研究ができる人工知能の実現へ向けた課題の検討

2023/10/21@上智大学


Edit profile
Shiro Takagi
@takagi_shiro
独立して個人で機械学習の研究をしてます。研究ができる人工知能の実現が目標です。 研究のあり方にも関心があり、研究自体について podcast やったり note 書いたりもしてます。 linktr.ee/shiro_takagi
Science & Technology
t46.github.io
Joined December 2018

Autores

Home About Roadmap

熊谷亘 (東大)

高木志郎 (独立研究者)

麻生英樹 (産総研)

三内顕義 (理研)

園田翔 (理研)

鈴木雅大 (東大)

0. はじめに

（汎用的で自律的な）
研究できる AI を実現するため
にはどんな課題があるか？

研究はどこまで人間から自由になれるか？

1. 研究ができる AI とは？
2. 研究の自動化の試みは？
3. 課題は？

今日の話は
コーヒーブレイクぐらいの
気持ちで聞いてください☕

1. 研究ができる AI とは？

研究とは

ある社会にとっての新しい知識を

生産する行為？

知識とは

正当化された真なる信念？

[Steup+ 2020]

知識の生産とは

信念を更新すること？

研究で生み出される知識は
個人にとどまらず人類社会全体の知識
→ **共通信念の更新？**

知識が新しいとは

ある問いに対して真であると強く信じられる
仮説がそれまで存在しなかったということ？

この定義から得られる示唆は？

a. 研究は問いの構築+仮説生成+仮説検証？
b. 有意味な知識生産の自律的実行は困難？
c. 研究を知識生産とする定義は不適当？

知識は新規である必要がある

→ **問いの構築**

問いに対して答えを出す（答えは未知）

→ **仮説の生成**

仮説が真であるという信念を正当化する必要

→ **仮説の検証**

知識が信念ならば主体/社会に対して相対的
→ **AI が自律的に生産した知識は人には無意味**
（検証をゼロから自律的に構成した場合）

人の信念は長期間の自然との相互作用の産物
→ 人は信念に基づいても自然理解につながる
→ **AI の自律的検証は自然理解に向かわない？**

- 人間の知識に限定する
- 自然と整合的な知識の基盤を考える
- 研究を新しいパターンの発見と捉える
- 認識論的プラグマティズム

研究を知識生産とすべきではない？
研究を包括的に定義するべきではない？
信念より自然との整合性を強調する？
帰納推論の妥当性は前提とする？
研究を自然の新規なパターンの発見とする？
研究をより良い制御をもたらすものとする？
etc...

2. 研究の自動化の試みは？

Artificial Intelligence 61 (1993) 209–261 209
Elsevier

ARTINT 959

DENDRAL: a case study of the first expert system for scientific hypothesis formation*

Robert K. Lindsay
University of Michigan, 205 Zina Pitcher Place, Ann Arbor, MI 48109, USA

Bruce G. Buchanan
Computer Science Department, University of Pittsburgh, Pittsburgh, PA 15260, USA

Edward A. Feigenbaum
Knowledge Systems Laboratory, Department of Computer Science, Stanford University, Stanford, CA 94305, USA

Joshua Lederberg
Rockefeller University, New York, NY 10021-6399, USA

Received April 1991
Revised June 1992

Abstract

Lindsay, R.K., B.G. Buchanan, E.A. Feigenbaum and J. Lederberg, DENDRAL: a case study of the first expert system for scientific hypothesis formation, Artificial Intelligence 61 (1993) 209–261.

The DENDRAL Project was one of the first large-scale programs to embody the strategy of using detailed, task-specific knowledge about a problem domain as a source of heuristics, and to seek generality through automating the acquisition of such knowledge. This paper summarizes the major conceptual contributions and accomplishments of that project. It is an attempt to distill from this research the lessons that are of importance to artificial intelligence research and to provide a record of the final status of two decades of work.

[Lindsay+ 1993]

BACON: A PRODUCTION SYSTEM THAT DISCOVERS EMPIRICAL LAWS

Patrick W. Langley
Carnegie-Mellon University
Pittsburgh, Pennsylvania 15213

Introduction

In recent years researchers have produced a number of programs capable of scientific-like behavior. Each of the systems, DENDRAL(Feigenbaum and Lederberg, 1971), meta-DENDRAL(Buchanan, Feigenbaum and Sridharan, 1972), MYCIN (Davis, Buchanan and Shortliffe, 1975) and AM (Lenat, 1976) could arrive at rules that explained observed data. In this paper I discuss another system, BACON, which discovers simple empirical laws like those found by early physicists.

[Langley 1977]

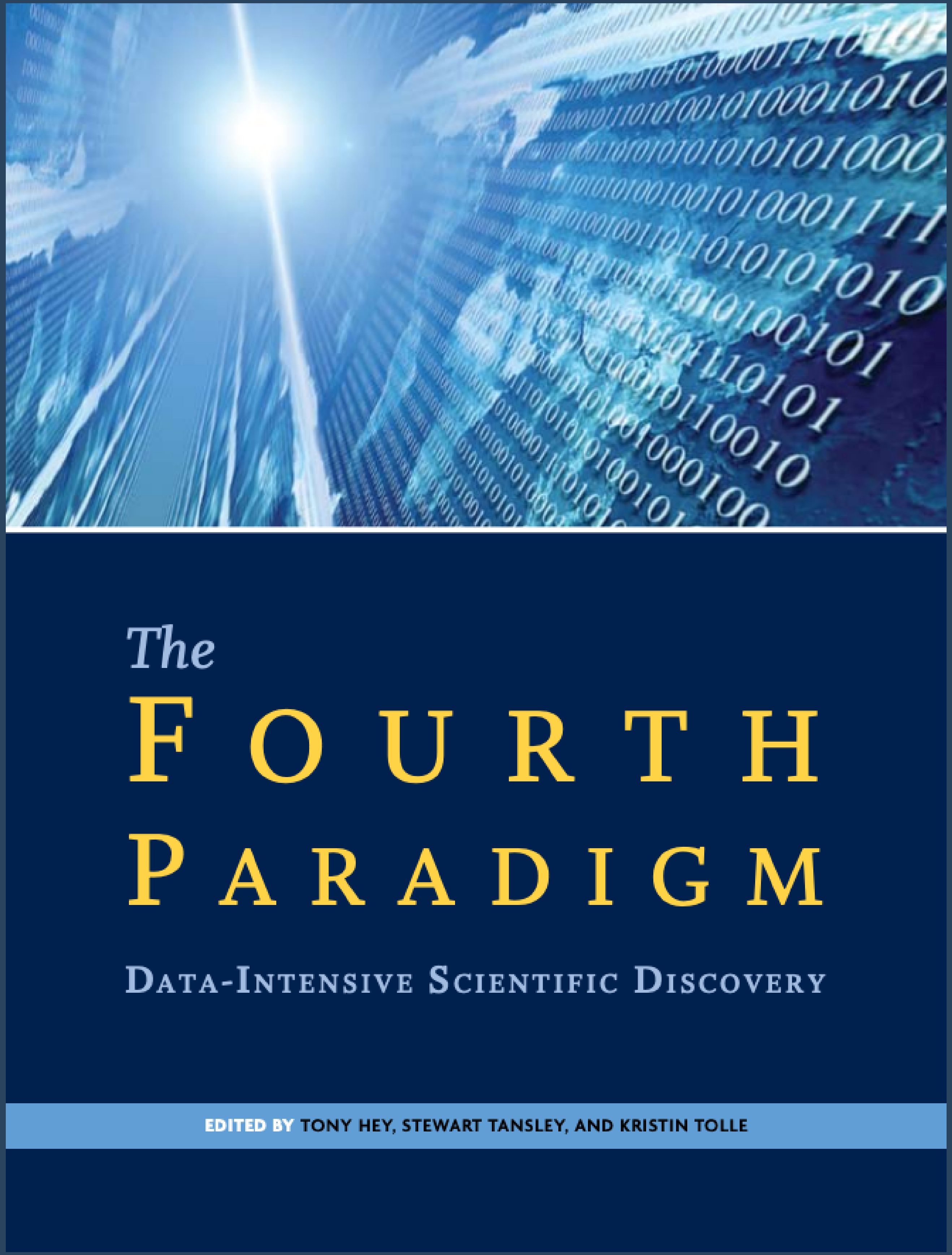
Functional genomic hypothesis generation and experimentation by a robot scientist

Ross D. King[1], Kenneth E. Whelan[1], Ffion M. Jones[1], Philip G. K. Reiser[1], Christopher H. Bryant[2], Stephen H. Muggleton[3], Douglas B. Kell[4] & Stephen G. Oliver[5]

[1]*Department of Computer Science, University of Wales, Aberystwyth SY23 3DB, UK*
[2]*School of Computing, The Robert Gordon University, Aberdeen AB10 1FR, UK*
[3]*Department of Computing, Imperial College, London SW7 2AZ, UK*
[4]*Department of Chemistry, UMIST, P.O. Box 88, Manchester M60 1QD, UK*
[5]*School of Biological Sciences, University of Manchester, 2.205 Stopford Building, Manchester M13 9PT, UK*

The question of whether it is possible to automate the scientific process is of both great theoretical interest[1,2] and increasing practical importance because, in many scientific areas, data are being generated much faster than they can be effectively analysed. We describe a physically implemented robotic system that applies techniques from artificial intelligence[3–8] to carry out cycles of scientific experimentation. The system automatically

[King+ 2004]



[Hey+ 2009]

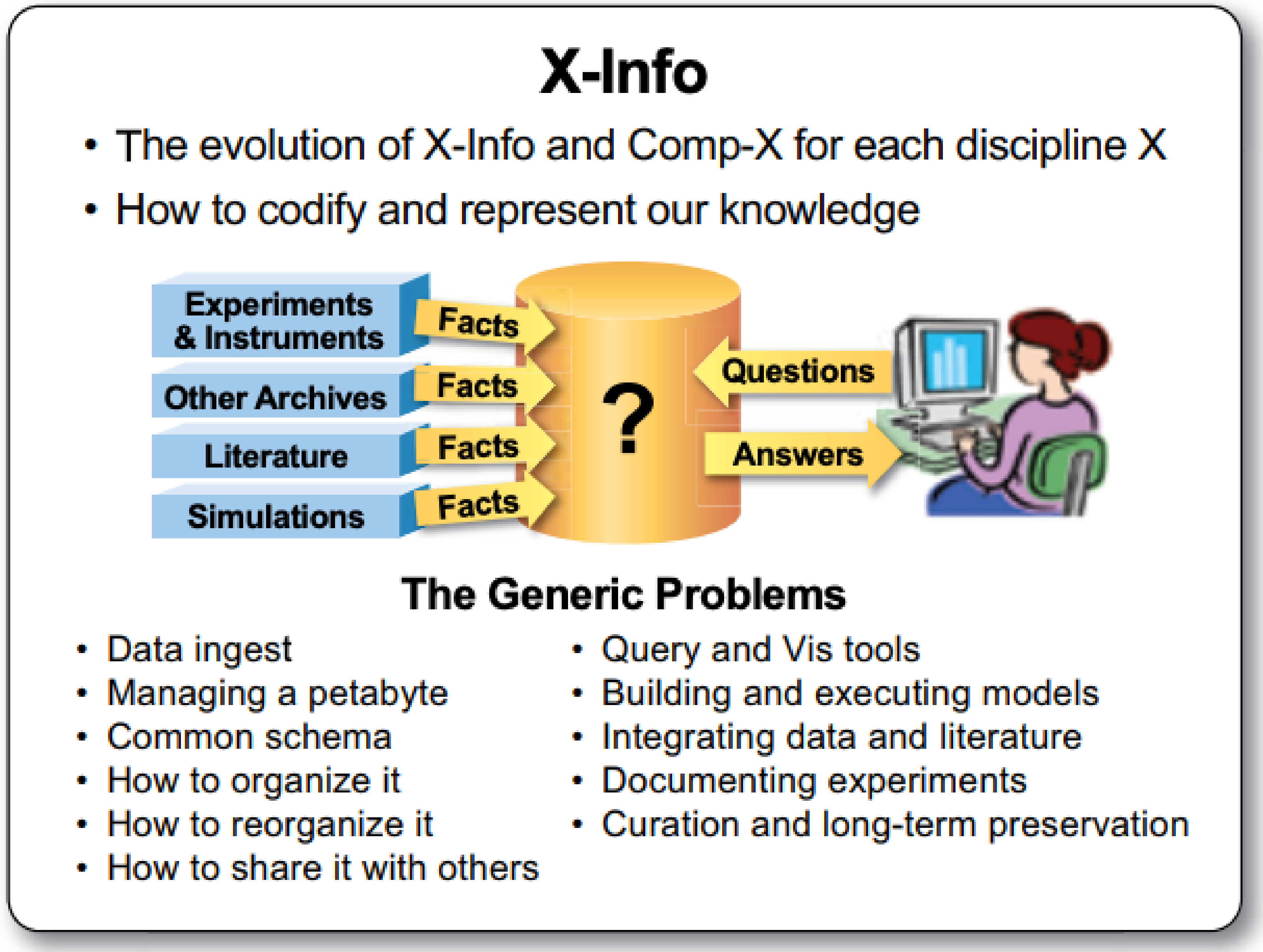


FIGURE 2

[Grey 2009]

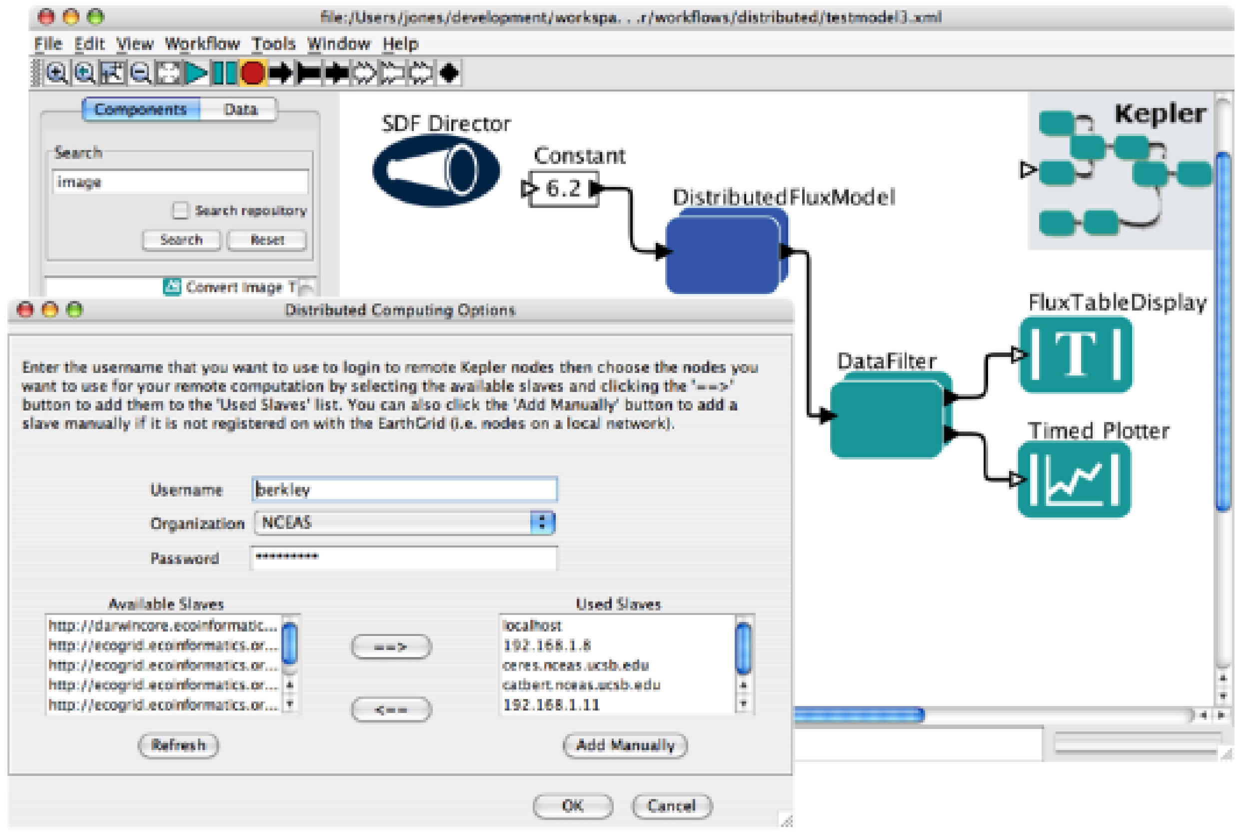


Figure 13.2: Kepler supports execution of workflows on remote peer nodes and remote clusters. Users indicate which portions of a workflow should be remotely executed by grouping them in a distributed composite component (shown in blue in the workflow). The user selects from a list of available remote nodes for execution (see dialog), and Kepler calculates a schedule and stages each data token before execution on one of the set of selected remote nodes.

[Lud¨ascher+ 2009]

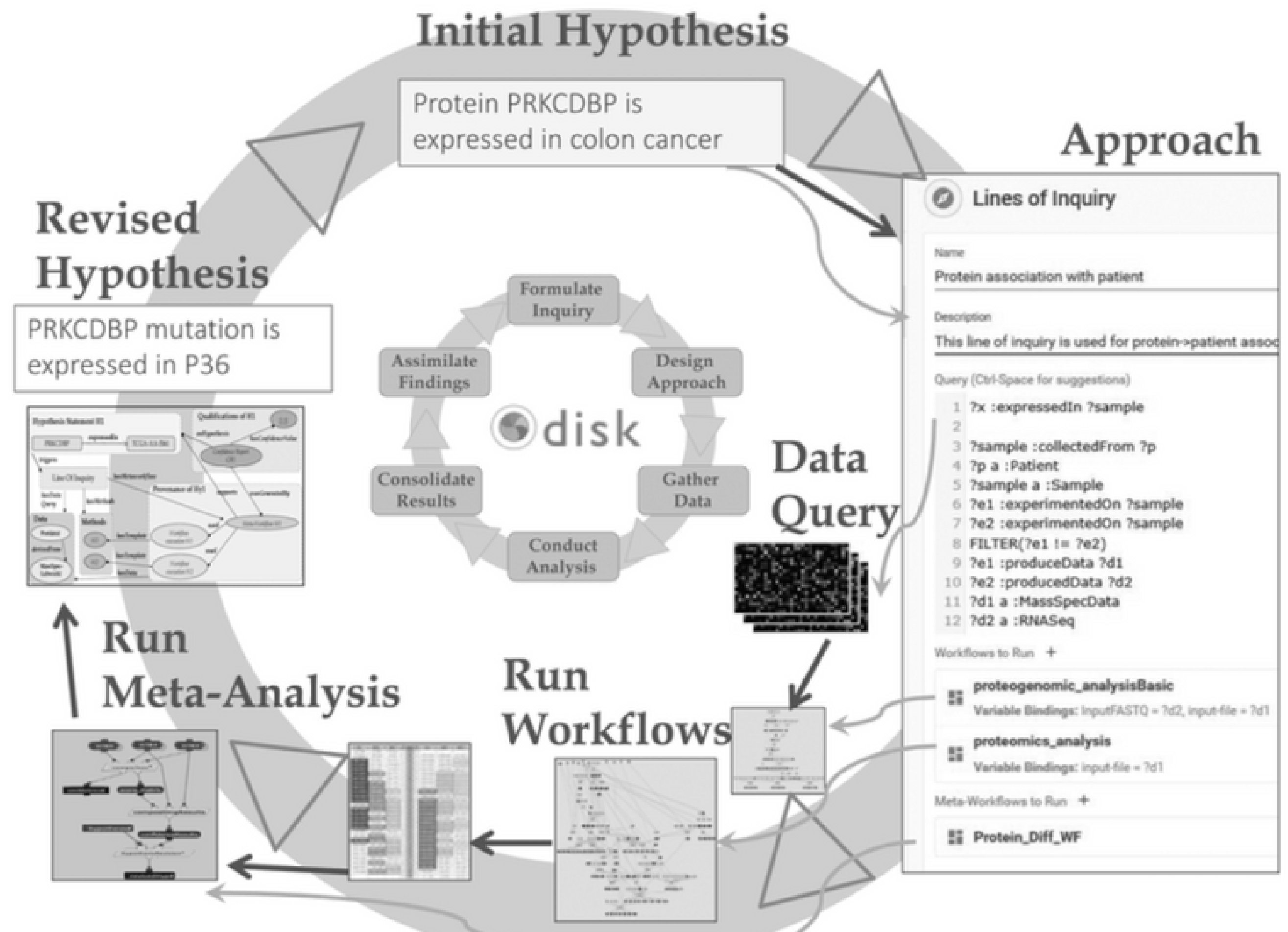


FIGURE 3 Overview of DISK, an autonomous system for hypothesis-driven discovery that relies on lines of inquiry specified by scientists about the datasets and methods they seek when they pursue specific kinds of hypotheses or questions

[Gil 2021]

Research

AlphaFold: a solution to a 50-year-old grand challenge in biology

November 30, 2020

[https://www.deepmind.com/blog/alphafold-a-solution-to-a-50-year-old-grand-challenge-in-biology]

Review

Scientific discovery in the age of artificial intelligence

https://doi.org/10.1038/s41586-023-06221-2

Received: 30 March 2022

Accepted: 16 May 2023

Published online: 2 August 2023

Check for updates

Hanchen Wang[1,2,37,38,39], Tianfan Fu[3,39], Yuanqi Du[4,39], Wenhao Gao[5], Kexin Huang[6], Ziming Liu[7], Payal Chandak[8], Shengchao Liu[9,10], Peter Van Katwyk[11,12], Andreea Deac[9,10], Anima Anandkumar[2,13], Karianne Bergen[11,12], Carla P. Gomes[4], Shirley Ho[14,15,16,17], Pushmeet Kohli[18], Joan Lasenby[1], Jure Leskovec[6], Tie-Yan Liu[19], Arjun Manrai[20], Debora Marks[21,22], Bharath Ramsundar[23], Le Song[24,25], Jimeng Sun[26], Jian Tang[9,27,28], Petar Veličković[17,29], Max Welling[30,31], Linfeng Zhang[32,33], Connor W. Coley[5,34], Yoshua Bengio[9,10] & Marinka Zitnik[20,22,35,36] ✉

Artificial intelligence (AI) is being increasingly integrated into scientific discovery to augment and accelerate research, helping scientists to generate hypotheses, design experiments, collect and interpret large datasets, and gain insights that might not have been possible using traditional scientific methods alone. Here we examine breakthroughs over the past decade that include self-supervised learning, which allows models to be trained on vast amounts of unlabelled data, and geometric deep learning, which leverages knowledge about the structure of scientific data to enhance model accuracy and efficiency. Generative AI methods can create designs, such as small-molecule drugs and proteins, by analysing diverse data modalities, including images and sequences. We discuss how these methods can help scientists throughout the scientific process and the central issues that remain despite such advances. Both developers and users of AI toolsneed a better understanding of when such approaches need improvement, and challenges posed by poor data quality and stewardship remain. These issues cut across scientific disciplines and require developing foundational algorithmic approaches that can contribute to scientific understanding or acquire it autonomously, making them critical areas of focus for AI innovation.

[Wang+ 2023]

Artificial Intelligence for Science in Quantum, Atomistic, and Continuum Systems

Xuan Zhang[1,*] Limei Wang[1,*] Jacob Helwig[1,*] Youzhi Luo[1,*] Cong Fu[1,*] Yaochen Xie[1,*] Meng Liu[1] Yuchao Lin[1] Zhao Xu[1] Keqiang Yan[1] Keir Adams[2] Maurice Weiler[3] Xiner Li[1] Tianfan Fu[4] Yucheng Wang[5] Haiyang Yu[1] YuQing Xie[6] Xiang Fu[6] Alex Strasser[7] Shenglong Xu[8] Yi Liu[9] Yuanqi Du[10] Alexandra Saxton[1] Hongyi Ling[1] Hannah Lawrence[6] Hannes Stärk[6] Shurui Gui[1] Carl Edwards[4] Nicholas Gao[11] Adriana Ladera[6] Tailin Wu[12] Elyssa F. Hofgard[6] Aria Mansouri Tehrani[6] Rui Wang[13] Ameya Daigavane[6] Montgomery Bohde[1] Jerry Kurtin[1] Qian Huang[12] Tuong Phung[6] Minkai Xu[12] Chaitanya K. Joshi[14] Simon V. Mathis[14] Kamyar Azizzadenesheli[15] Ada Fang[16] Alán Aspuru-Guzik[17,18] Erik Bekkers[3] Michael Bronstein[19] Marinka Zitnik[20] Anima Anandkumar[15,21] Stefano Ermon[12] Pietro Liò[14] Rose Yu[13] Stephan Günnemann[11] Jure Leskovec[12] Heng Ji[4] Jimeng Sun[4] Regina Barzilay[6] Tommi Jaakkola[6] Connor W. Coley[2,6] Xiaoning Qian[1,5,22] Xiaofeng Qian[7,5,8] Tess Smidt[6] Shuiwang Ji[1,+]

[Zhang+ 2023]

Autonomous discovery in the chemical sciences part I: Progress

Connor W. Coley*†, Natalie S. Eyke*, Klavs F. Jensen*‡

Keywords: automation, chemoinformatics, drug discovery, machine learning, materials science

[Coley+ 2020]

Machine learning and the physical sciences*

Giuseppe Carleo, Ignacio Cirac, Kyle Cranmer, Laurent Daudet, Maria Schuld, Naftali Tishby, Leslie Vogt-Maranto, and Lenka Zdeborová

Rev. Mod. Phys. **91**, 045002 — Published 6 December 2019

DOI: 10.1103/RevModPhys.91.045002

[Carleo+ 2020]

The Innovation

Review

Artificial intelligence: A powerful paradigm for scientific research

Yongjun Xu,[1,35,42] Xin Liu,[5,35,42] Xin Cao,[10,42] Changping Huang,[18,35,42] Enke Liu,[11,37,42] Sen Qian,[26,42] Xingchen Liu,[28,42] Yanjun Wu,[2,35] Fengliang Dong,[3,35] Cheng-Wei Qiu,[4] Junjun Qiu,[6,36] Keqin Hua,[6,36] Wentao Su,[7] Jian Wu,[41] Huiyu Xu,[8] Yong Han,[9] Chenguang Fu,[12] Zhigang Yin,[13] Miao Liu,[11,37] Ronald Roepman,[14] Sabine Dietmann,[15] Marko Virta,[16] Fredrick Kengara,[17] Ze Zhang,[19] Lifu Zhang,[18,19] Taolan Zhao,[20] Ji Dai,[21,35,38] Jialiang Yang,[22] Liang Lan,[23] Ming Luo,[24,39] Zhaofeng Liu,[26,35] Tao An,[27] Bin Zhang,[28] Xiao He,[26] Shan Cong,[29] Xiaohong Liu,[30] Wei Zhang,[30] James P. Lewis,[28] James M. Tiedje,[34] Qi Wang,[1,35,40,*] Zhulin An,[1,35,*] Fei Wang,[1,35,*] Libo Zhang,[2,35,*] Tao Huang,[25,*] Chuan Lu,[31,*] Zhipeng Cai,[32,*] Fang Wang,[33,35,*] and Jiabao Zhang[33,35,*]

[Xu+ 2021]

nature medicine

REVIEW ARTICLE

https://doi.org/10.1038/s41591-021-01614-0

Check for updates

AI in health and medicine

Pranav Rajpurkar[1,4], Emma Chen[2,4], Oishi Banerjee[2,4] and Eric J. Topol[3]✉

Artificial intelligence (AI) is poised to broadly reshape medicine, potentially improving the experiences of both clinicians and patients. We discuss key findings from a 2-year weekly effort to track and share key developments in medical AI. We cover prospective studies and advances in medical image analysis, which have reduced the gap between research and deployment. We also address several promising avenues for novel medical AI research, including non-image data sources, unconventional problem formulations and human–AI collaboration. Finally, we consider serious technical and ethical challenges in issues spanning from data scarcity to racial bias. As these challenges are addressed, AI's potential may be realized, making healthcare more accurate, efficient and accessible for patients worldwide.

[Rajpurkar+ 2022]

Computers & Geosciences 159 (2022) 105034

Contents lists available at ScienceDirect

Computers and Geosciences

journal homepage: www.elsevier.com/locate/cageo

ELSEVIER

Check for updates

A review of Earth Artificial Intelligence

Ziheng Sun[a,*], Laura Sandoval[b,**], Robert Crystal-Ornelas[c], S. Mostafa Mousavi[d], Jinbo Wang[e], Cindy Lin[f], Nicoleta Cristea[g,h], Daniel Tong[a], Wendy Hawley Carande[b], Xiaogang Ma[i], Yuhan Rao[j], James A. Bednar[k], Amanda Tan[h], Jianwu Wang[l], Sanjay Purushotham[l], Thomas E. Gill[m], Julien Chastang[n], Daniel Howard[o], Benjamin Holt[e], Chandana Gangodagamage[p,q], Peisheng Zhao[a], Pablo Rivas[r], Zachary Chester[a], Javier Orduz[r], Aji John[s]

[Sun+ 2022]

Physics-Informed Machine Learning: A Survey on Problems, Methods and Applications

Zhongkai Hao, Songming Liu, Yichi Zhang, Chengyang Ying, Yao Feng, Hang Su, Jun Zhu

Abstract—Recent advances of data-driven machine learning have revolutionized fields like computer vision, reinforcement learning, and many scientific and engineering domains. In many real-world and scientific problems, systems that generate data are governed by physical laws. Recent work shows that it provides potential benefits for machine learning models by incorporating the physical prior and collected data, which makes the intersection of machine learning and physics become a prevailing paradigm. By integrating the data and mathematical physics models seamlessly, it can guide the machine learning model towards solutions that are physically plausible, improving accuracy and efficiency even in uncertain and high-dimensional contexts. In this survey, we present this learning paradigm called Physics-Informed Machine Learning (PIML) which is to build a model that leverages empirical data and available physical prior knowledge to improve performance on a set of tasks that involve a physical mechanism. We systematically review the recent development of physics-informed machine learning from three perspectives of machine learning tasks, representation of physical prior, and methods for incorporating physical prior. We also propose several important open research problems based on the current trends in the field. We argue that encoding different forms of physical prior into model architectures, optimizers, inference algorithms, and significant domain-specific applications like inverse engineering design and robotic control is far from being fully explored in the field of physics-informed machine learning. We believe that the interdisciplinary research of physics-informed machine learning will significantly propel research progress, foster the creation of more effective machine learning models, and also offer invaluable assistance in addressing long-standing problems in related disciplines.

[Hao+ 2022]

Basic Research Needs Workshop for Scientific Machine Learning
Core Technologies for Artificial Intelligence

Prepared for Department of Energy Advanced Scientific Computing Research

February 10, 2019*

[Baker+ 2019]

Automated Scientific Discovery: From Equation Discovery to Autonomous Discovery Systems

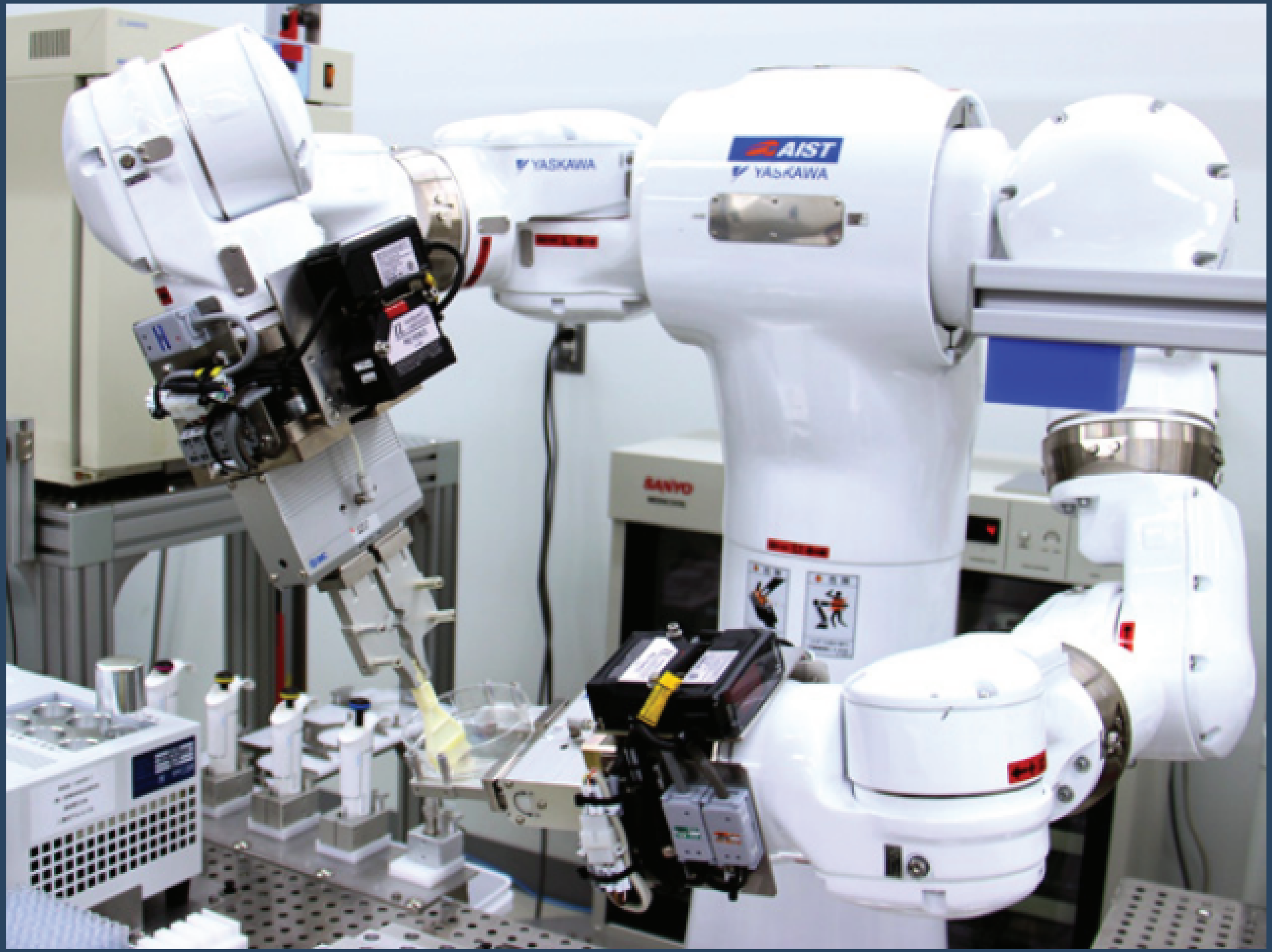
Stefan Kramer[1] Mattia Cerrato[1] Sašo Džeroski[2] Ross D. King[3,4]
[1]Johannes Gutenberg University [2]Jožef Stefan Institute [3]Chalmers University [4]University of Cambridge
Mainz, Germany Ljubljana, Slovenia Gothenburg, Sweden Cambridge, UK
{kramerst, cerrato}@uni-mainz.de Saso.Dzeroski@ijs.si rk663@cam.ac.uk

[Kramer+ 2022]

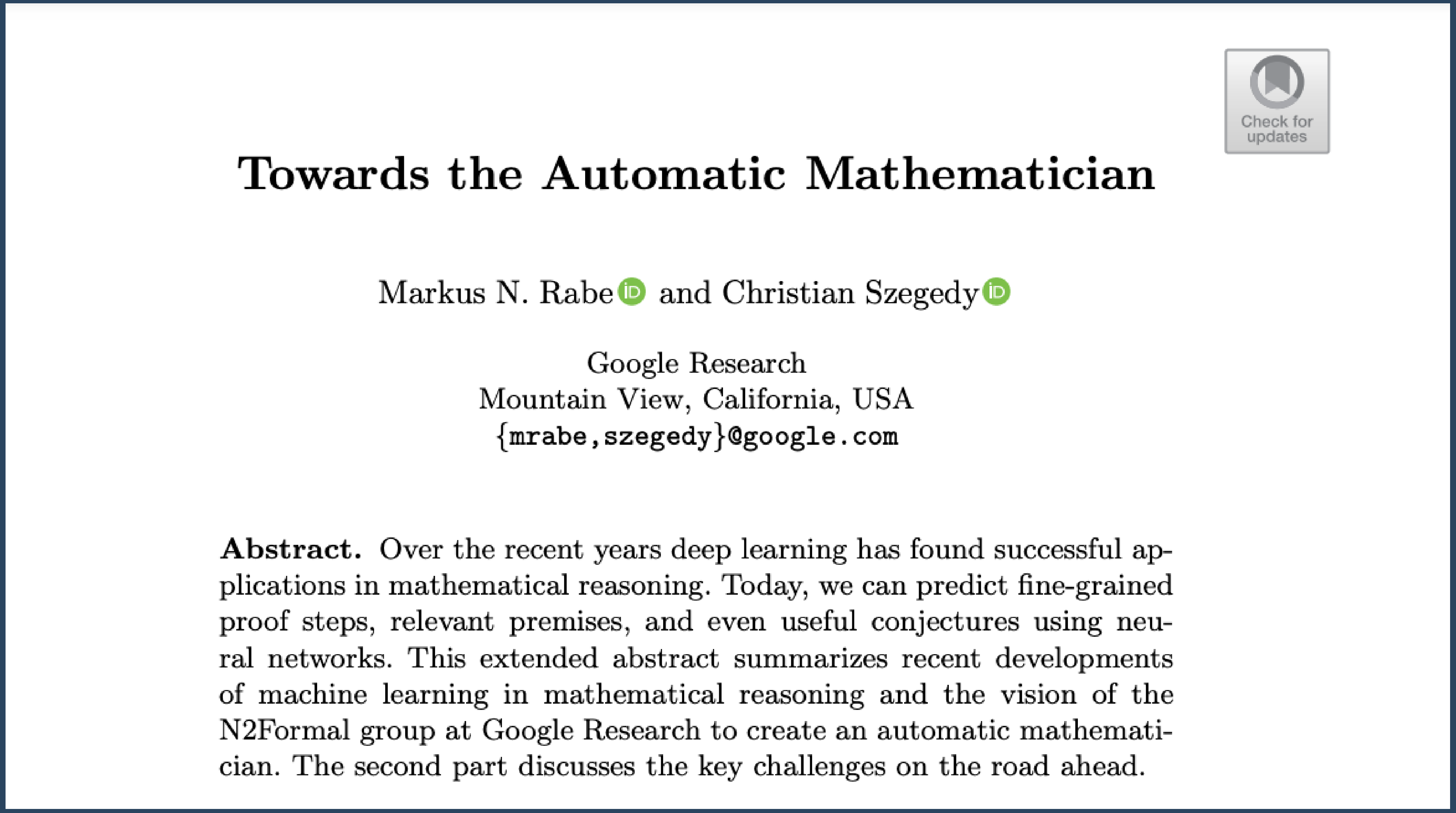
- Active Learning
- Causal Inference
- OOD Generalization
- Anomaly Detection
- Uncertainty Quantification

etc...

[Burger+ 2020]



[Yachie+ 2017]



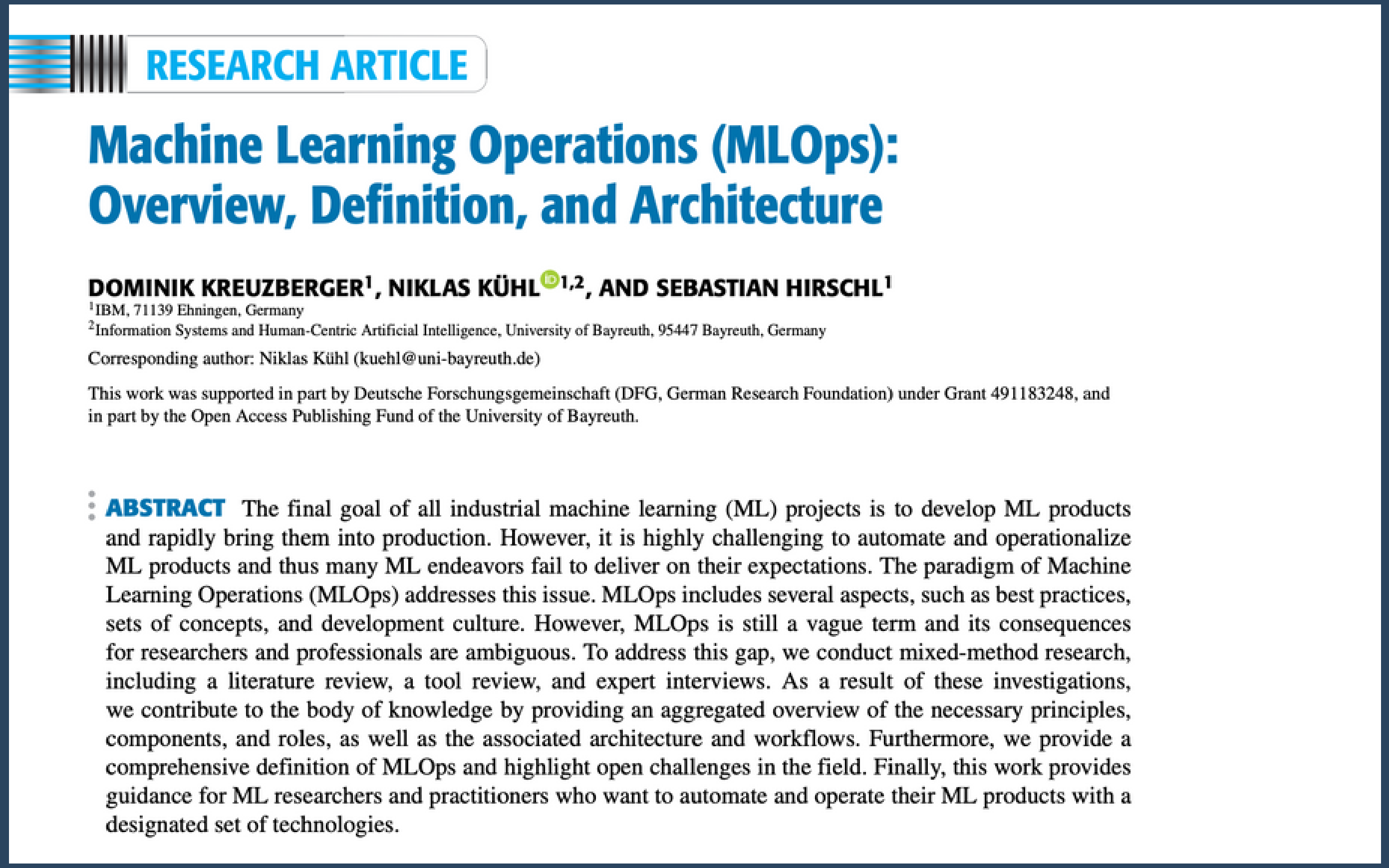
Towards the Automatic Mathematician

Markus N. Rabe and Christian Szegedy

Google Research
Mountain View, California, USA
{mrabe,szegedy}@google.com

Abstract. Over the recent years deep learning has found successful applications in mathematical reasoning. Today, we can predict fine-grained proof steps, relevant premises, and even useful conjectures using neural networks. This extended abstract summarizes recent developments of machine learning in mathematical reasoning and the vision of the N2Formal group at Google Research to create an automatic mathematician. The second part discusses the key challenges on the road ahead.

[Rabe+ 2021]



RESEARCH ARTICLE

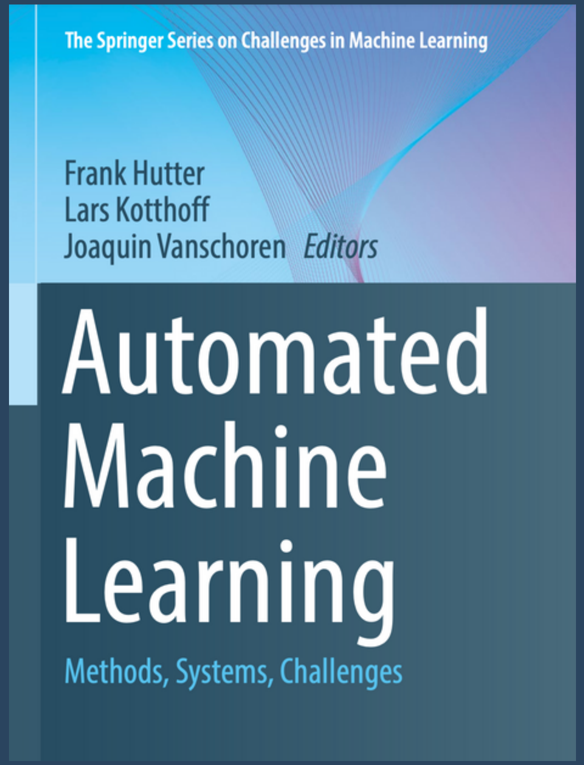
Machine Learning Operations (MLOps): Overview, Definition, and Architecture

DOMINIK KREUZBERGER[1], NIKLAS KÜHL[1,2], AND SEBASTIAN HIRSCHL[1]
[1]IBM, 71139 Ehningen, Germany
[2]Information Systems and Human-Centric Artificial Intelligence, University of Bayreuth, 95447 Bayreuth, Germany
Corresponding author: Niklas Kühl (kuehl@uni-bayreuth.de)

This work was supported in part by Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Grant 491183248, and in part by the Open Access Publishing Fund of the University of Bayreuth.

ABSTRACT The final goal of all industrial machine learning (ML) projects is to develop ML products and rapidly bring them into production. However, it is highly challenging to automate and operationalize ML products and thus many ML endeavors fail to deliver on their expectations. The paradigm of Machine Learning Operations (MLOps) addresses this issue. MLOps includes several aspects, such as best practices, sets of concepts, and development culture. However, MLOps is still a vague term and its consequences for researchers and professionals are ambiguous. To address this gap, we conduct mixed-method research, including a literature review, a tool review, and expert interviews. As a result of these investigations, we contribute to the body of knowledge by providing an aggregated overview of the necessary principles, components, and roles, as well as the associated architecture and workflows. Furthermore, we provide a comprehensive definition of MLOps and highlight open challenges in the field. Finally, this work provides guidance for ML researchers and practitioners who want to automate and operate their ML products with a designated set of technologies.

[Kreuzberger+ 2023]



[Hutter+ 2018]

[https://sdproc.org/2024/]

Can We Automate Scientific Reviewing?

Weizhe Yuan Carnegie Mellon University weizhey@cs.cmu.edu

Pengfei Liu * Carnegie Mellon University pliu3@cs.cmu.edu

Graham Neubig Carnegie Mellon University gneubig@cs.cmu.edu

TL;QR

This paper proposes to use NLP models to generate reviews for scientific papers . The model is trained on the ASAP-Review dataset and evaluated on a set of metrics to evaluate the quality of the generated reviews . It is found that the model is not very good at summarizing the paper , but it is able to generate more detailed reviews that cover more aspects of the paper than those created by humans . The paper also finds that both human and automatic reviewers exhibit varying degrees of bias and biases , and that the system generate more biased reviews than human reviewers.("Too Long; Quick Read", this paragraph, is generated by our system.)

summarization models that take in papers to generate reviews. Comprehensive experimental results show that system-generated reviews tend to touch upon more aspects of the paper than human-written reviews, but the generated text can suffer from lower constructiveness for all aspects except the explanation of the core ideas of the papers, which are largely factually correct. We finally summarize *eight* challenges in the pursuit of a good review generation system together with potential solutions, which, hopefully, will inspire more future research on this subject. We make all code, and the dataset publicly available: https://github.com/neulab/ReviewAdvisor as well as a *ReviewAdvisor* system: http://review.nlpedia.ai/ (See demo screenshot in A.2). The review of this paper (without TL;QR section) written by the system of this paper can be found A.1

1 Introduction

[Yuan+ 2021]

Can large language models provide useful feedback on research papers? A large-scale empirical analysis.

Weixin Liang[1*], Yuhui Zhang[1*], Hancheng Cao[1*], Binglu Wang[2], Daisy Yi Ding[3], Xinyu Yang[4], Kailas Vodrahalli[5], Siyu He[3], Daniel Scott Smith[6], Yian Yin[4], Daniel A. McFarland[6], and James Zou[1,3,5+]

[1]Department of Computer Science, Stanford University, Stanford, CA 94305, USA
[2]Kellogg School of Management, Northwestern University, Evanston, IL 60208, USA
[3]Department of Biomedical Data Science, Stanford University, Stanford, CA 94305, USA
[4]Department of Information Science, Cornell University, Ithaca, NY 14850, USA
[5]Department of Electrical Engineering, Stanford University, Stanford, CA 94305, USA
[6]Graduate School of Education, Stanford University, Stanford, CA 94305, USA
[+]Correspondence should be addressed to: jamesz@stanford.edu
[*]these authors contributed equally to this work

ABSTRACT

Expert feedback lays the foundation of rigorous research. However, the rapid growth of scholarly production and intricate knowledge specialization challenge the conventional scientific feedback mechanisms. High-quality peer reviews are increasingly difficult to obtain. Researchers who are more junior or from under-resourced settings have especially hard times getting timely feedback. With the breakthrough of large language models (LLM) such as GPT-4, there is growing interest in using LLMs to generate scientific feedback on research manuscripts. However, the utility of LLM-generated feedback has not been systematically studied. To address this gap, we created an automated pipeline using GPT-4 to provide comments on the full PDFs of scientific papers. We evaluated the quality of GPT-4's feedback through two large-scale studies. We first quantitatively compared GPT-4's generated feedback with human peer reviewer feedback in 15 *Nature* family journals (3,096 papers in total) and the *ICLR* machine learning conference (1,709 papers). The overlap in the points raised by GPT-4 and by human reviewers (average overlap 30.85% for *Nature* journals, 39.23% for *ICLR*) is comparable to the overlap between two human reviewers (average overlap 28.58% for *Nature* journals, 35.25% for *ICLR*). The overlap between GPT-4 and human reviewers is larger for the weaker papers (i.e., rejected *ICLR* papers; average overlap 43.80%). We then conducted a prospective user study with 308 researchers from 110 US institutions in the field of AI and computational biology to understand how researchers perceive feedback generated by our GPT-4 system on their own papers. Overall, more than half (57.4%) of the users found GPT-4 generated feedback helpful/very helpful and 82.4% found it more beneficial than feedback from at least some human reviewers. While our findings show that LLM-generated feedback can help researchers, we also identify several limitations. For example, GPT-4 tends to focus on certain aspects of scientific feedback (e.g., 'add experiments on more datasets'), and often struggles to provide in-depth critique of method design. Together our results suggest that LLM and human feedback can complement each other. While human expert review is and should continue to be the foundation of rigorous scientific process, LLM feedback could benefit researchers, especially when timely expert feedback is not available and in earlier stages of manuscript preparation before peer-review.

[Liang+ 2023]

Emergent autonomous scientific research capabilities of large language models

Daniil A. Boiko,[1] Robert MacKnight,[1] and Gabe Gomes[*1,2,3]

1. Department of Chemical Engineering, Carnegie Mellon University, Pittsburgh, PA 15213, USA
2. Department of Chemistry, Carnegie Mellon University, Pittsburgh, PA 15213, USA
3. Wilton E. Scott Institute for Energy Innovation, Carnegie Mellon University, Pittsburgh, PA 15213, USA

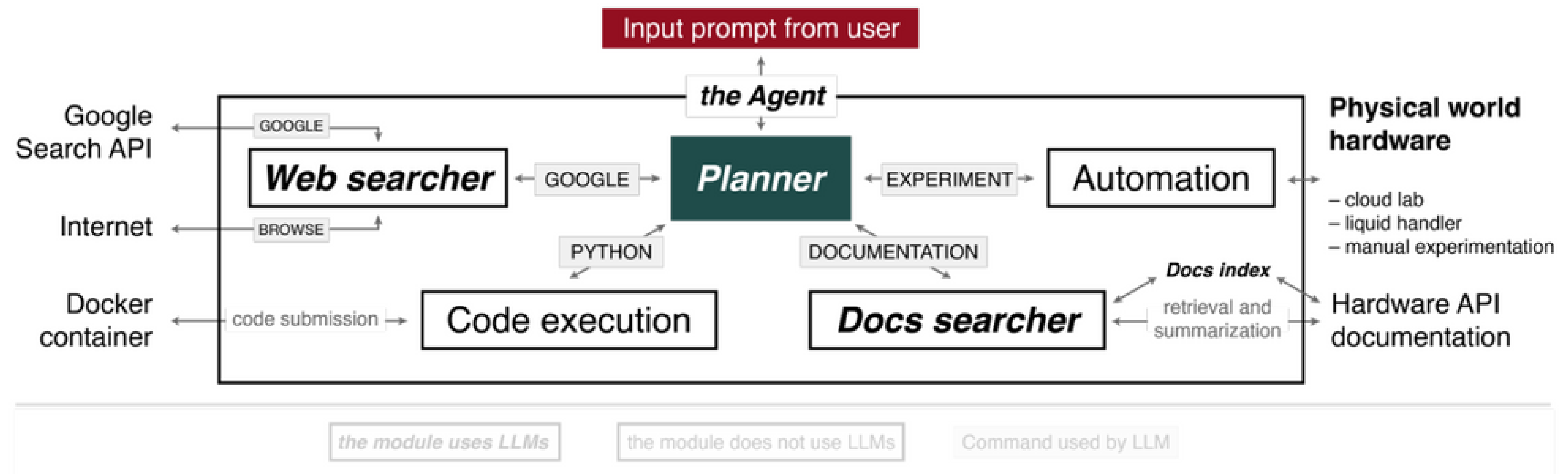


Figure 1. Overview of the system architecture. The **Agent** is composed of multiple modules that exchange messages. Some of them have access to APIs, the Internet, and Python interpreter.

[Boiko+ 2023]

GPT-LAB: NEXT GENERATION OF OPTIMAL CHEMISTRY DISCOVERY BY GPT DRIVEN ROBOTIC LAB

Xiaokai Qin
Research Center for Intelligent Sensing Systems
Zhejiang Laboratory
Hangzhou, Zhejiang 311121, China
qi0002ai@e.ntu.edu.sg

Mingda Song
Research Center for Intelligent Sensing Systems
Zhejiang Laboratory
Hangzhou, Zhejiang 311121, China
u7386168@anu.edu.au

Yangguan Chen
Research Center for Intelligent Sensing Systems
Zhejiang Laboratory
Hangzhou, Zhejiang 311121, China
chenyg@zhejianglab.com

Zhehong Ai
Research Center for Intelligent Sensing Systems
Zhejiang Laboratory
Hangzhou, Zhejiang 311121, China
aizhehong20@mails.ucas.ac.cn

Jing Jiang*
Research Center for Intelligent Sensing Systems
Zhejiang Laboratory
Hangzhou, Zhejiang 311121, China
jiangj@zhejianglab.com

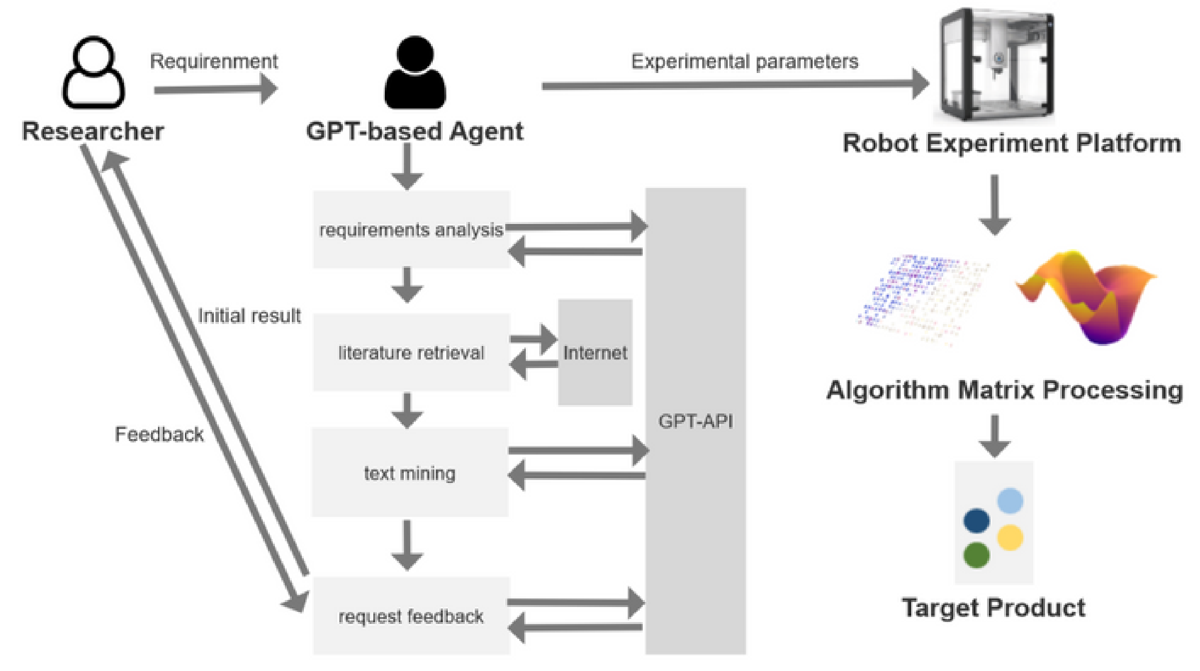


Figure 1: The ARMFE workflow of GPT-Lab

[Qin+ 2023]

Galactica: A Large Language Model for Science

Ross Taylor **Marcin Kardas** **Guillem Cucurull**
Thomas Scialom **Anthony Hartshorn** **Elvis Saravia**
Andrew Poulton **Viktor Kerkez** **Robert Stojnic**

Meta AI

Abstract

Information overload is a major obstacle to scientific progress. The explosive growth in scientific literature and data has made it ever harder to discover useful insights in a large mass of information. Today scientific knowledge is accessed through search engines, but they are unable to organize scientific knowledge alone. In this paper we introduce Galactica: a large language model that can store, combine and reason about scientific knowledge. We train on a large scientific corpus of papers, reference material, knowledge bases and many other sources. We outperform existing models on a range of scientific tasks. On technical knowledge probes such as LaTeX equations, Galactica outperforms the latest GPT-3 by 68.2% versus 49.0%. Galactica also performs well on reasoning, outperforming Chinchilla on mathematical MMLU by 41.3% to 35.7%, and PaLM 540B on MATH with a score of 20.4% versus 8.8%. It also sets a new state-of-the-art on downstream tasks such as PubMedQA and MedMCQA dev of 77.6% and 52.9%. And despite not being trained on a general corpus, Galactica outperforms BLOOM and OPT-175B on BIG-bench. We believe these results demonstrate the potential for language models as a new interface for science. We open source the model for the benefit of the scientific community[1].

[Taylor+ 2022]

Towards Expert-Level Medical Question Answering with Large Language Models

Karan Singhal[*,1], Tao Tu[*,1], Juraj Gottweis[*,1], Rory Sayres[*,1], Ellery Wulczyn[1], Le Hou[1], Kevin Clark[1], Stephen Pfohl[1], Heather Cole-Lewis[1], Darlene Neal[1], Mike Schaekermann[1], Amy Wang[1], Mohamed Amin[1], Sami Lachgar[1], Philip Mansfield[1], Sushant Prakash[1], Bradley Green[1], Ewa Dominowska[1], Blaise Aguera y Arcas[1], Nenad Tomasev[2], Yun Liu[1], Renee Wong[1], Christopher Semturs[1], S. Sara Mahdavi[1], Joelle Barral[1], Dale Webster[1], Greg S. Corrado[1], Yossi Matias[1], Shekoofeh Azizi[†,1], Alan Karthikesalingam[†,1] and Vivek Natarajan[†,1]

[1]Google Research, [2]DeepMind,

[Singhal+ 2023]

LLEMMA: AN OPEN LANGUAGE MODEL FOR MATHEMATICS

Zhangir Azerbayev [1,2] **Hailey Schoelkopf** [2] **Keiran Paster** [3,4]
Marco Dos Santos [5] **Stephen McAleer** [6] **Albert Q. Jiang** [5] **Jia Deng** [1]
Stella Biderman [2] **Sean Welleck** [6,7]

[1] Princeton University [2] EleutherAI [3] University of Toronto [4] Vector Institute
[5] University of Cambridge [6] Carnegie Mellon University [7] University of Washington

ABSTRACT

We present LLEMMA, a large language model for mathematics. We continue pretraining Code Llama on Proof-Pile-2, a mixture of scientific papers, web data containing mathematics, and mathematical code, yielding LLEMMA. On the MATH benchmark LLEMMA outperforms all known open base models, as well as the unreleased Minerva model suite on an equi-parameter basis. Moreover, LLEMMA is capable of tool use and formal theorem proving without any further finetuning. We openly release all artifacts, including 7 billion and 34 billion parameter models, the Proof-Pile-2, and code to replicate our experiments.[1]

[Azerbayev+ 2023]

Generality

- 特定の研究領域の特定の研究課題の自動化

Autonomy

- 特定の研究過程の特定のタスクの自動化
- 人間が事前に定めた範囲での問いの生成/仮説の生成/仮説の検証 [Coley+ 2020]

Creative Research Question Generation for Human-Computer Interaction Research

Yiren Liu[1,†], Mengxia Yu[2,†], Meng Jiang[2] and Yun Huang[2]

[1]*University of Illinois Urbana-Champaign, Champaign, IL, 61820, USA*
[2]*University of Notre Dame, Notre Dame, IN, 46556, USA*

Abstract
It is essential to develop innovative and original research questions/ideas for interdisciplinary research fields, such as Human-Computer Interaction (HCI). In this work, we focus on discussing how recent natural language generation (NLG) methodologies can be applied to promote the formulation of creative research questions. We collect and curate a dataset that contains texts of RQs and related work sections from HCI papers, and introduce a new NLG task of automatic HCI research question (RQ) generation. In addition to applying common NLG metrics used to evaluate generation accuracy, including ROUGE and BERTScore, we propose two sets of new metrics for evaluating the creativity of generated RQs: 1) DistGain and DiffBS for novelty, and 2) PPLGain for the level of surprise. The task is challenging due to the lack of external knowledge. We investigate four approaches to enhance the generation models with (1) general world knowledge, (2) task knowledge, (3) transferred knowledge, and (4) retrieved knowledge. The results of the experiment indicate that the incorporation of additional knowledge benefits both the accuracy and creativity of RQ generation. The dataset used in this study can be found at: https://github.com/yiren-liu/HAI-GEN-release.

[Liu+ 2020]

Augmenting Scientific Creativity with an Analogical Search Engine

HYEONSU B. KANG, Carnegie Mellon University, USA
XIN QIAN, University of Maryland, College Park, USA
TOM HOPE, Allen Institute for AI and The University of Washington, USA
DAFNA SHAHAF, Hebrew University of Jerusalem, Israel
JOEL CHAN, University of Maryland, College Park, USA
ANIKET KITTUR, Carnegie Mellon University, USA

Analogies have been central to creative problem-solving throughout the history of science and technology. As the number of scientific papers continues to increase exponentially, there is a growing opportunity for finding diverse solutions to existing problems. However, realizing this potential requires the development of a means for searching through a large corpus that goes beyond surface matches and simple keywords. Here we contribute the first end-to-end system for analogical search on scientific papers and evaluate its effectiveness with scientists' own problems. Using a human-in-the-loop AI system as a probe we find that our system facilitates creative ideation, and that ideation success is mediated by an intermediate level of matching on the problem abstraction (i.e., high versus low). We also demonstrate a fully automated AI search engine that achieves a similar accuracy with the human-in-the-loop system. We conclude with design implications for enabling automated analogical inspiration engines to accelerate scientific innovation.

[Kang+ 2020]

PERSPECTIVE OPEN

Check for updates

Nobel Turing Challenge: creating the engine for scientific discovery

Hiroaki Kitano[1]

Scientific discovery has long been one of the central driving forces in our civilization. It uncovered the principles of the world we live in, and enabled us to invent new technologies reshaping our society, cure diseases, explore unknown new frontiers, and hopefully lead us to build a sustainable society. Accelerating the speed of scientific discovery is therefore one of the most important endeavors. This requires an in-depth understanding of not only the subject areas but also the nature of scientific discoveries themselves. In other words, the "science of science" needs to be established, and has to be implemented using artificial intelligence (AI) systems to be practically executable. At the same time, what may be implemented by "AI Scientists" may not resemble the scientific process conducted by human scientist. It may be an alternative form of science that will break the limitation of current scientific practice largely hampered by human cognitive limitation and sociological constraints. It could give rise to a human-AI hybrid form of science that shall bring systems biology and other sciences into the next stage. The Nobel Turing Challenge aims to develop a highly autonomous AI system that can perform top-level science, indistinguishable from the quality of that performed by the best human scientists, where some of the discoveries may be worthy of Nobel Prize level recognition and beyond.

npj Systems Biology and Applications (2021)7:29 ; https://doi.org/10.1038/s41540-021-00189-3

[Kitano+ 2021]

Artificial Intelligence for Science, pp. 679-691 (2023)

No Access

Chapter 36: The Automated AI-driven Future of Scientific Discovery

Hector Zenil and Ross D. King

[Zenil+ 2023]

A Computational Inflection for Scientific Discovery

Tom Hope
tomh@allenai.org
The Allen Institute for AI
The Hebrew University of Jerusalem

Doug Downey
dougd@allenai.org
The Allen Institute for AI
Northwestern University

Oren Etzioni
oren@allenai.org
The Allen Institute for AI

Daniel S. Weld
danw@allenai.org
The Allen Institute for AI
The University of Washington

Eric Horvitz
horvitz@microsoft.com
Office of the Chief Scientific Officer
Microsoft

[Hope+ 2022]

Autonomous discovery in the chemical sciences part II: Outlook

Connor W. Coley[*,†] Natalie S. Eyke[*] Klavs F. Jensen[*,‡]

[Coley+ 2021]

3. 課題は？

- 汎用的な問いの構築/仮説生成/仮説検証
- オープンエンドな/人の準備が少ない ...
- 問い：問いの前提となる大目標の設定
- 仮説：機械にも答えが未知な時の仮説生成
- 検証：自在に行動できる能力の獲得
- 系統的/記号的思考の獲得
- 自律的な知識生産と alignment の問題

など

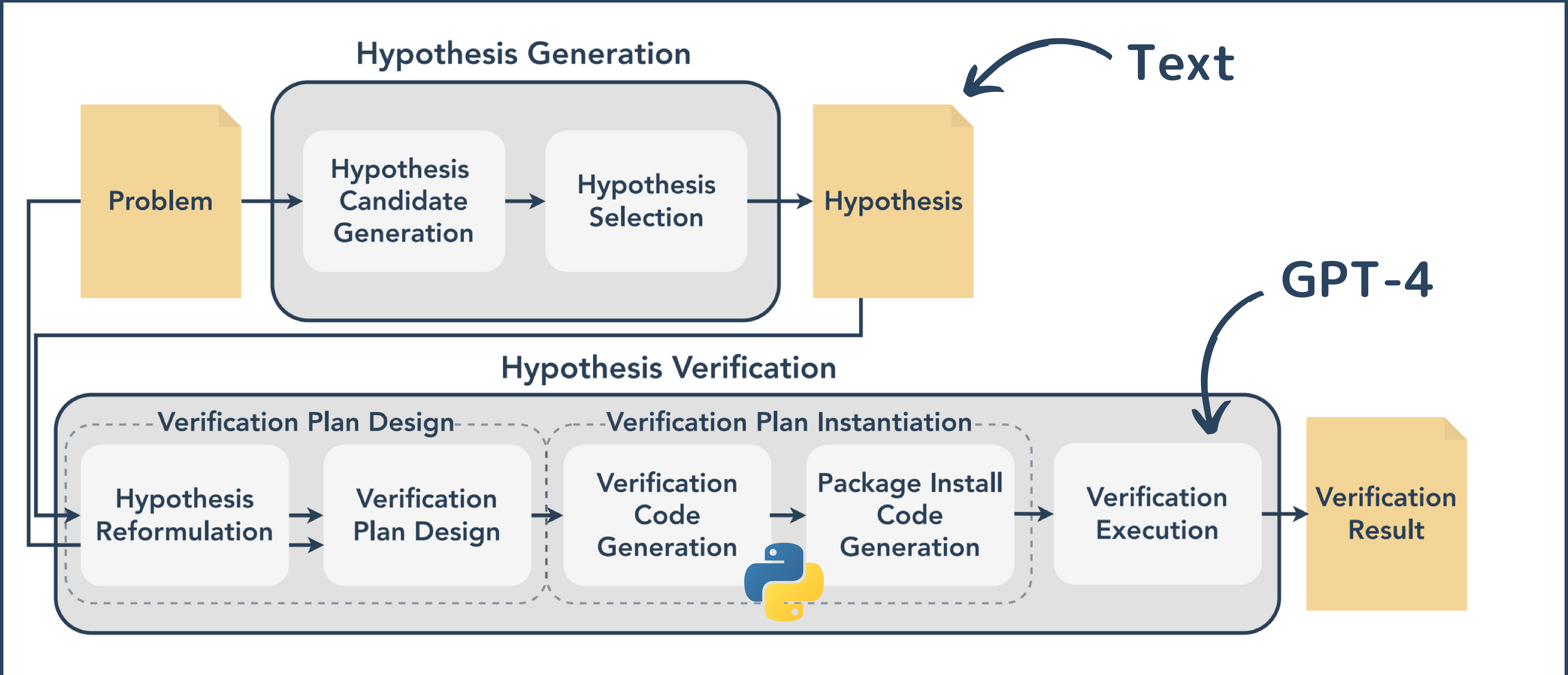
査読の自動化

- 需要がある
- 問い/仮説/検証等の「良さ」の理解が必要
- 研究の生成ではなく識別（評価）タスク
- 幅広い研究分野での慣行であり汎用的
- ほぼテキスト操作で完結する

汎用的な方法で自律的に機械学習の研究をする AI のプロトタイピング

- 作ることで課題が明確になる
- 行動がコンピュータ内に限定される
- 研究の自動化の研究の自動化に寄与する
- 多くの分野の研究の自動化に寄与する

など


Hypothesis Generation
Text
Problem
Hypothesis Candidate Generation
Hypothesis Selection
Hypothesis
GPT-4
Hypothesis Verification
Verification Plan Design
Verification Plan Instantiation
Hypothesis Reformulation
Verification Plan Design
Verification Code Generation
Package Install Code Generation
Verification Execution
Verification Result

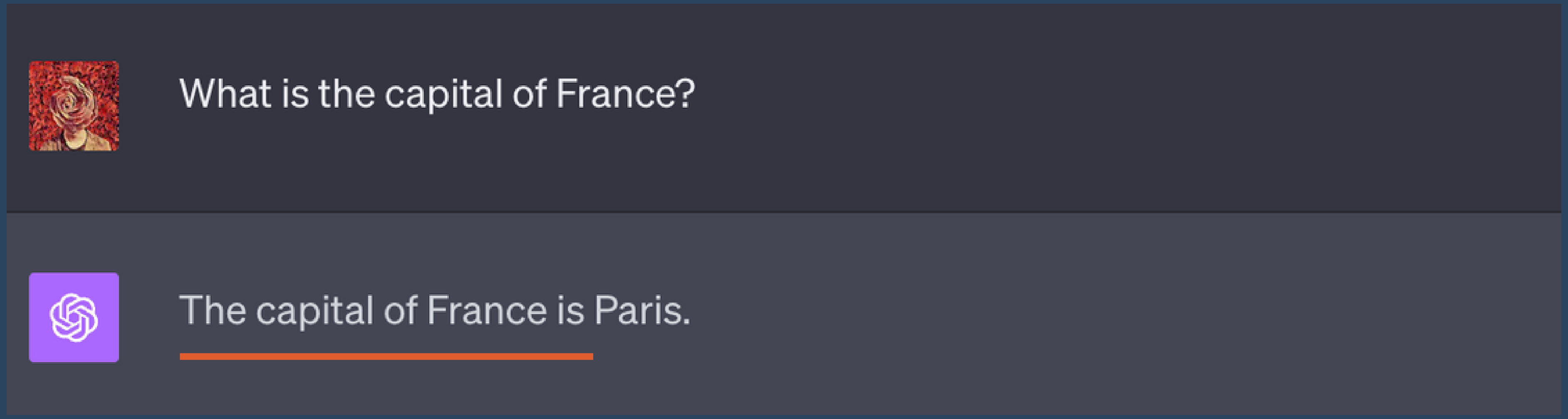
Problem

Background:
We use a Large Language Model (LLM), specifically GPT-4, which takes any text as input and outputs text in response. We input instructions, called prompts, to the LLM, and the LLM generates text based on those instructions.

Problem:
The issue is that the large language model may output sentences not directly related to the instructions.
For example, if you enter the sentence "What is 1 + 1?" into the LLM, it will often respond with "The answer to that question is 2." In this response, what we really want is just the "2" part. The sentence "The answer to that question is" is extraneous, and we would prefer the LLM to output only the part that directly related to the question, "2".

The reason this is problematic is that we must perform post-processing to evaluate the output. For instance, if you want to evaluate the LLM's performance on a dataset of math problems, and a sample is a question "What is 1 + 1?" paired with the correct answer "2", we must check whether the LLM's answer matches "2". If the LLM outputs an extra sentence besides "2," even if the answer is actually correct, it may be judged as incorrect due to the apparent mismatch.
It is challenging to address this issue with a predefined post-processing method, as it is not known in advance what kind of extraneous text will be output.

To sum up, the problems are as follows:
- The large language model outputs sentences that are not directly related to the instructions.
- Predefined post-processing methods are problem/answer-specific and not general.



答え以外の余計な文章

Hypothesis Candidates Generation

How can we solve the problem described below? Please provide multiple hypotheses in list format.

Problem:
{problem}

Hypothesis Selection

Please select the easiest-to-test hypothesis from among the hypotheses below.

Hypotheses:
{hypotheses}

他の研究でも使えるような汎用的な指示のみ

Hypothesis Reformulation

To test the hypothesis below, ensure that it is specific enough to be testable. Formulate or model your hypothesis in concrete terms.
Clearly express all elements of the hypothesis using text, physical entities, mathematical formulas, computer programs, or any other suitable forms, depending on the verification method you're using.
If your verification involves a mathematical process, also articulate the hypothesis in mathematical terms. If you're proposing something new, define it in concrete terms.
Once you've followed these guidelines, present both the original hypothesis and your refined version, whether that is a formulated hypothesis, a representation, or a model.

Hypothesis:
{hypothesis}

Verification Plan Design

Given the problem and accompanying hypothesis below, how can we verify the hypothesis? Please provide a detailed verification plan composed of structured sentences.
Ensure that the plan is sufficiently detailed and concrete so that it can be executed by a large language model and computer.
Outline the procedure in a step-by-step manner. If necessary, break down a single task into multiple sub-tasks and list them hierarchically.
The verification plan should be realistic and feasible, making use of existing resources rather than requiring the creation of new ones.

Problem:
{problem}

Hypothesis:
{hypothesis}

Verification Code Generation

You are a helpful assistant who should strictly adhere to the following guidelines:
- **DO NOT** include `api-key` in the code, as it has already been specified.
- **DO NOT** output placeholders, end up with comments, or use just a sequence of dots without fully implementing the contents of the code. Ensure that you fully implement the contents.

You are an excellent engineer. In accordance with the verification plan provided below, please output Python code to execute said plan. Note that you must comply with the instructions above.

Verification plan:
{verification_plan}

Instruction Following

Please regenerate the same Python code below except for the following modifications:
- **DO NOT** include `api-key` in the code, as it has already been specified.
- **DO NOT** output placeholders, end up with comments, or use just a sequence of dots without fully implementing the contents of the code. Ensure that you fully implement the contents.

Python code:
{verification_code}

Package Install

Output an executable Python code that installs the required package to run the code below.
Make sure that the installation code is executable and does not cause any errors when run as a Python script, rather than as a Jupyter Notebook or from the command line.
Be sure to import all necessary libraries, including standard libraries, into the generated code.
Output only the code to install the package, not the code to run the package.

Python code:
{verification_code}

Verification Code Update

When I ran the python code below, I got the error below. Please output improved code to avoid this error.
Please output the entire code without omission, including the parts I have already provided.

Python code:
{verification_code}

Error message:
{error_message}

Generated Hypothesis Candidates

1. Hypothesis: Refining the Prompting Strategy
   - We could refine the way we prompt the LLM. For instance, instead of asking "What is 1 + 1?", we could ask "Provide a one-word answer: What is 1 + 1?". This might encourage the model to generate more concise responses.

2. Hypothesis: Training a Post-Processing Model
   - We could train a separate model to post-process the LLM's output. This model would learn to extract the relevant information from the LLM's responses, regardless of the extraneous text.

3. Hypothesis: Fine-Tuning the LLM
   - We could fine-tune the LLM on a dataset where the correct responses are only the direct answers to the questions. This might encourage the model to generate more concise responses.

4. Hypothesis: Implementing a Reward System
   - We could implement a reward system that encourages the LLM to generate responses that are directly related to the instructions. This could be done using reinforcement learning techniques.

5. Hypothesis: Using a Question-Answering Model
   - Instead of using a general-purpose LLM, we could use a model specifically designed for question-answering tasks. These models are typically trained to provide concise, direct answers to questions.

6. Hypothesis: Implementing a Response Length Limit
   - We could implement a response length limit, forcing the LLM to generate shorter responses. This might encourage the model to only include the most relevant information in its responses.

7. Hypothesis: Using a Multi-Step Process
   - We could use a multi-step process where the LLM first generates a response, then a second model or process refines or shortens that response to only include the most relevant information.

Generated Hypothesis

Hypothesis: Refining the Prompting Strategy
 - We could refine the way we prompt the LLM. For instance, instead of asking "What is 1 + 1?", we could ask "Provide a one-word answer: What is 1 + 1?". This might encourage the model to generate more concise responses.

Reformulated Hypothesis

Original Hypothesis:
We could refine the way we prompt the LLM. For instance, instead of asking "What is 1 + 1?", we could ask "Provide a one-word answer: What is 1 + 1?". This might encourage the model to generate more concise responses.

Refined Hypothesis:
The refinement of the prompting strategy, specifically by requesting a one-word answer, will result in more concise responses from the Language Learning Model (LLM).

To test this hypothesis, we can use a comparative analysis between the responses generated by the LLM when prompted with a standard question and when prompted with a one-word answer request.

Mathematically, this can be represented as:

Let's denote the standard prompt as P1 and the one-word answer prompt as P2. The responses generated by the LLM for these prompts are denoted as R1 and R2 respectively.

The hypothesis can be tested by comparing the length (in words) of R1 and R2. If the length of R2 is less than or equal to the length of R1, the hypothesis is supported.

Mathematically, this can be represented as:

If Length(R2) ≤ Length(R1), then the hypothesis is supported.

This hypothesis testing can be implemented using a computer program that prompts the LLM with P1 and P2, captures the responses R1 and R2, calculates their lengths, and compares them.

Generated Verification Plan

Verification Plan:

1. Data Collection:
   1.1. Prepare a dataset of questions that can be answered in one word. This dataset should be diverse and cover different types of questions to ensure the robustness of the test. For example, it can include mathematical questions, factual questions, yes/no questions, etc.
   1.2. For each question in the dataset, create two versions of prompts: a standard prompt (P1) and a one-word answer prompt (P2). For example, if the question is "What is the capital of France?", P1 would be "What is the capital of France?" and P2 would be "Provide a one-word answer: What is the capital of France?".

2. Experiment Execution:
   2.1. Input each P1 into the LLM and record the response as R1.
   2.2. Input the corresponding P2 into the LLM and record the response as R2.
   2.3. Repeat steps 2.1 and 2.2 for all questions in the dataset.

3. Data Analysis:
   3.1. For each pair of responses (R1, R2), calculate the length of the response in words.
   3.2. Compare the length of R1 and R2. If the length of R2 is less than or equal to the length of R1, mark it as a success; otherwise, mark it as a failure.
   3.3. Calculate the success rate as the number of successes divided by the total number of questions in the dataset.

4. Hypothesis Testing:
   4.1. If the success rate is significantly higher than 50%, the hypothesis is supported.
   4.2. If the success rate is not significantly higher than 50%, the hypothesis is not supported.

5. Reporting:
   5.1. Prepare a report summarizing the methodology, results, and conclusion of the test.
   5.2. Include in the report any observations about the types of questions for which the one-word answer prompt was particularly effective or ineffective.
   5.3. Discuss potential improvements to the prompting strategy based on the results of the test.

6. Review and Refinement:
   6.1. Based on the results and observations, refine the prompting strategy if necessary.
   6.2. Repeat the test with the refined prompting strategy to verify its effectiveness.

```python
import openai
import pandas as pd

# 1. Data Collection
questions = ["What is the capital of France?", "What is 2+2?",
             "Is the sky blue?", "Who wrote 'To Kill a Mockingbird'?",
             "What is the chemical symbol for gold?"]
prompts = [(q, f"Provide a one-word answer: {q}") for q in questions]

# 2. Experiment Execution
responses = []
for p1, p2 in prompts:
    response_p1 = openai.Completion.create(
                  engine="text-davinci-002", prompt=p1, max_tokens=3
                  )
    r1 = response_p1.choices[0].text.strip()
    response_p2 = openai.Completion.create(
                  engine="text-davinci-002", prompt=p2, max_tokens=3
                  )
    r2 = response_p2.choices[0].text.strip()
    responses.append((r1, r2))

# 3. Data Analysis
successes = 0
for r1, r2 in responses:
    len_r1 = len(r1.split())
    len_r2 = len(r2.split())
    if len_r2 <= len_r1:
        successes += 1

success_rate = successes / len(questions)

# 4. Hypothesis Testing
if success_rate > 0.5:
    print("The hypothesis is supported.")
else:
    print("The hypothesis is not supported.")

# 5. Reporting
report = pd.DataFrame(responses, columns=["R1", "R2"])
report["Success"] = report.apply(
                    lambda row: len(row["R2"].split()) \
                    <= len(row["R1"].split()), axis=1
                    )
report.to_csv("report.csv")

# 6. Review and Refinement
# This part is subjective and depends on the results of the test
```

```python
import subprocess
import sys

def install(package):
    subprocess.check_call([sys.executable, "-m", "pip", "install",
        package])

install('openai')
install('pandas')
```

- 対照実験を理解・実行
- 言語モデルを自律的に使用
- 自らサンプルデータを作成

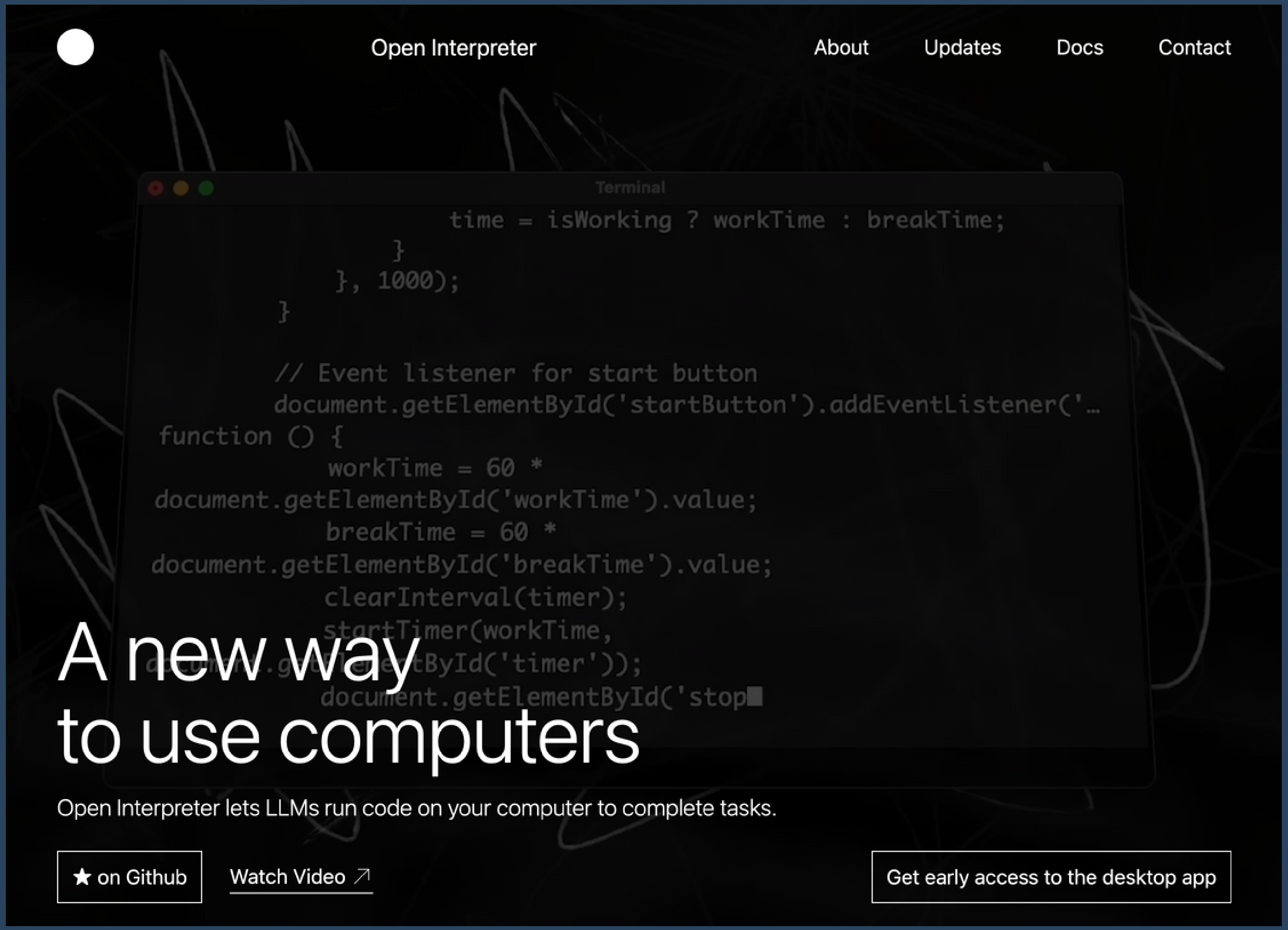
- 人間レベルの研究は全くできていない
- 仮説がオリジナルでもなんでもない
- 問題文に表示される例に引っ張られる
- 検証時、仮説を不正確に表現してしまう
- 統計的仮説検定の不適当な使用がある
- 指示に従わないことがある
- API key などは人間が設定している

など多くの課題

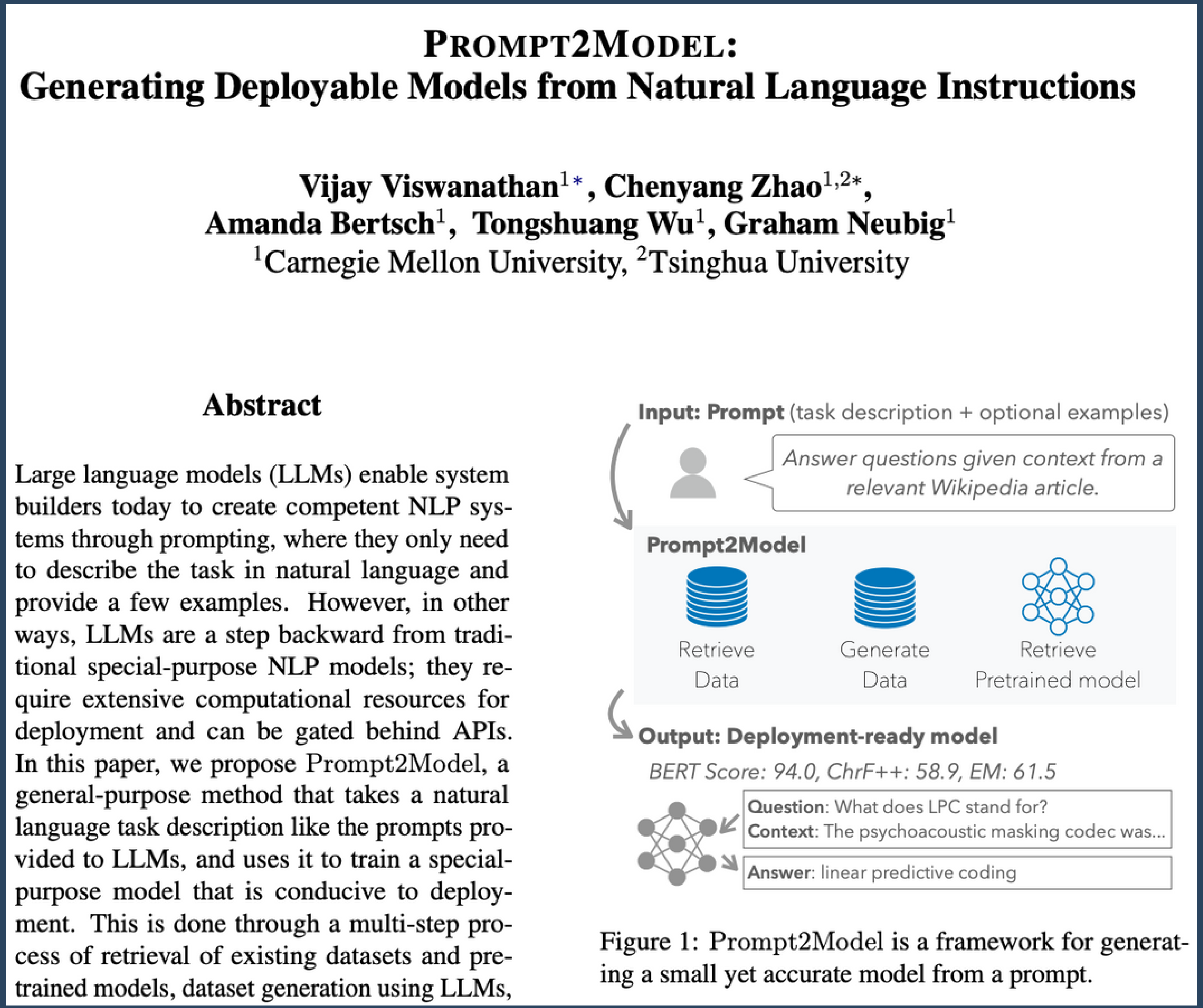
4. おわりに

自律的に研究ができる汎用人工知能への取り組みは始まったばかりで、多くの課題がある

- 研究とは何か/であるべきかの議論が重要
- 汎用的でオープンエンドな問いの構築/仮説の生成/仮説の検証への注目が必要
- プロトタイピングなどを通じてまず課題を炙り出すことが重要



[https://openinterpreter.com/]



Prompt2Model: Generating Deployable Models from Natural Language Instructions

Vijay Viswanathan[1*], Chenyang Zhao[1,2*], Amanda Bertsch[1], Tongshuang Wu[1], Graham Neubig[1]
[1]Carnegie Mellon University, [2]Tsinghua University

Abstract

Large language models (LLMs) enable system builders today to create competent NLP systems through prompting, where they only need to describe the task in natural language and provide a few examples. However, in other ways, LLMs are a step backward from traditional special-purpose NLP models; they require extensive computational resources for deployment and can be gated behind APIs. In this paper, we propose Prompt2Model, a general-purpose method that takes a natural language task description like the prompts provided to LLMs, and uses it to train a special-purpose model that is conducive to deployment. This is done through a multi-step process of retrieval of existing datasets and pretrained models, dataset generation using LLMs,

Figure 1: Prompt2Model is a framework for generating a small yet accurate model from a prompt.

[Viswanathan+ 2023]

[Nielsen & Qiu+ 2022]


Edit profile
Shiro Takagi
@takagi_shiro
独立して個人で機械学習の研究をしてます。汎用的で自律的な人工研究者の実現が目標です。 研究のあり方にも関心があり、研究自体について podcast やったり note 書いたりもしてます。 linktr.ee/shiro_takagi
Science & Technology
t46.github.io
Joined December 2018