead the public and researchers themselves regarding the capabilities of these LMs, but might encourage new research directions that do not necessarily depend on having larger LMs. As we discuss in §5, LMs are not









- Sam Altman (OpenAI CEO) in the podcast "Honestly with Bari Weiss", 27.04.2023:
  - "People have this experience the first time they talk to ChatGPT which is like: 'Wow! The experts, the linguists, they can argue about the definition of the word 'understanding', but it *feels like* this thing understands me, it feels like this things trying to help [...]', and that's powerful."

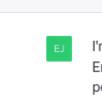




- Al tools for programming are...
  - ...no worse than using a calculator instead of mental arithmetic, or
  - ...as bad as physical inactivity for the body?







### < > A https://chat.openai.com

Model: GPT-5

I'm making a presentation on "The Future of Software Engineering in the light of LLMs". Can you provide some perspectives on this topic?

As an Al model, I can help you with tasks related to language, but when it comes to generating original content you would be better off making your own presentations instead of making a word calculator do it for you.

### Send a message.

G

-

ChatGPT may produce inaccurate information about people, places, or facts. ChatGPT May 3 Version

# So, I asked ChatGPT...



- Software engineers are not replaced, but augmented
- A new skill set for using AI and LLMs
- Active discussion around the ethics of LLMs



There's too much effort trying to create autonomous machines rather than trying to create machines that are useful tools for humans.

– Prof. Emily M. Bender





Link to slides

- How can we (and should we) integrate LLM-based tools into our current software engineering practices at SINTEF?
- What are the benefits and potential pitfalls of increasing our use of LLMs?

