L’IA, et nous ?

Pr. Boris Beaude
Université de Lausanne
Laboratoire d’étude des sciences et des techniques
Mercredi 1er novembre 2023

Google Cloud - Inside Google Cloud TPU Data Center - 29 août 2023

1. Un problème de définition

KASPAROV vs
DEEP BLUE
the rematch
1:57
1:58
AUDEMARS PIGUET

1:57
1:58
AP
AUDEMARS PIGUET
AP

1:57

the rematch
1:57
1:58

2022
San Francisco
2017
Phoenix East Valley
2015
Austin

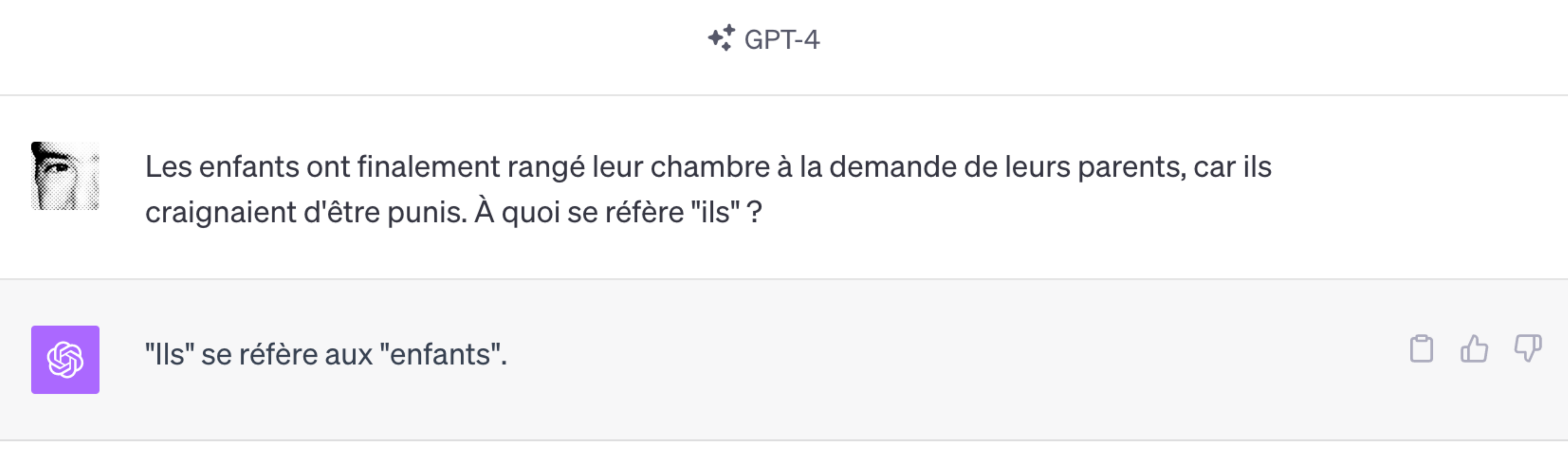
PIENSE
THINK
ΣΚΕΨΟΥ
DENKE
PENSER
$2,000
$7,000
$5,000
Ken
WATSON
BRAD
Franz Liszt 96%
Chopin 10%
Ferruccio Busoni 6%

IBM Deep Blue - 1997

DeepMind - 2016

IBM Watson - 2011

Alphabet - Waymo - 2022



Les enfants ont finalement rangé leur chambre à la demande de leurs parents, car ils craignaient d'être punis. À quoi se réfère "ils" ?

"Ils" se réfère aux "enfants".

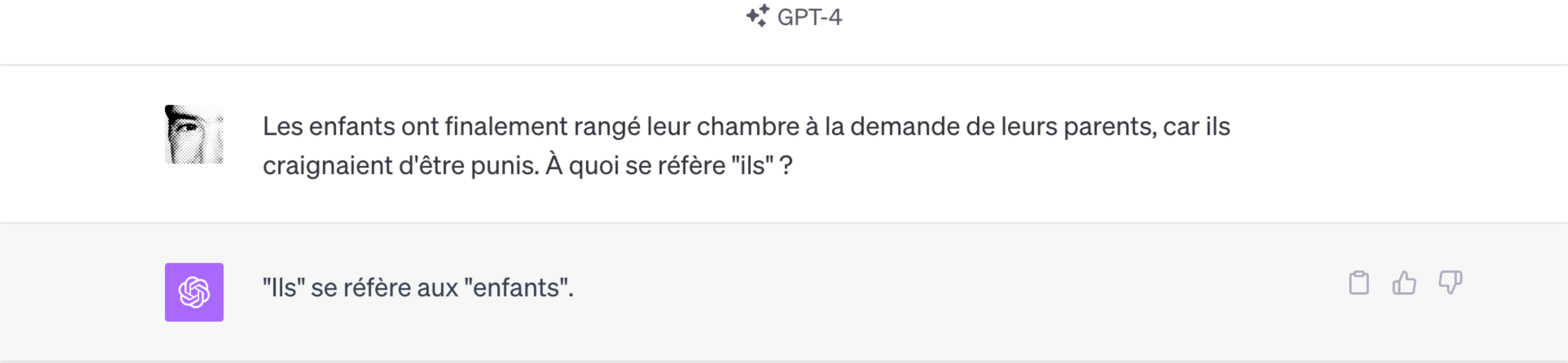
Open AI - ChatGPT 4

Midjourney - 2022

AI effect ?

Problématique de la qualification de l’intelligence

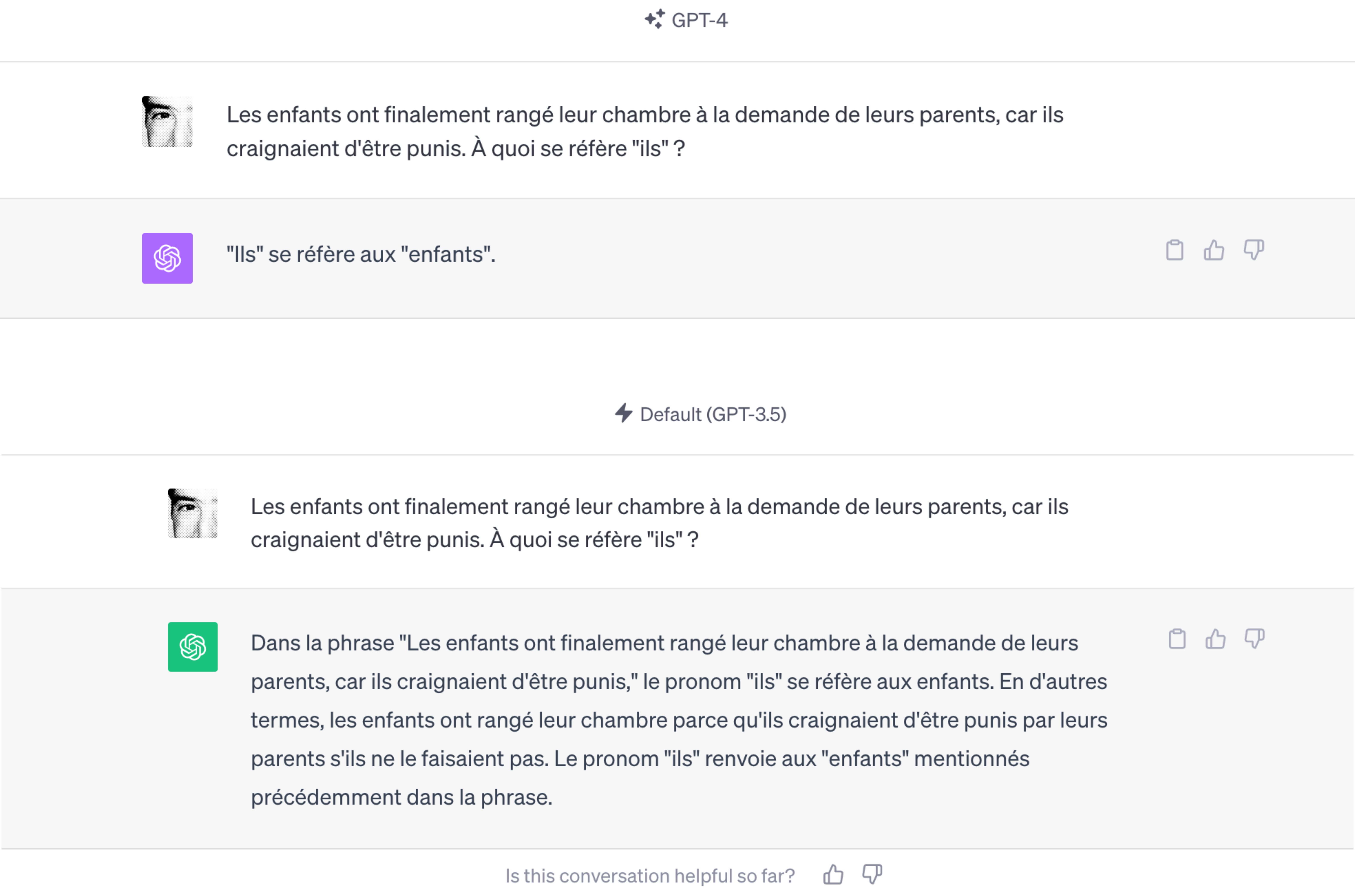
- La reconnaissance de caractères (Lecun - Tri postal - 1989)
- Les échecs (IBM Deep Blue - Kasparov - 1997)
- L’érudition - Jeopardy (IBM Watson - Jeopardy - 2011)
- Le Go (DeepMind - AlphaGo Lee - 2016)
- La compréhension d’un texte ? (Alibaba - 2018)
- La traduction ? (DeepL - 2017)
- L’art ? (Dall-E - 2021, Midjourney - 2022)
- La conduite automobile ? (Waymo - 2022)
- Le dialogue généraliste ? (ChatGPT - 2022)
- La désambiguïsation lexicale ? (GPT 3.5 - 2022)



Les enfants ont finalement rangé leur chambre à la demande de leurs parents, car ils craignaient d'être punis. À quoi se réfère "ils" ?

"Ils" se réfère aux "enfants".

Exemple de désambiguïsation lexicale



GPT-4

Les enfants ont finalement rangé leur chambre à la demande de leurs parents, car ils craignaient d'être punis. À quoi se réfère "ils" ?

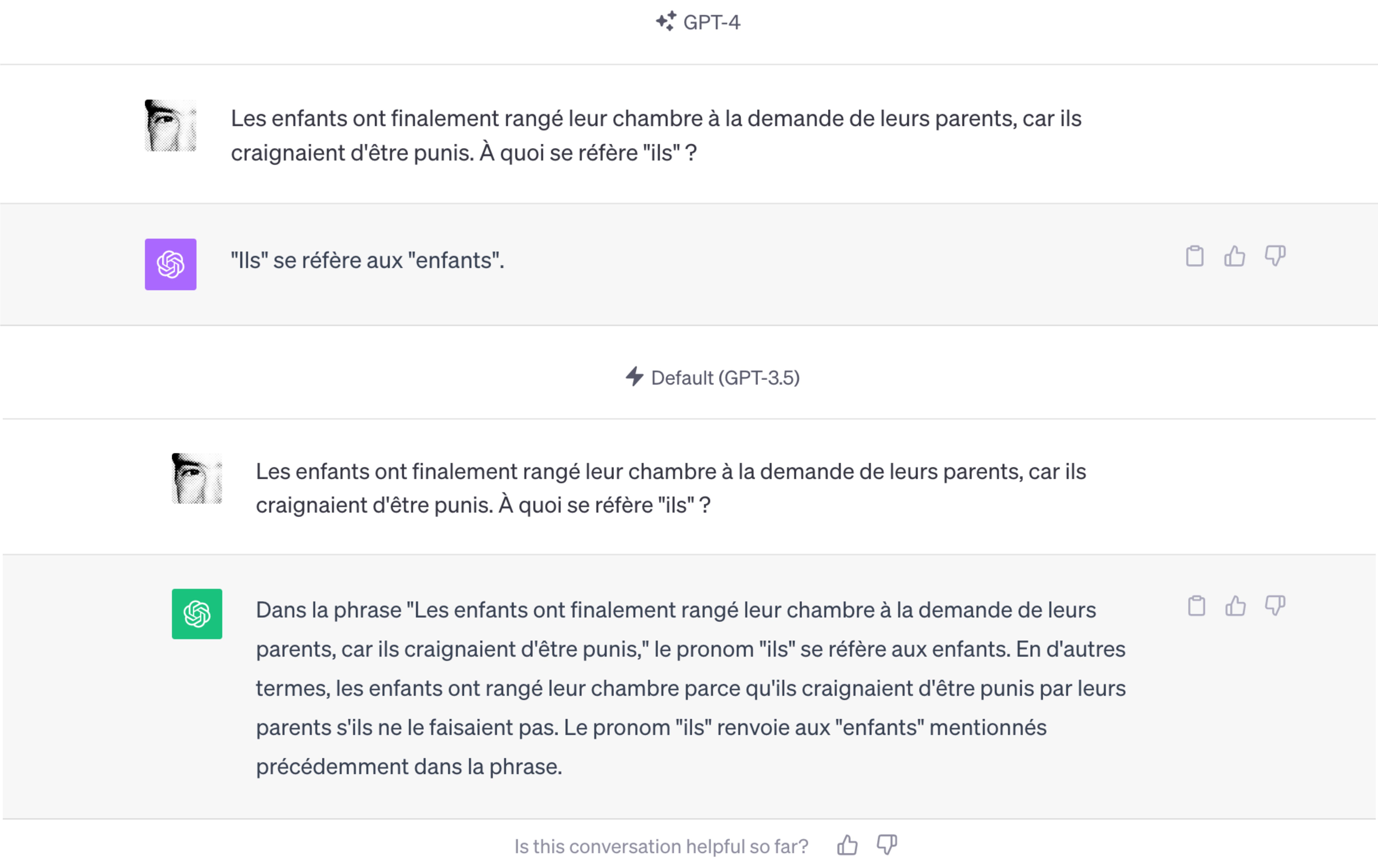
"Ils" se réfère aux "enfants".

Default (GPT-3.5)

Les enfants ont finalement rangé leur chambre à la demande de leurs parents, car ils craignaient d'être punis. À quoi se réfère "ils" ?

Dans la phrase "Les enfants ont finalement rangé leur chambre à la demande de leurs parents, car ils craignaient d'être punis," le pronom "ils" se réfère aux enfants. En d'autres termes, les enfants ont rangé leur chambre parce qu'ils craignaient d'être punis par leurs parents s'ils ne le faisaient pas. Le pronom "ils" renvoie aux "enfants" mentionnés précédemment dans la phrase.

Exemple de désambiguïsation lexicale



GPT-4

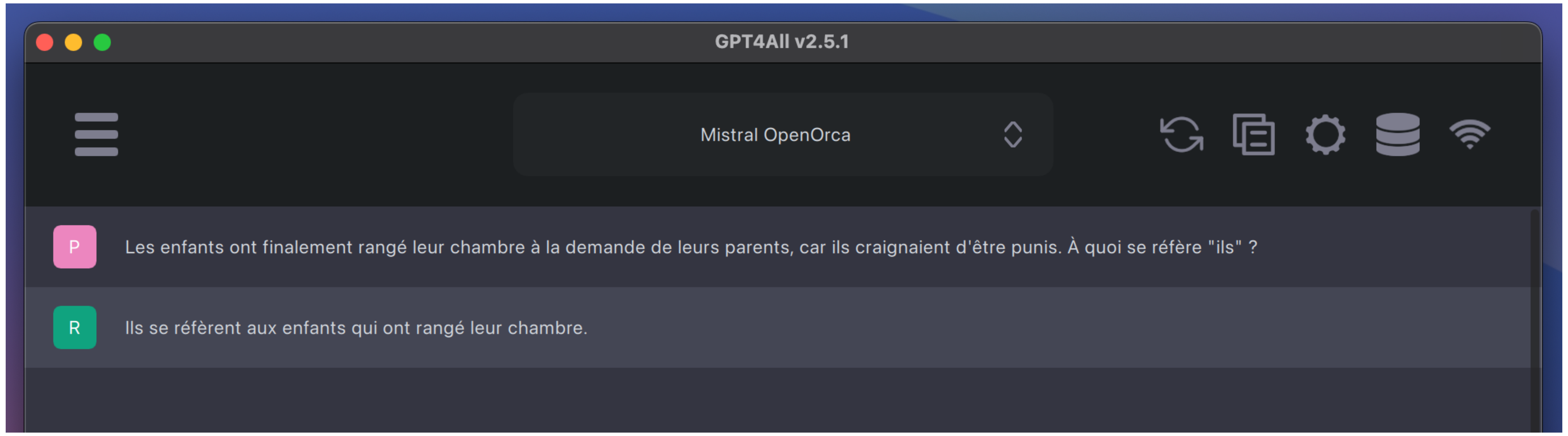
Les enfants ont finalement rangé leur chambre à la demande de leurs parents, car ils craignaient d'être punis. À quoi se réfère "ils" ?

"Ils" se réfère aux "enfants".

Default (GPT-3.5)

Les enfants ont finalement rangé leur chambre à la demande de leurs parents, car ils craignaient d'être punis. À quoi se réfère "ils" ?

Dans la phrase "Les enfants ont finalement rangé leur chambre à la demande de leurs parents, car ils craignaient d'être punis," le pronom "ils" se réfère aux enfants. En d'autres termes, les enfants ont rangé leur chambre parce qu'ils craignaient d'être punis par leurs parents s'ils ne le faisaient pas. Le pronom "ils" renvoie aux "enfants" mentionnés précédemment dans la phrase.



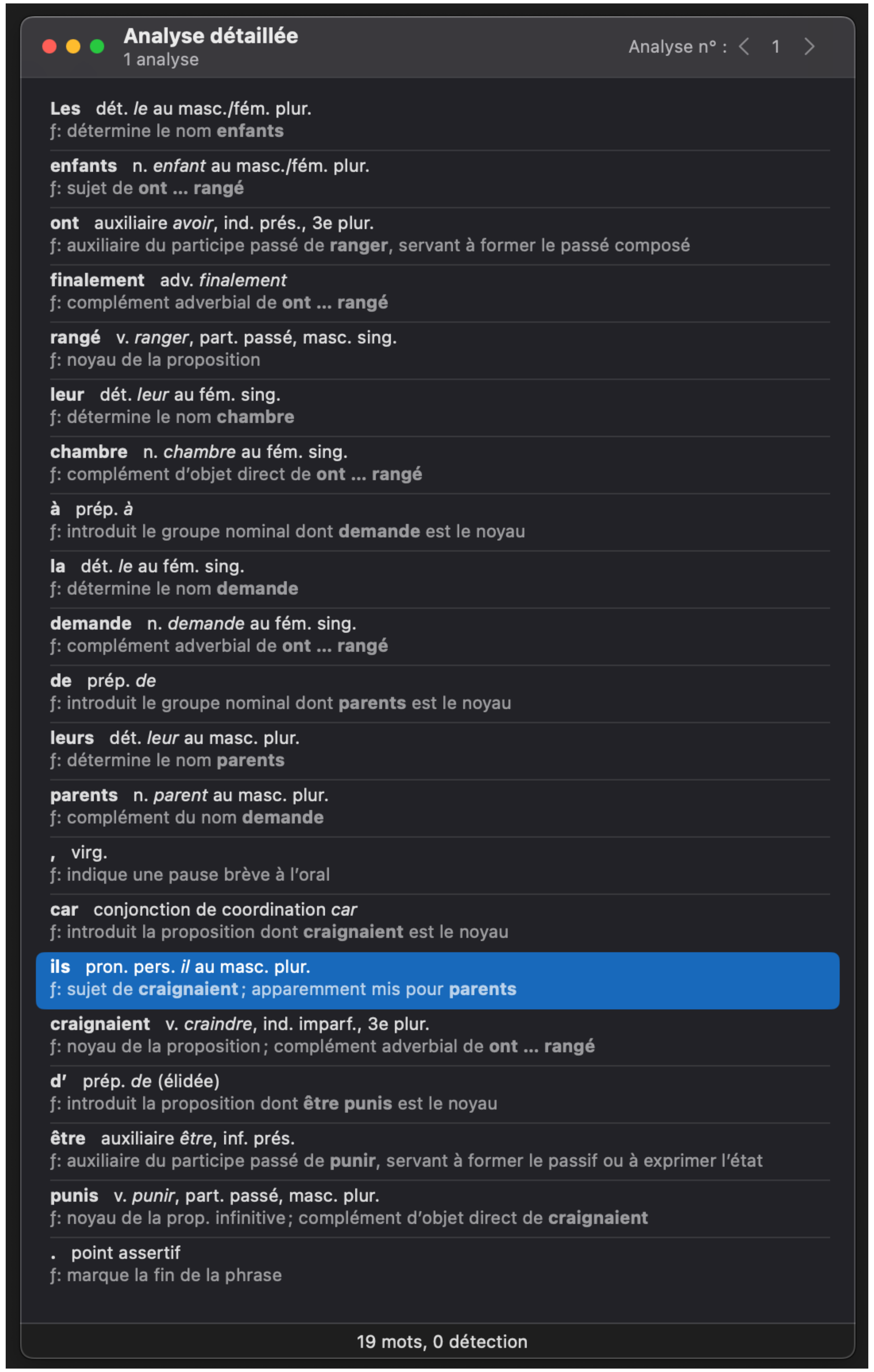
GPT4All v2.5.1

Mistral OpenOrca

Les enfants ont finalement rangé leur chambre à la demande de leurs parents, car ils craignaient d'être punis. À quoi se réfère "ils" ?

Ils se réfèrent aux enfants qui ont rangé leur chambre.

Exemple de désambiguïsation lexicale



Les dét. *le* au masc./fém. plur.
ƒ: détermine le nom **enfants**

enfants n. *enfant* au masc./fém. plur.
ƒ: sujet de **ont … rangé**

ont auxiliaire *avoir*, ind. prés., 3e plur.
ƒ: auxiliaire du participe passé de **ranger**, servant à former le passé composé

finalement adv. *finalement*
ƒ: complément adverbial de **ont … rangé**

rangé v. *ranger*, part. passé, masc. sing.
ƒ: noyau de la proposition

leur dét. *leur* au fém. sing.
ƒ: détermine le nom **chambre**

chambre n. *chambre* au fém. sing.
ƒ: complément d'objet direct de **ont … rangé**

à prép. *à*
ƒ: introduit le groupe nominal dont **demande** est le noyau

la dét. *le* au fém. sing.
ƒ: détermine le nom **demande**

demande n. *demande* au fém. sing.
ƒ: complément adverbial de **ont … rangé**

de prép. *de*
ƒ: introduit le groupe nominal dont **parents** est le noyau

leurs dét. *leur* au masc. plur.
ƒ: détermine le nom **parents**

parents n. *parent* au masc. plur.
ƒ: complément du nom **demande**

, virg.
ƒ: indique une pause brève à l'oral

car conjonction de coordination *car*
ƒ: introduit la proposition dont **craignaient** est le noyau

ils pron. pers. *il* au masc. plur.
ƒ: sujet de **craignaient** ; apparemment mis pour **parents**

craignaient v. *craindre*, ind. imparf., 3e plur.
ƒ: noyau de la proposition ; complément adverbial de **ont … rangé**

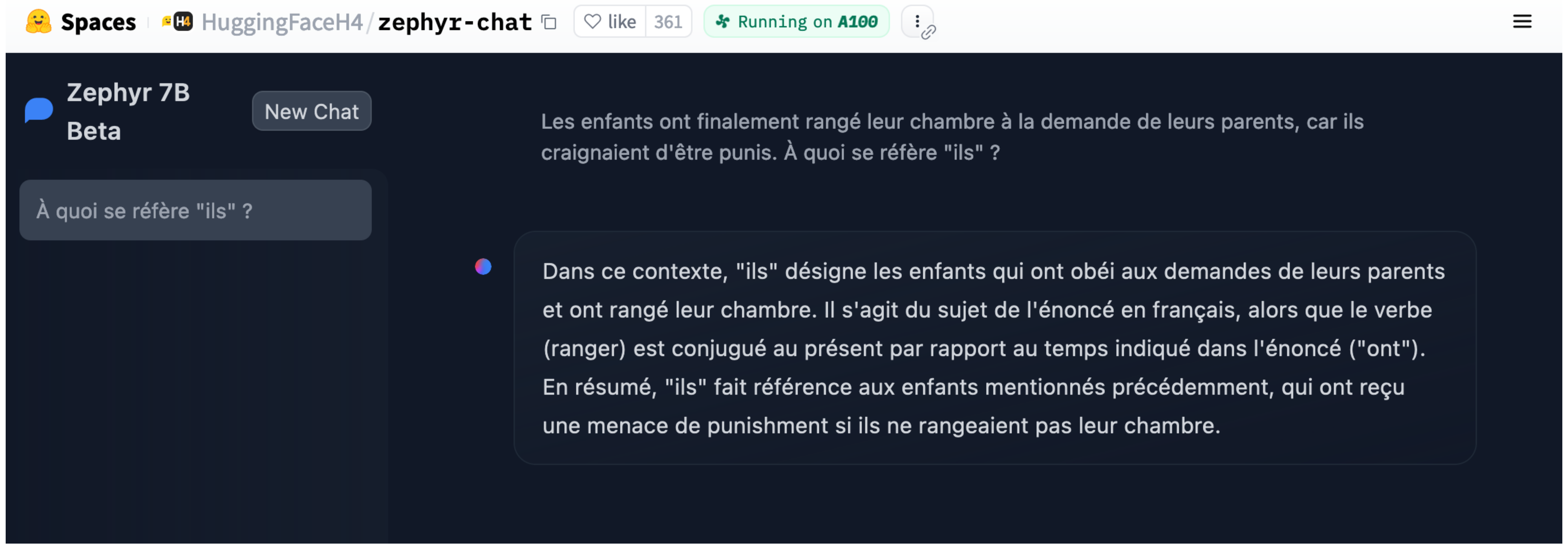
d' prép. *de* (élidée)
ƒ: introduit la proposition dont **être punis** est le noyau

être auxiliaire *être*, inf. prés.
ƒ: auxiliaire du participe passé de **punir**, servant à former le passif ou à exprimer l'état

punis v. *punir*, part. passé, masc. plur.
ƒ: noyau de la prop. infinitive ; complément d'objet direct de **craignaient**

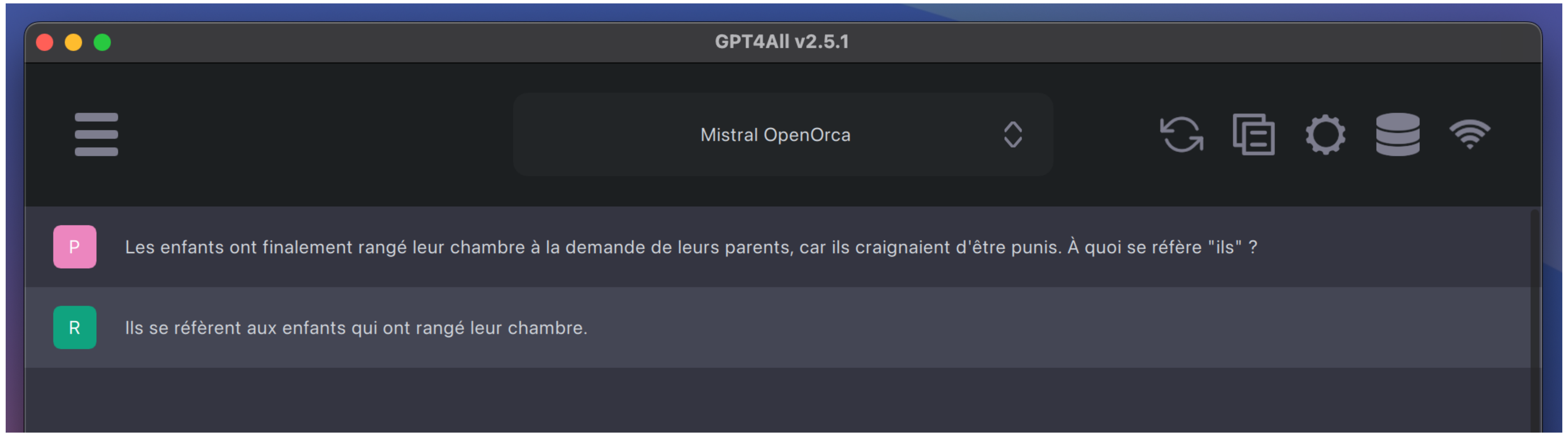
. point assertif
ƒ: marque la fin de la phrase

Druide - Antidote - 2023



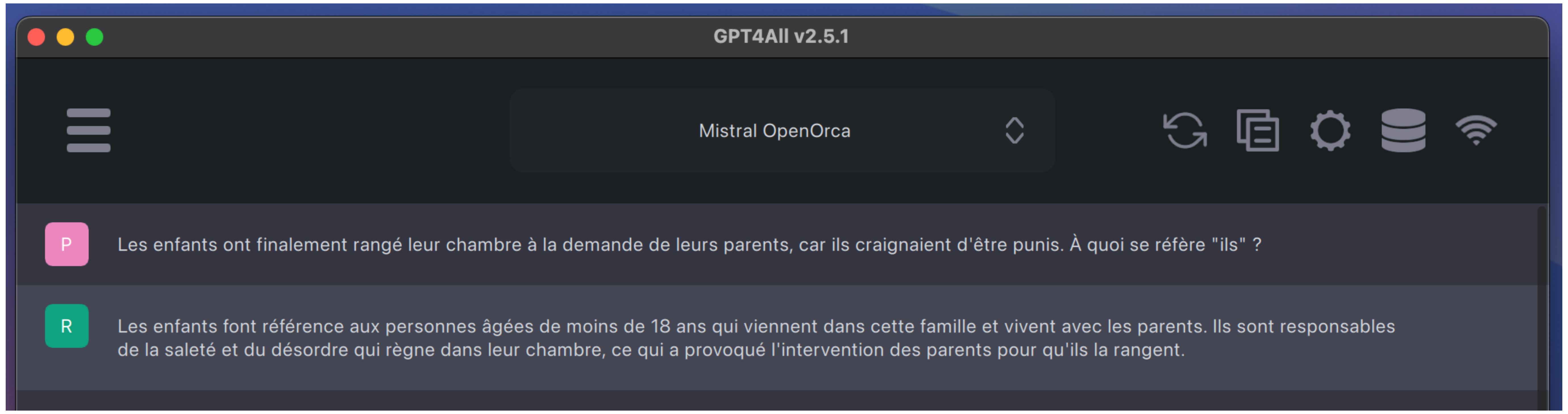
Les enfants ont finalement rangé leur chambre à la demande de leurs parents, car ils craignaient d'être punis. À quoi se réfère "ils" ?

Dans ce contexte, "ils" désigne les enfants qui ont obéi aux demandes de leurs parents et ont rangé leur chambre. Il s'agit du sujet de l'énoncé en français, alors que le verbe (ranger) est conjugué au présent par rapport au temps indiqué dans l'énoncé ("ont"). En résumé, "ils" fait référence aux enfants mentionnés précédemment, qui ont reçu une menace de punishment si ils ne rangeaient pas leur chambre.



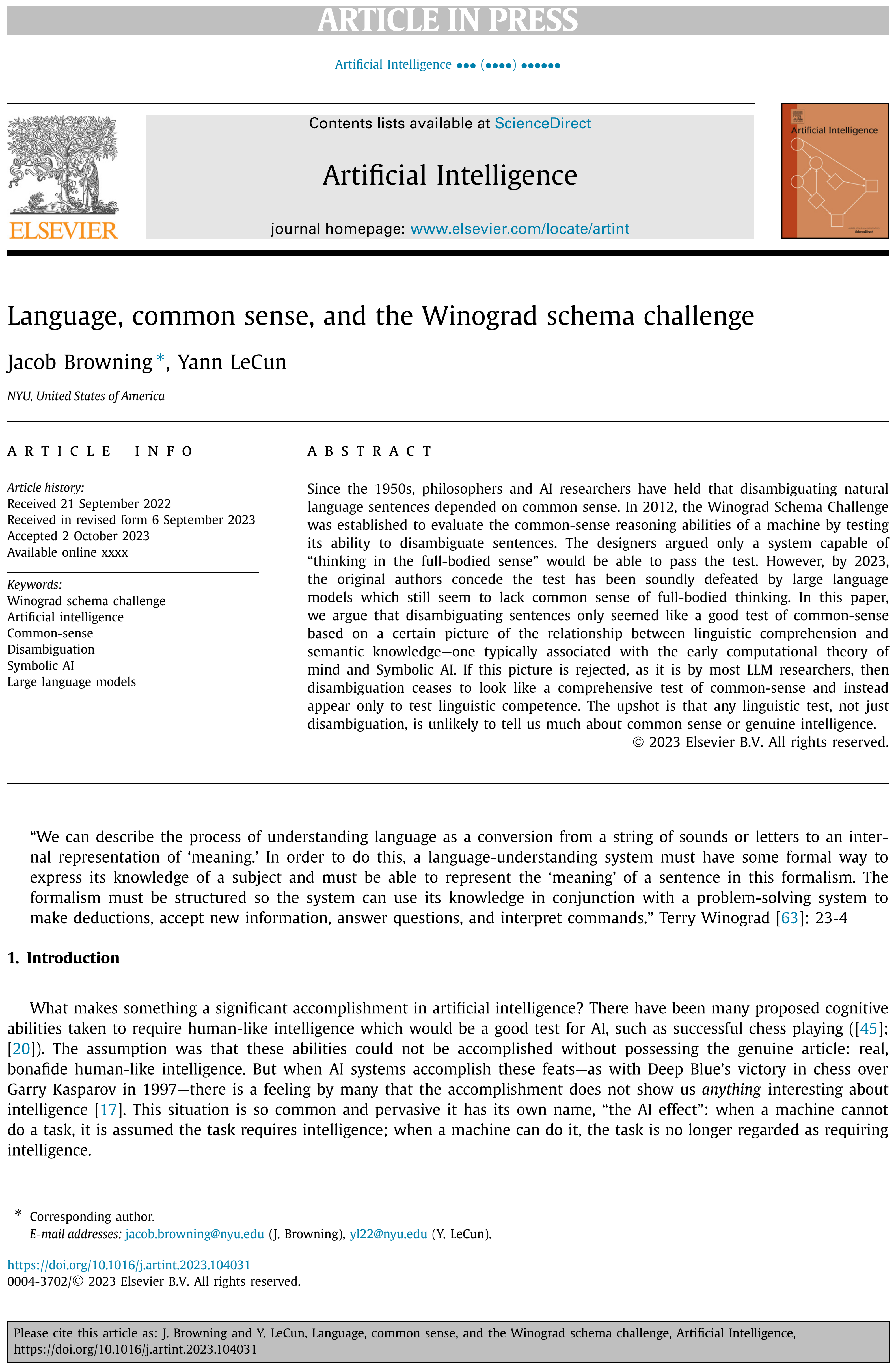
P Les enfants ont finalement rangé leur chambre à la demande de leurs parents, car ils craignaient d'être punis. À quoi se réfère "ils" ?

R Ils se réfèrent aux enfants qui ont rangé leur chambre.



P Les enfants ont finalement rangé leur chambre à la demande de leurs parents, car ils craignaient d'être punis. À quoi se réfère "ils" ?

R Les enfants font référence aux personnes âgées de moins de 18 ans qui viennent dans cette famille et vivent avec les parents. Ils sont responsables de la saleté et du désordre qui règne dans leur chambre, ce qui a provoqué l'intervention des parents pour qu'ils la rangent.



ARTICLE IN PRESS

Artificial Intelligence ••• (••••) ••••••

Contents lists available at ScienceDirect

Artificial Intelligence

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

Language, common sense, and the Winograd schema challenge

Jacob Browning*, Yann LeCun

NYU, United States of America

A R T I C L E   I N F O

Article history:
Received 21 September 2022
Received in revised form 6 September 2023
Accepted 2 October 2023
Available online xxxx

Keywords:
Winograd schema challenge
Artificial intelligence
Common-sense
Disambiguation
Symbolic AI
Large language models

A B S T R A C T

Since the 1950s, philosophers and AI researchers have held that disambiguating natural language sentences depended on common sense. In 2012, the Winograd Schema Challenge was established to evaluate the common-sense reasoning abilities of a machine by testing its ability to disambiguate sentences. The designers argued only a system capable of "thinking in the full-bodied sense" would be able to pass the test. However, by 2023, the original authors concede the test has been soundly defeated by large language models which still seem to lack common sense of full-bodied thinking. In this paper, we argue that disambiguating sentences only seemed like a good test of common-sense based on a certain picture of the relationship between linguistic comprehension and semantic knowledge—one typically associated with the early computational theory of mind and Symbolic AI. If this picture is rejected, as it is by most LLM researchers, then disambiguation ceases to look like a comprehensive test of common-sense and instead appear only to test linguistic competence. The upshot is that any linguistic test, not just disambiguation, is unlikely to tell us much about common sense or genuine intelligence.

© 2023 Elsevier B.V. All rights reserved.

> "We can describe the process of understanding language as a conversion from a string of sounds or letters to an internal representation of 'meaning.' In order to do this, a language-understanding system must have some formal way to express its knowledge of a subject and must be able to represent the 'meaning' of a sentence in this formalism. The formalism must be structured so the system can use its knowledge in conjunction with a problem-solving system to make deductions, accept new information, answer questions, and interpret commands." Terry Winograd [63]: 23-4

1. Introduction

What makes something a significant accomplishment in artificial intelligence? There have been many proposed cognitive abilities taken to require human-like intelligence which would be a good test for AI, such as successful chess playing ([45]; [20]). The assumption was that these abilities could not be accomplished without possessing the genuine article: real, bonafide human-like intelligence. But when AI systems accomplish these feats—as with Deep Blue's victory in chess over Garry Kasparov in 1997—there is a feeling by many that the accomplishment does not show us *anything* interesting about intelligence [17]. This situation is so common and pervasive it has its own name, "the AI effect": when a machine cannot do a task, it is assumed the task requires intelligence; when a machine can do it, the task is no longer regarded as requiring intelligence.

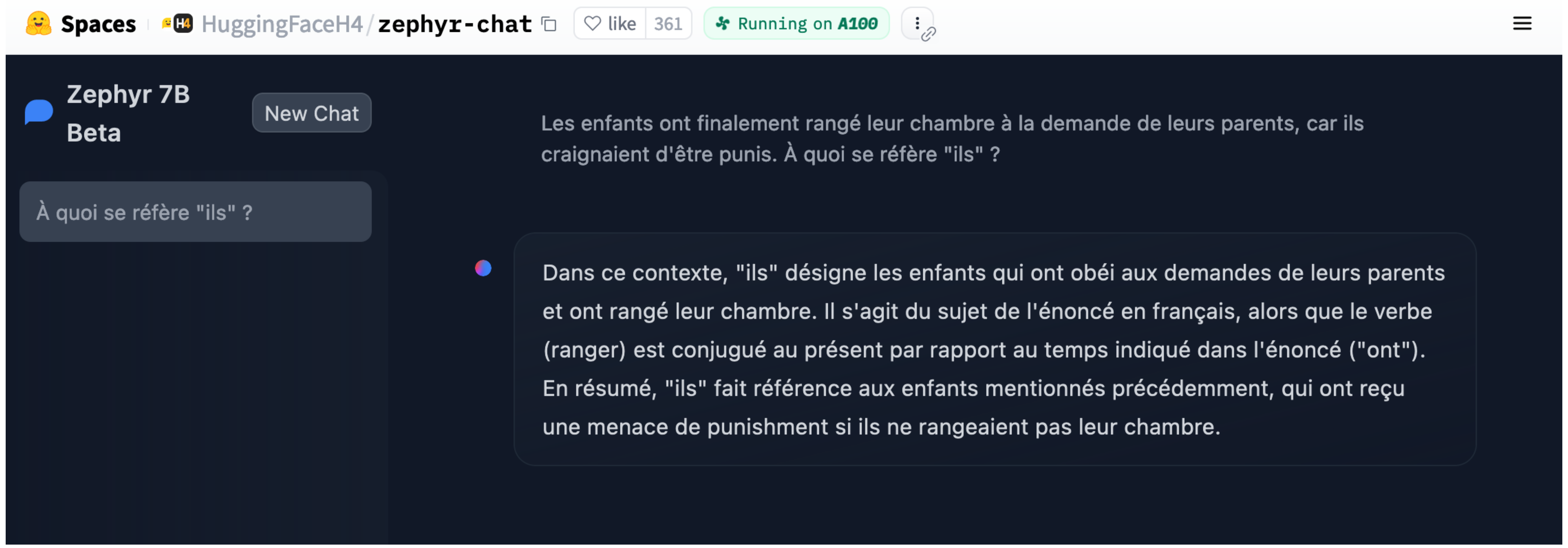
---

* Corresponding author.
E-mail addresses: jacob.browning@nyu.edu (J. Browning), yl22@nyu.edu (Y. LeCun).

https://doi.org/10.1016/j.artint.2023.104031
0004-3702/© 2023 Elsevier B.V. All rights reserved.

Please cite this article as: J. Browning and Y. LeCun, Language, common sense, and the Winograd schema challenge, Artificial Intelligence, https://doi.org/10.1016/j.artint.2023.104031

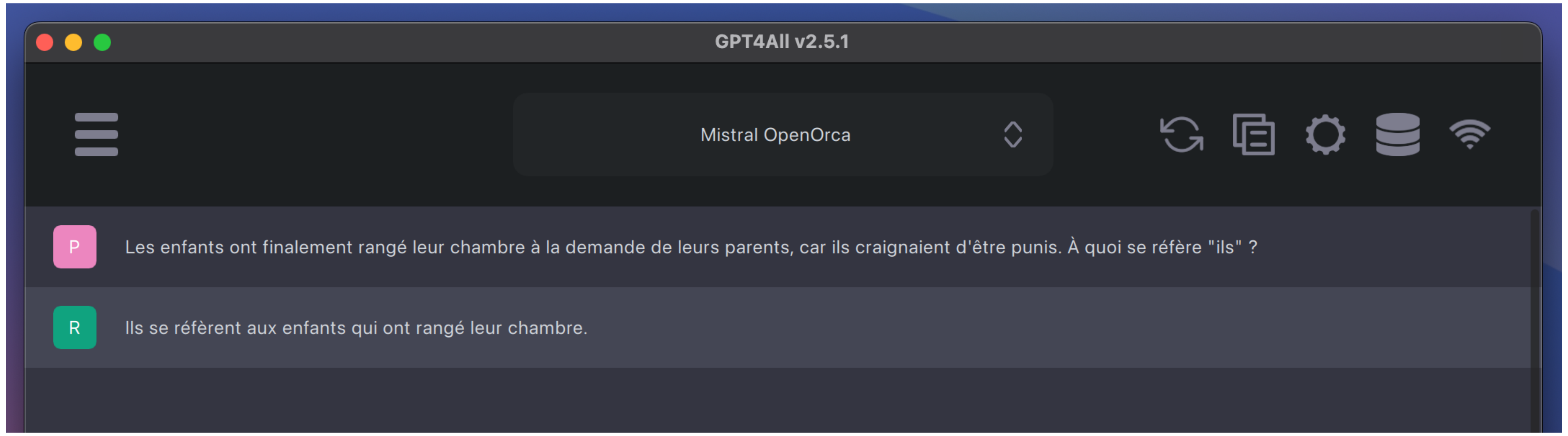
Winograd schema challenge



À quoi se réfère "ils" ?

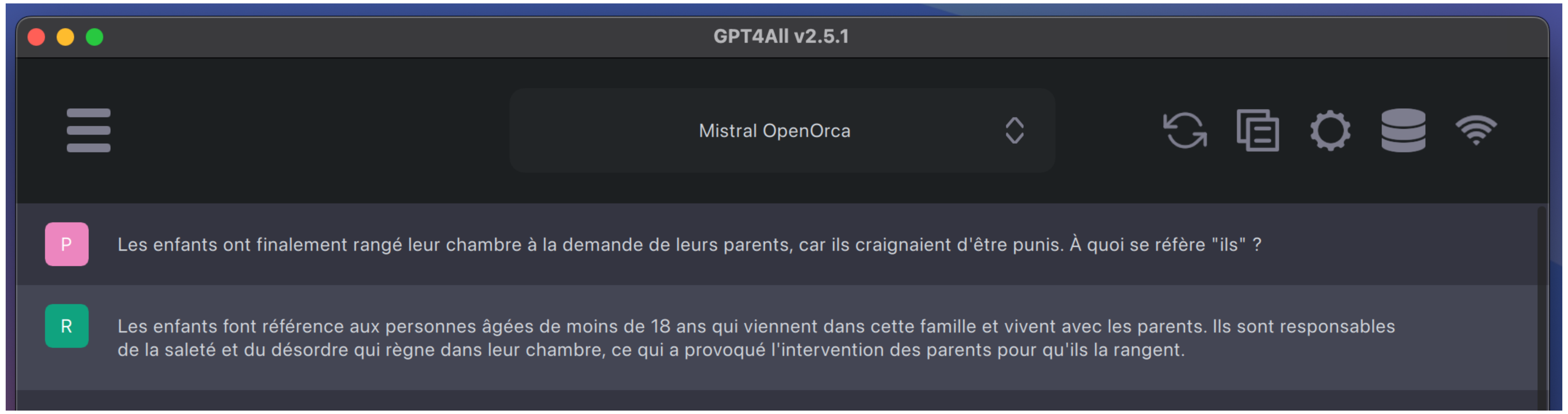
Les enfants ont finalement rangé leur chambre à la demande de leurs parents, car ils craignaient d'être punis. À quoi se réfère "ils" ?

Dans ce contexte, "ils" désigne les enfants qui ont obéi aux demandes de leurs parents et ont rangé leur chambre. Il s'agit du sujet de l'énoncé en français, alors que le verbe (ranger) est conjugué au présent par rapport au temps indiqué dans l'énoncé ("ont"). En résumé, "ils" fait référence aux enfants mentionnés précédemment, qui ont reçu une menace de punishment si ils ne rangeaient pas leur chambre.



P Les enfants ont finalement rangé leur chambre à la demande de leurs parents, car ils craignaient d'être punis. À quoi se réfère "ils" ?

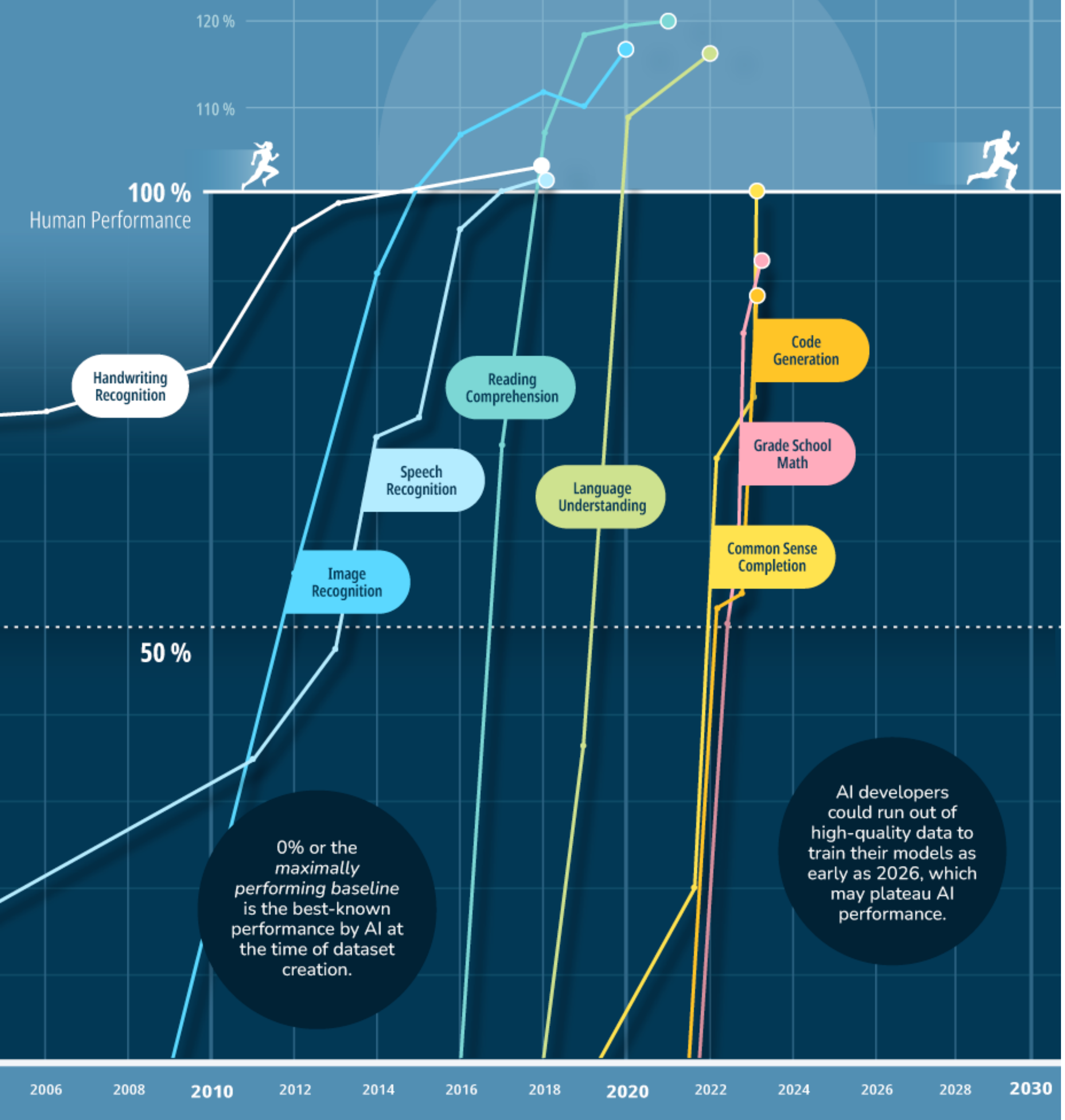
R Ils se réfèrent aux enfants qui ont rangé leur chambre.



P Les enfants ont finalement rangé leur chambre à la demande de leurs parents, car ils craignaient d'être punis. À quoi se réfère "ils" ?

R Les enfants font référence aux personnes âgées de moins de 18 ans qui viennent dans cette famille et vivent avec les parents. Ils sont responsables de la saleté et du désordre qui règne dans leur chambre, ce qui a provoqué l'intervention des parents pour qu'ils la rangent.

Jacob Browning, Yann LeCun, Language, common sense, and the Winograd schema challenge, Artificial Intelligence, 6 Octobre 2023 & Boris Beaude - 01/10/2023 - CC-BY-SA

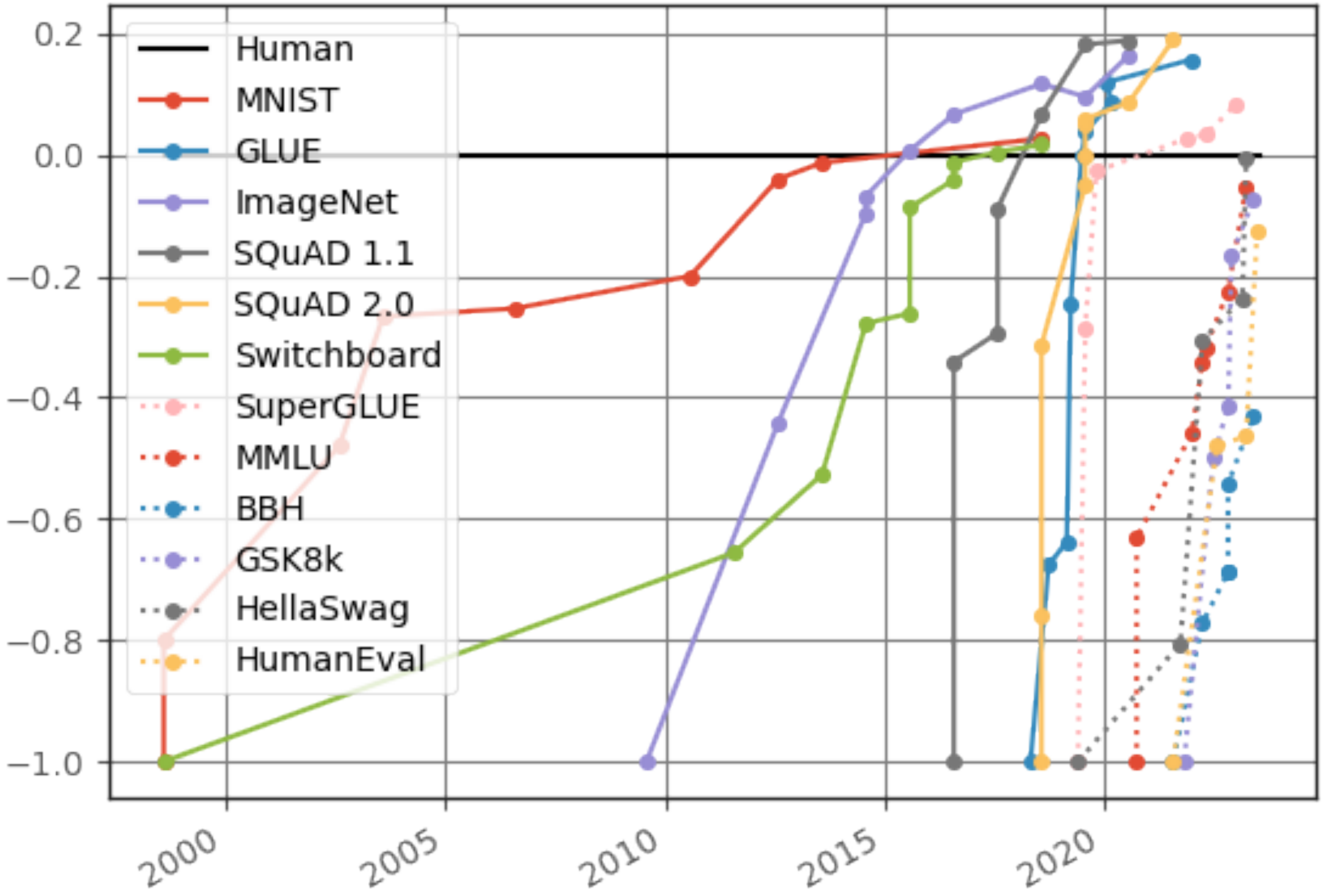


0% or the *maximally performing baseline* is the best-known performance by AI at the time of dataset creation.

AI developers could run out of high-quality data to train their models as early as 2026, which may plateau AI performance.

Handwriting Recognition	2018	MNIST
Speech Recognition	2017	Switchboard
Image Recognition	2015	ImageNet
Reading Comprehension	2018	SQuAD 1.1, 2.0
Language Understanding	2020	GLUE
Common Sense Completion	2023	HellaSwag
Grade School Math	N/A	GSK8k
Code Generation	N/A	HumanEval

Visual Capitalist via Douwe Kiela & Tristan Thrush & Kawin Ethayarajh & Amanpreet Singh, Plotting Progress in AI, Contextual AI, 31 juillet 2023



Handwriting Recognition	2018	MNIST
Speech Recognition	2017	Switchboard
Image Recognition	2015	ImageNet
Reading Comprehension	2018	SQuAD 1.1, 2.0
Language Understanding	2020	GLUE
Common Sense Completion	2023	HellaSwag
Grade School Math	N/A	GSK8k
Code Generation	N/A	HumanEval

Visual Capitalist via Douwe Kiela & Tristan Thrush & Kawin Ethayarajh & Amanpreet Singh, Plotting Progress in AI, Contextual AI, 31 juillet 2023

Que s’est-il passé ?

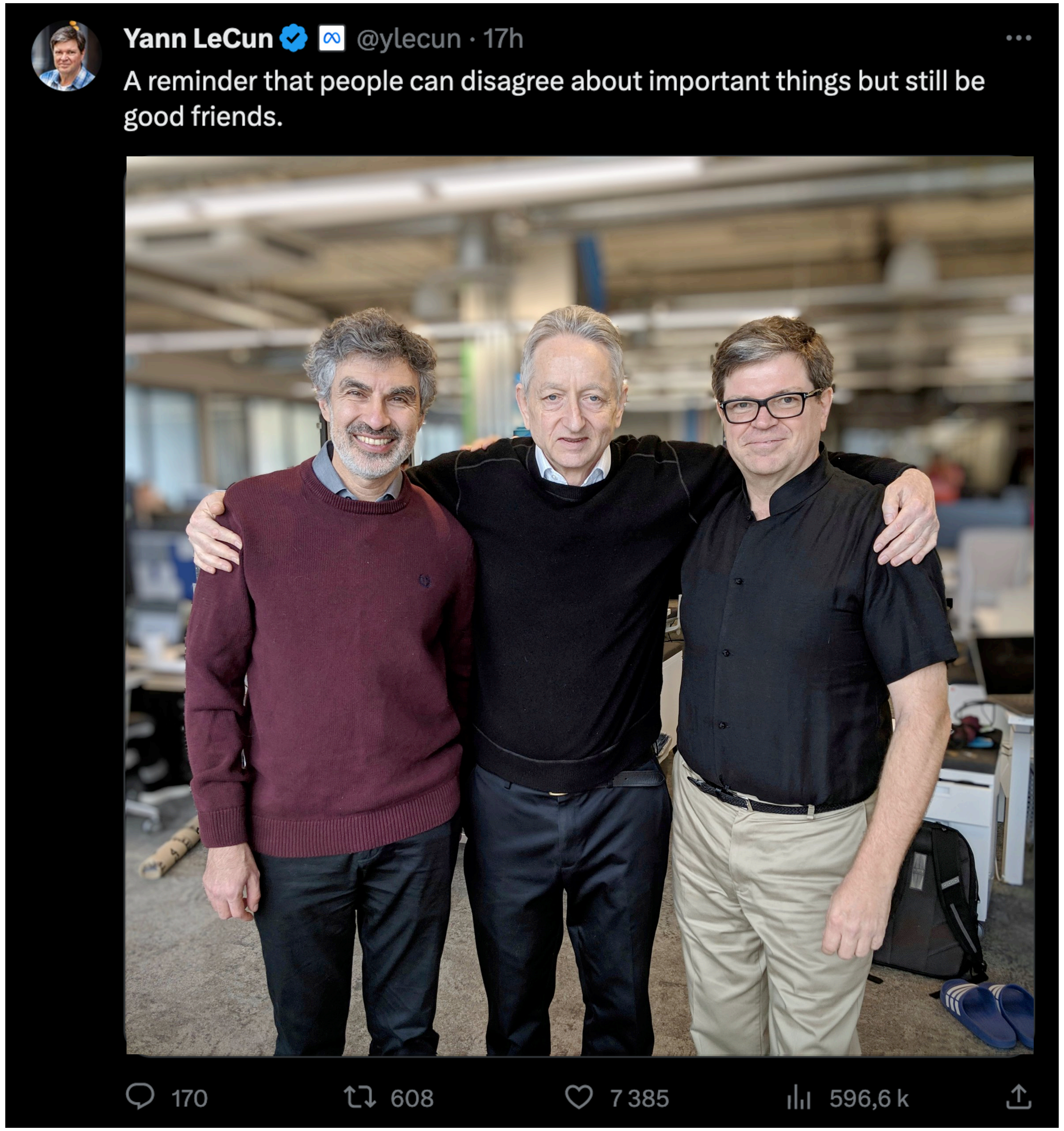
2. Une histoire mouvementée

2. Une histoire mouvementée

Yoshua Bengio
Canadien - Mila

Geoffrey Hinton
Canadien / Britanique - Google AI

Yann LeCun
Français - Meta AI



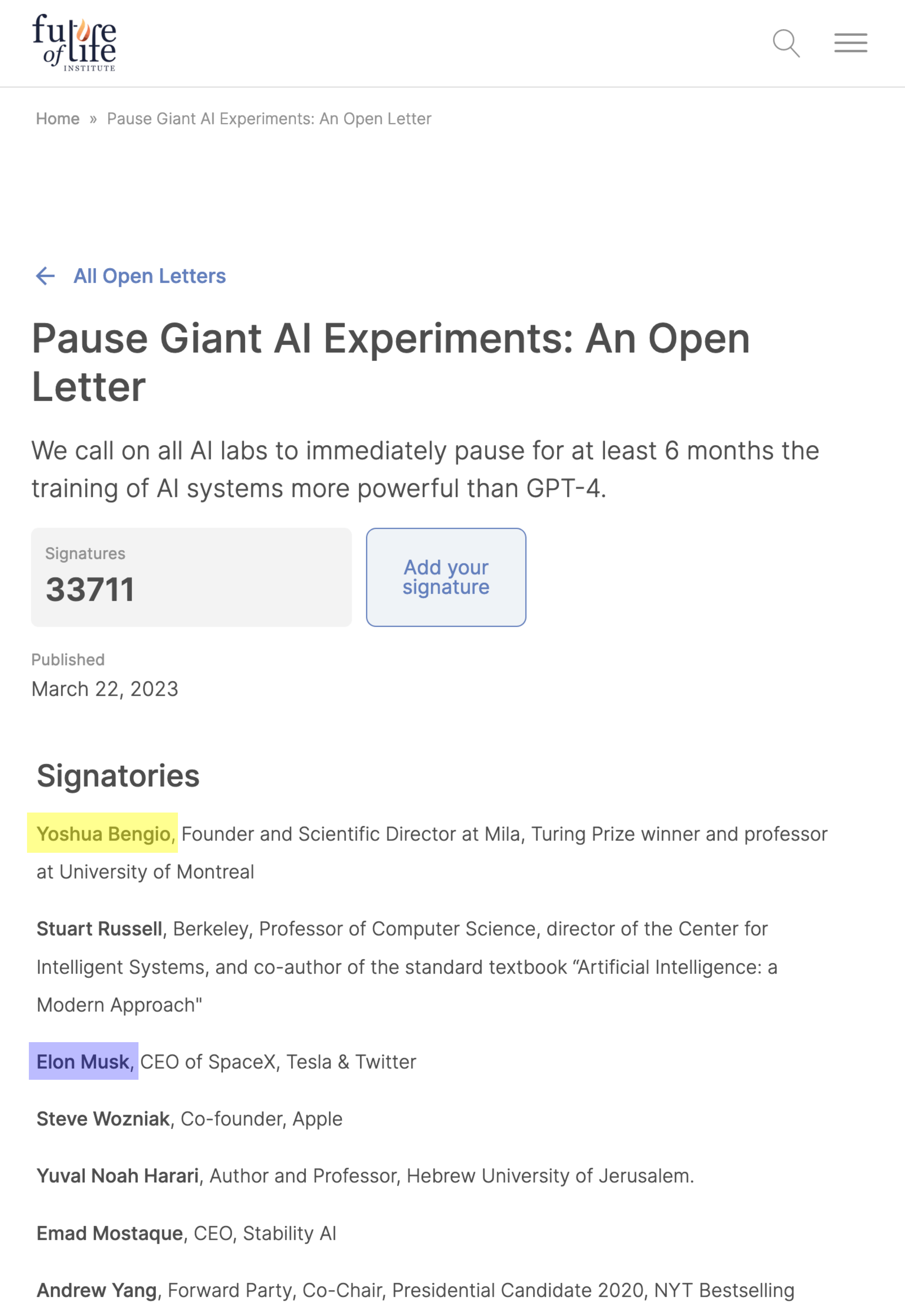
Yann LeCun @ylecun · 17h

A reminder that people can disagree about important things but still be good friends.

170 608 7 385 596,6 k

Yann LeCun - Post sur son compte Twitter - 21 octobre 2023

Geoffrey Hinton démissionne de Google AI



Home » Pause Giant AI Experiments: An Open Letter

← All Open Letters

Pause Giant AI Experiments: An Open Letter

We call on all AI labs to immediately pause for at least 6 months the training of AI systems more powerful than GPT-4.

Signatures
33711

Add your signature

Published
March 22, 2023

Signatories

Yoshua Bengio, Founder and Scientific Director at Mila, Turing Prize winner and professor at University of Montreal

Stuart Russell, Berkeley, Professor of Computer Science, director of the Center for Intelligent Systems, and co-author of the standard textbook "Artificial Intelligence: a Modern Approach"

Elon Musk, CEO of SpaceX, Tesla & Twitter

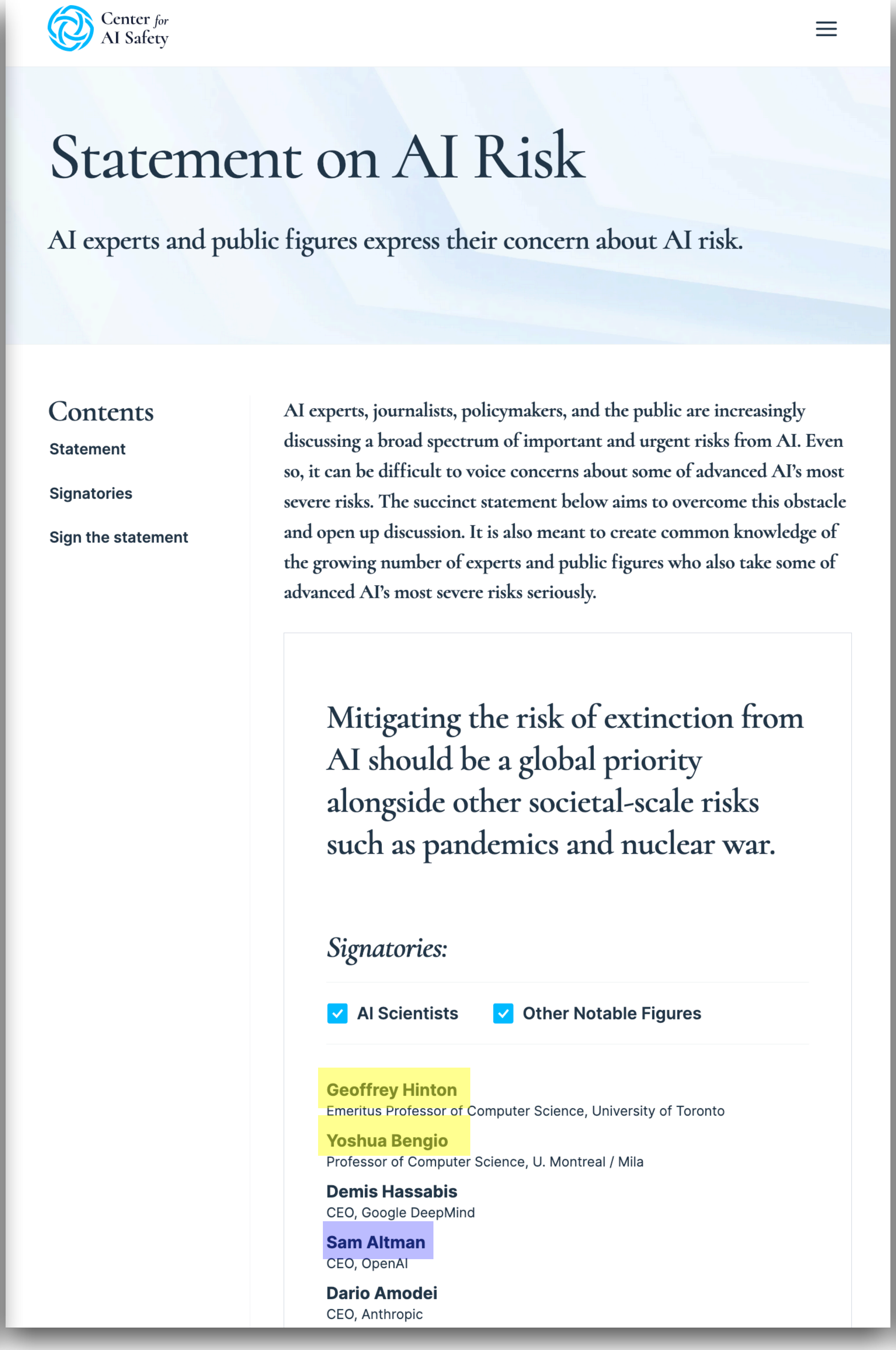
Steve Wozniak, Co-founder, Apple

Yuval Noah Harari, Author and Professor, Hebrew University of Jerusalem.

Emad Mostaque, CEO, Stability AI

Andrew Yang, Forward Party, Co-Chair, Presidential Candidate 2020, NYT Bestselling

Mars 2023



Statement on AI Risk

AI experts and public figures express their concern about AI risk.

Contents
- Statement
- Signatories
- Sign the statement

AI experts, journalists, policymakers, and the public are increasingly discussing a broad spectrum of important and urgent risks from AI. Even so, it can be difficult to voice concerns about some of advanced AI's most severe risks. The succinct statement below aims to overcome this obstacle and open up discussion. It is also meant to create common knowledge of the growing number of experts and public figures who also take some of advanced AI's most severe risks seriously.

> Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war.

Signatories:

☑ AI Scientists ☑ Other Notable Figures

Geoffrey Hinton
Emeritus Professor of Computer Science, University of Toronto

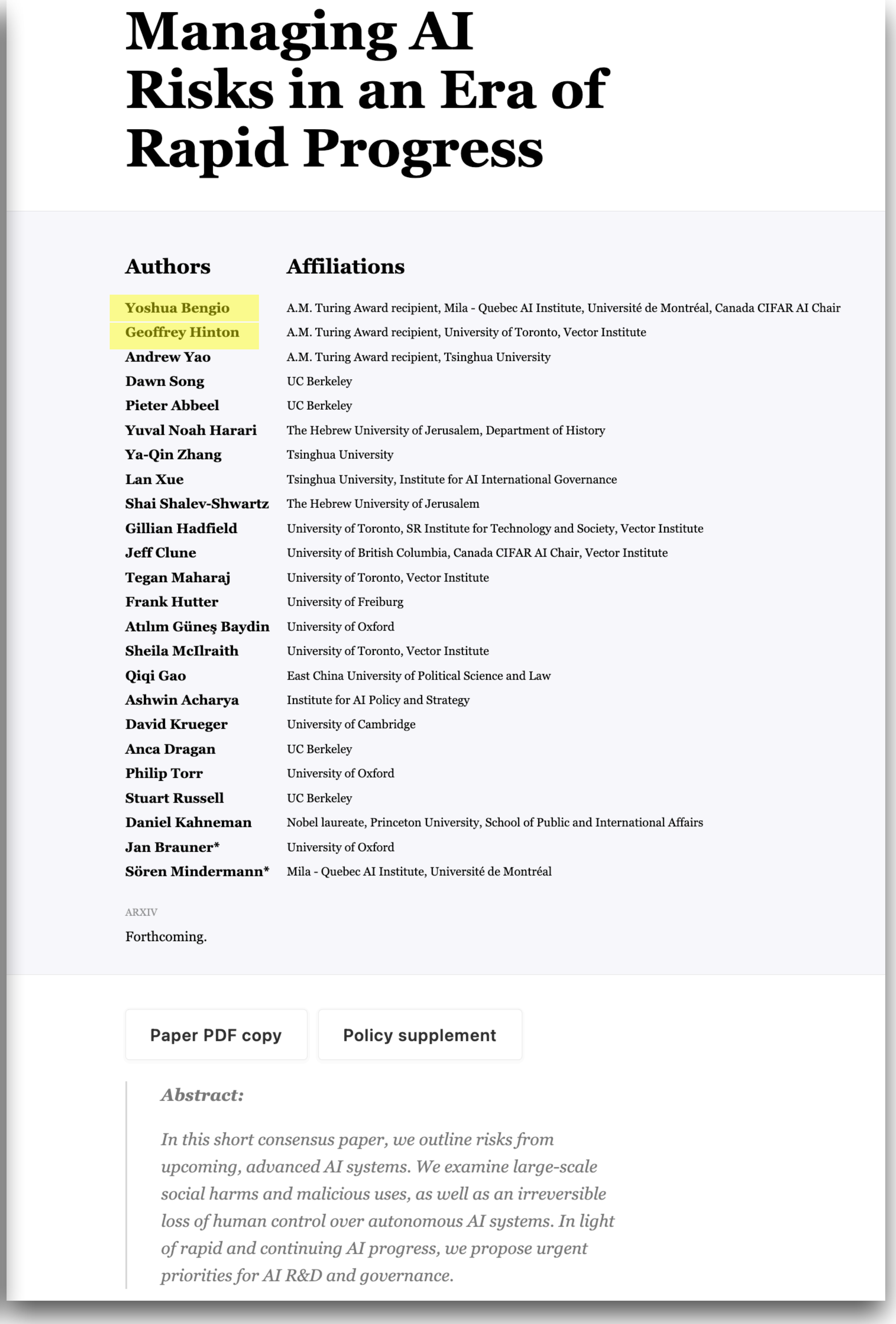
Yoshua Bengio
Professor of Computer Science, U. Montreal / Mila

Demis Hassabis
CEO, Google DeepMind

Sam Altman
CEO, OpenAI

Dario Amodei
CEO, Anthropic

Mai, 2023



Managing AI Risks in an Era of Rapid Progress

Authors	Affiliations
Yoshua Bengio	A.M. Turing Award recipient, Mila - Quebec AI Institute, Université de Montréal, Canada CIFAR AI Chair
Geoffrey Hinton	A.M. Turing Award recipient, University of Toronto, Vector Institute
Andrew Yao	A.M. Turing Award recipient, Tsinghua University
Dawn Song	UC Berkeley
Pieter Abbeel	UC Berkeley
Yuval Noah Harari	The Hebrew University of Jerusalem, Department of History
Ya-Qin Zhang	Tsinghua University
Lan Xue	Tsinghua University, Institute for AI International Governance
Shai Shalev-Shwartz	The Hebrew University of Jerusalem
Gillian Hadfield	University of Toronto, SR Institute for Technology and Society, Vector Institute
Jeff Clune	University of British Columbia, Canada CIFAR AI Chair, Vector Institute
Tegan Maharaj	University of Toronto, Vector Institute
Frank Hutter	University of Freiburg
Atılım Güneş Baydin	University of Oxford
Sheila McIlraith	University of Toronto, Vector Institute
Qiqi Gao	East China University of Political Science and Law
Ashwin Acharya	Institute for AI Policy and Strategy
David Krueger	University of Cambridge
Anca Dragan	UC Berkeley
Philip Torr	University of Oxford
Stuart Russell	UC Berkeley
Daniel Kahneman	Nobel laureate, Princeton University, School of Public and International Affairs
Jan Brauner*	University of Oxford
Sören Mindermann*	Mila - Quebec AI Institute, Université de Montréal

ARXIV
Forthcoming.

Paper PDF copy | Policy supplement

Abstract:

In this short consensus paper, we outline risks from upcoming, advanced AI systems. We examine large-scale social harms and malicious uses, as well as an irreversible loss of human control over autonomous AI systems. In light of rapid and continuing AI progress, we propose urgent priorities for AI R&D and governance.

Octobre 2023

REVIEW

doi:10.1038/nature14539

Deep learning

Yann LeCun[1,2], Yoshua Bengio[3] & Geoffrey Hinton[4,5]

Deep learning allows computational models that are composed of multiple processing layers to learn representations of data with multiple levels of abstraction. These methods have dramatically improved the state-of-the-art in speech recognition, visual object recognition, object detection and many other domains such as drug discovery and genomics. Deep learning discovers intricate structure in large data sets by using the backpropagation algorithm to indicate how a machine should change its internal parameters that are used to compute the representation in each layer from the representation in the previous layer. Deep convolutional nets have brought about breakthroughs in processing images, video, speech and audio, whereas recurrent nets have shone light on sequential data such as text and speech.

Machine-learning technology powers many aspects of modern society: from web searches to content filtering on social networks to recommendations on e-commerce websites, and it is increasingly present in consumer products such as cameras and smartphones. Machine-learning systems are used to identify objects in images, transcribe speech into text, match news items, posts or products with users' interests, and select relevant results of search. Increasingly, these applications make use of a class of techniques called deep learning.

Conventional machine-learning techniques were limited in their ability to process natural data in their raw form. For decades, constructing a pattern-recognition or machine-learning system required careful engineering and considerable domain expertise to design a feature extractor that transformed the raw data (such as the pixel values of an image) into a suitable internal representation or feature vector from which the learning subsystem, often a classifier, could detect or classify patterns in the input.

Representation learning is a set of methods that allows a machine to be fed with raw data and to automatically discover the representations needed for detection or classification. Deep-learning methods are representation-learning methods with multiple levels of representation, obtained by composing simple but non-linear modules that each transform the representation at one level (starting with the raw input) into a representation at a higher, slightly more abstract level. With the composition of enough such transformations, very complex functions can be learned. For classification tasks, higher layers of representation amplify aspects of the input that are important for discrimination and suppress irrelevant variations. An image, for example, comes in the form of an array of pixel values, and the learned features in the first layer of representation typically represent the presence or absence of edges at particular orientations and locations in the image. The second layer typically detects motifs by spotting particular arrangements of edges, regardless of small variations in the edge positions. The third layer may assemble motifs into larger combinations that correspond to parts of familiar objects, and subsequent layers would detect objects as combinations of these parts. The key aspect of deep learning is that these layers of features are not designed by human engineers: they are learned from data using a general-purpose learning procedure.

Deep learning is making major advances in solving problems that have resisted the best attempts of the artificial intelligence community for many years. It has turned out to be very good at discovering intricate structures in high-dimensional data and is therefore applicable to many domains of science, business and government. In addition to beating records in image recognition[1–4] and speech recognition[5–7], it has beaten other machine-learning techniques at predicting the activity of potential drug molecules[8], analysing particle accelerator data[9,10], reconstructing brain circuits[11], and predicting the effects of mutations in non-coding DNA on gene expression and disease[12,13]. Perhaps more surprisingly, deep learning has produced extremely promising results for various tasks in natural language understanding[14], particularly topic classification, sentiment analysis, question answering[15] and language translation[16,17].

We think that deep learning will have many more successes in the near future because it requires very little engineering by hand, so it can easily take advantage of increases in the amount of available computation and data. New learning algorithms and architectures that are currently being developed for deep neural networks will only accelerate this progress.

Supervised learning

The most common form of machine learning, deep or not, is supervised learning. Imagine that we want to build a system that can classify images as containing, say, a house, a car, a person or a pet. We first collect a large data set of images of houses, cars, people and pets, each labelled with its category. During training, the machine is shown an image and produces an output in the form of a vector of scores, one for each category. We want the desired category to have the highest score of all categories, but this is unlikely to happen before training. We compute an objective function that measures the error (or distance) between the output scores and the desired pattern of scores. The machine then modifies its internal adjustable parameters to reduce this error. These adjustable parameters, often called weights, are real numbers that can be seen as 'knobs' that define the input–output function of the machine. In a typical deep-learning system, there may be hundreds of millions of these adjustable weights, and hundreds of millions of labelled examples with which to train the machine.

To properly adjust the weight vector, the learning algorithm computes a gradient vector that, for each weight, indicates by what amount the error would increase or decrease if the weight were increased by a tiny amount. The weight vector is then adjusted in the opposite direction to the gradient vector.

The objective function, averaged over all the training examples, can

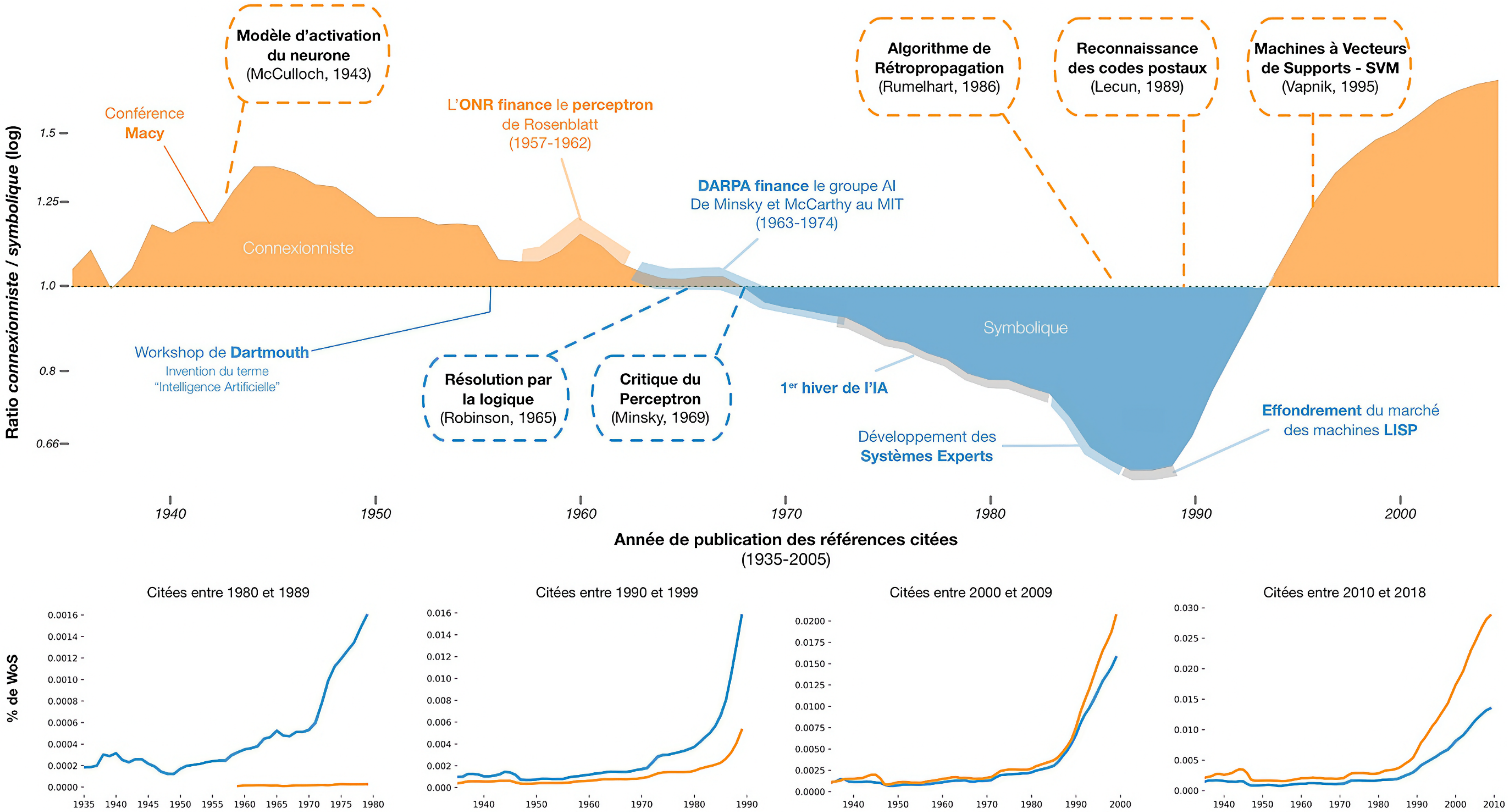
---

[1]Facebook AI Research, 770 Broadway, New York, New York 10003 USA. [2]New York University, 715 Broadway, New York, New York 10003, USA. [3]Department of Computer Science and Operations Research Université de Montréal, Pavillon André-Aisenstadt, PO Box 6128 Centre-Ville STN Montréal, Quebec H3C 3J7, Canada. [4]Google, 1600 Amphitheatre Parkway, Mountain View, California 94043, USA. [5]Department of Computer Science, University of Toronto, 6 King's College Road, Toronto, Ontario M5S 3G4, Canada.

© 2015 Macmillan Publishers Limited. All rights reserved

2015

Yann LeCun, Yoshua Bengio & Geoffrey Hinton, *Deep learning, Nature* 2015 (vol.521), pp. 436-444.



Modèle d'activation du neurone
(McCulloch, 1943)

Conférence **Macy**

L'**ONR finance** le **perceptron** de Rosenblatt (1957-1962)

DARPA finance le groupe AI
De Minsky et McCarthy au MIT
(1963-1974)

Algorithme de Rétropropagation
(Rumelhart, 1986)

Reconnaissance des codes postaux
(Lecun, 1989)

Machines à Vecteurs de Supports - SVM
(Vapnik, 1995)

Workshop de **Dartmouth**
Invention du terme
"Intelligence Artificielle"

Résolution par la logique
(Robinson, 1965)

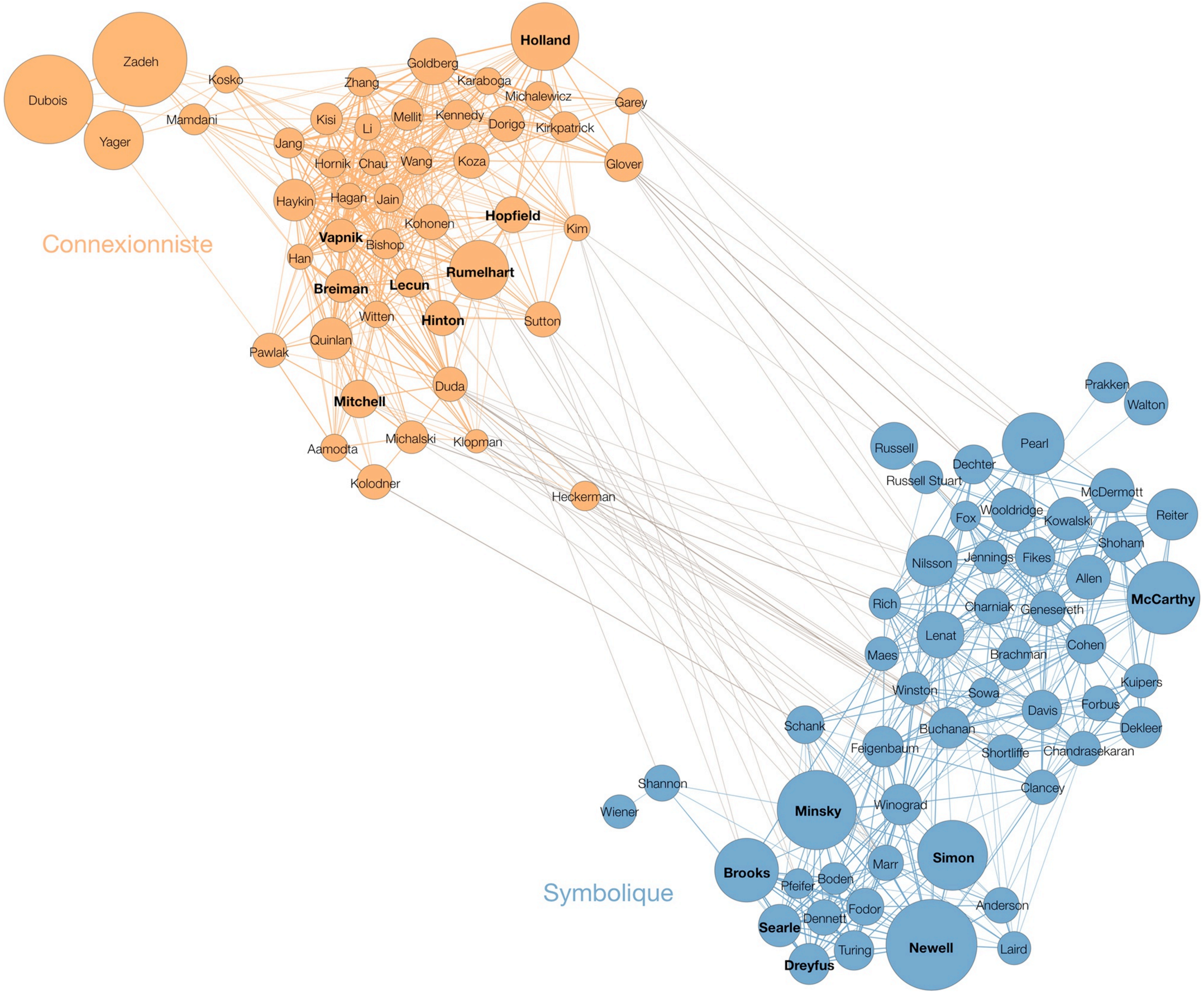
Critique du Perceptron
(Minsky, 1969)

1er hiver de l'IA

Développement des **Systèmes Experts**

Effondrement du marché des machines **LISP**

Dominique Cardon, Jean-Philippe Cointet, Antoine Mazières, *La revanche des neurones*, Réseaux 2018/5 (n° 211), pp. 173 à 220.



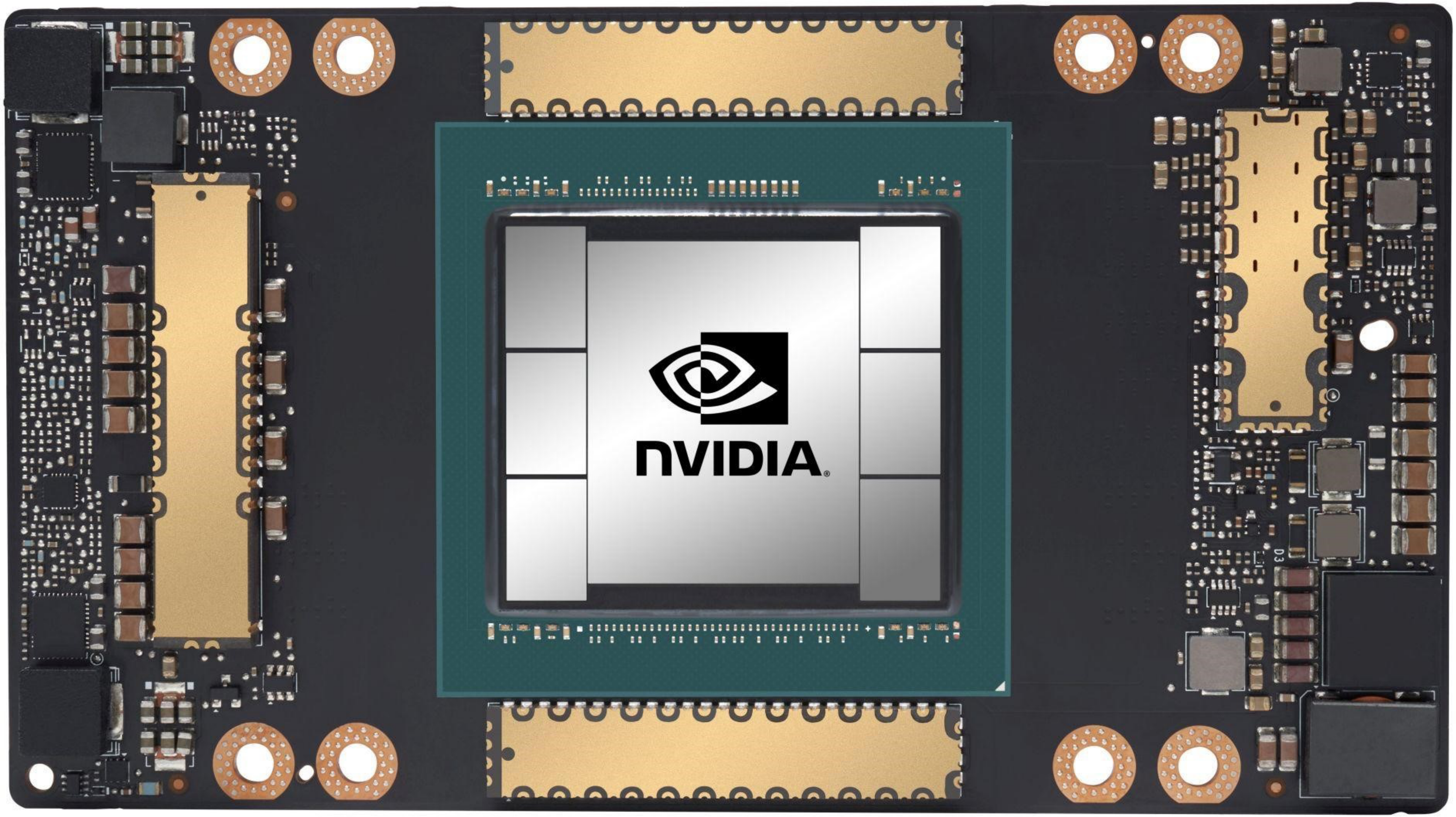
Dominique Cardon, Jean-Philippe Cointet, Antoine Mazières, *La revanche des neurones*, Réseaux 2018/5 (n° 211), pp. 173 à 220.

Principes du Deep Learning

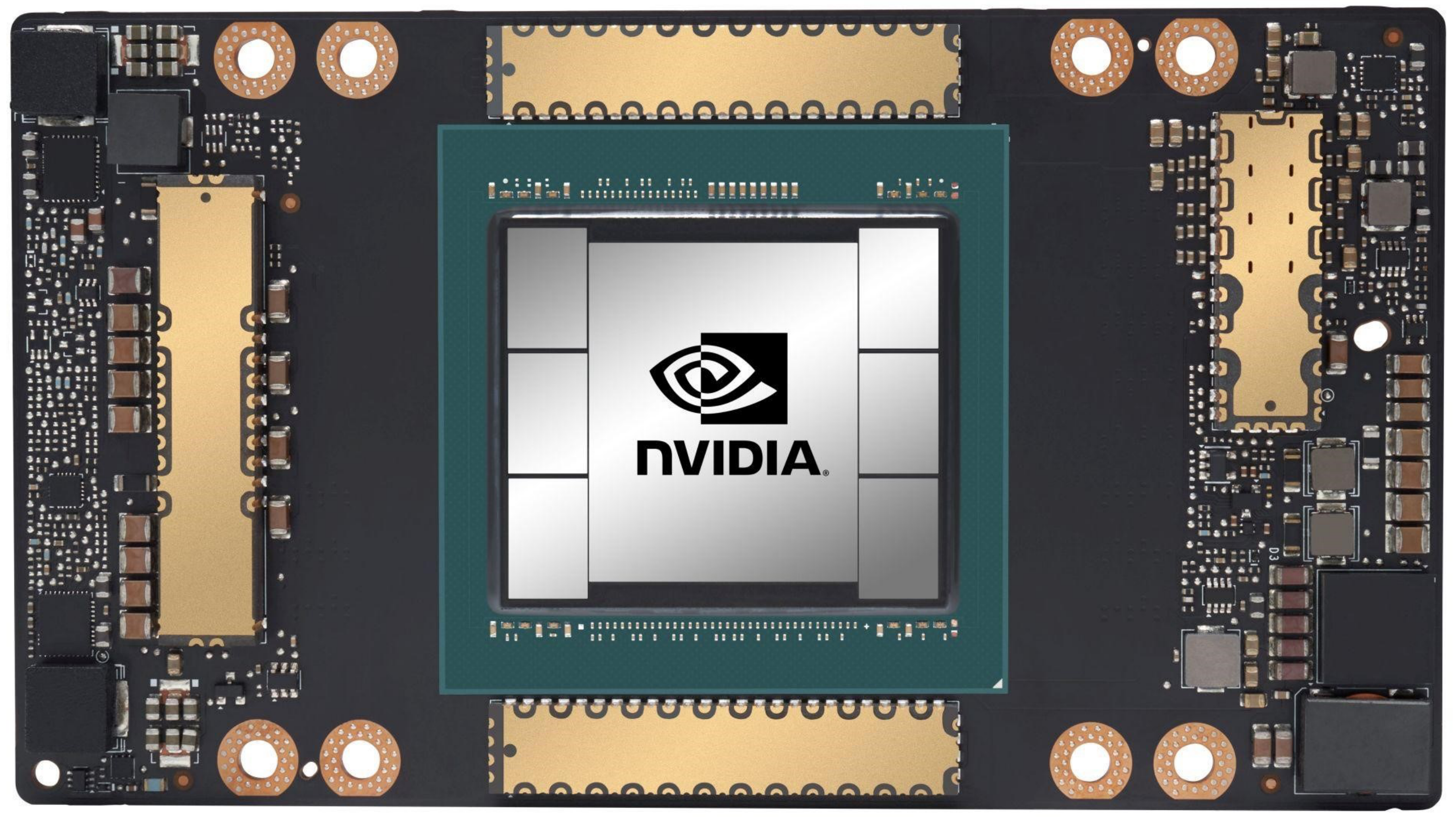
Quelles nouveautés ?

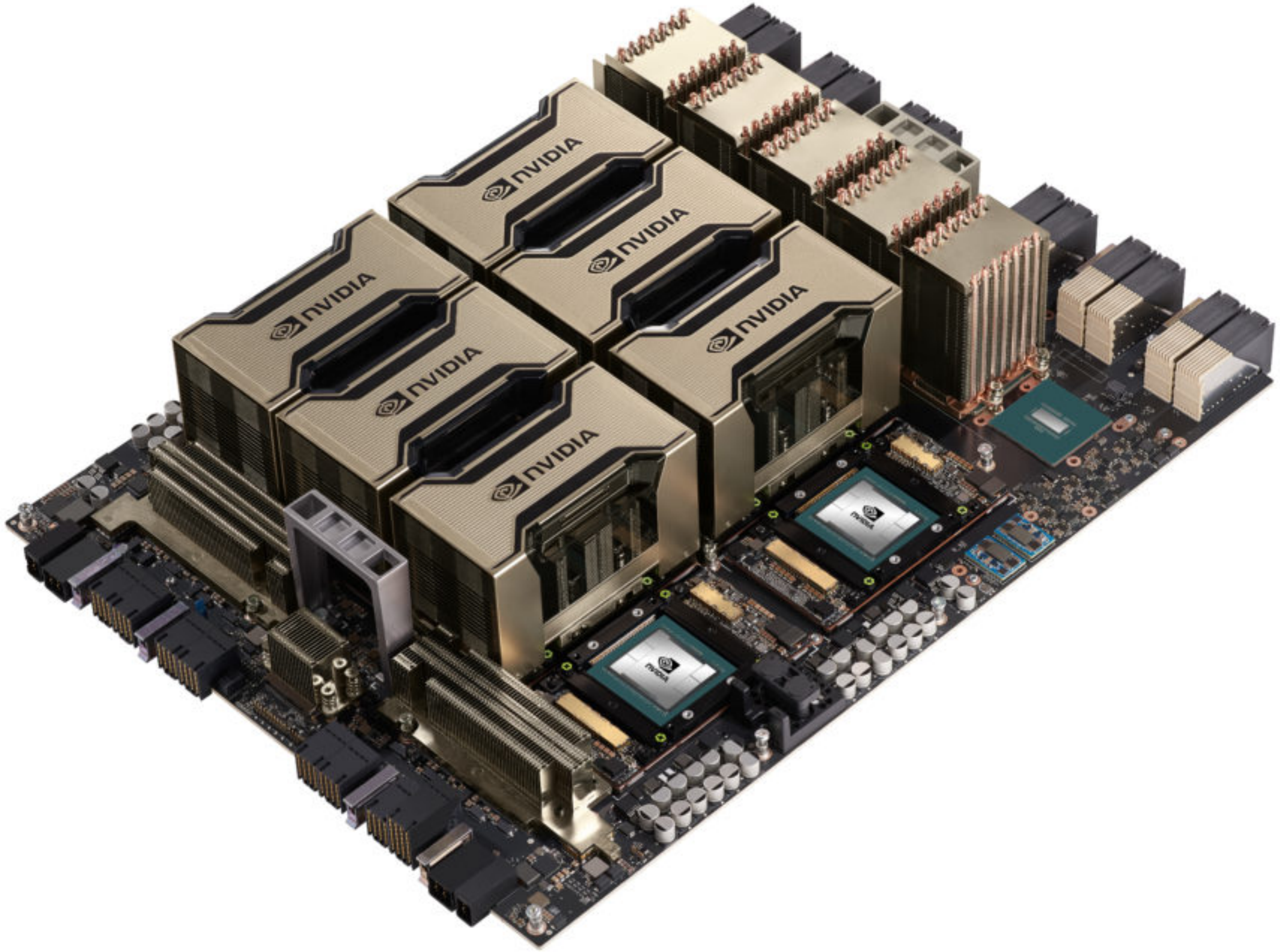
- Puissance de traitement - l’augmentation exponentielle des capacités
- Données - l’abondance de textes, d’images et de vidéos
- Algorithmes - le passage de l’expertise à l’expérience
- Modèles économiques - la tentation de l’ouverture

La puissance de calcul

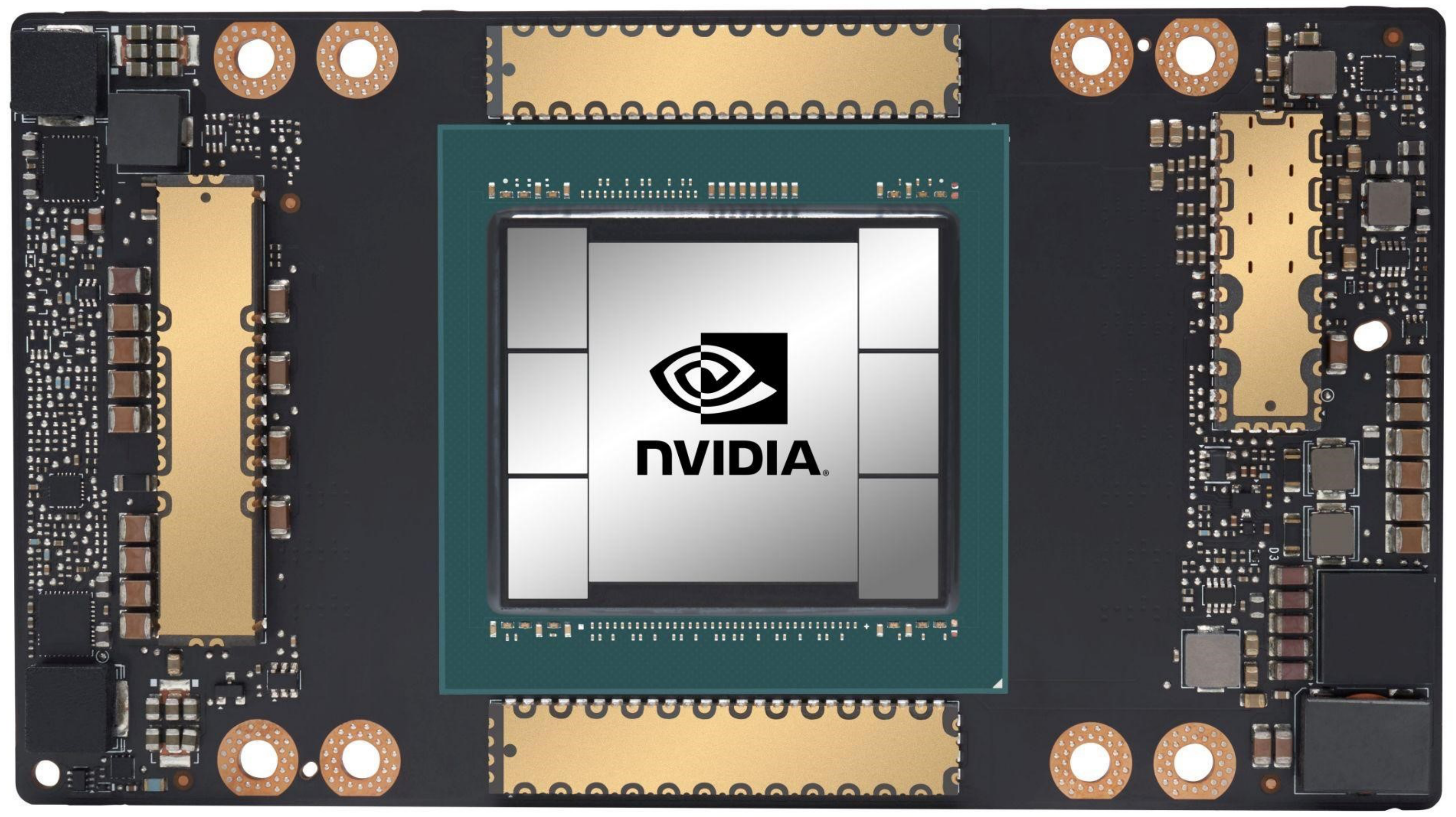


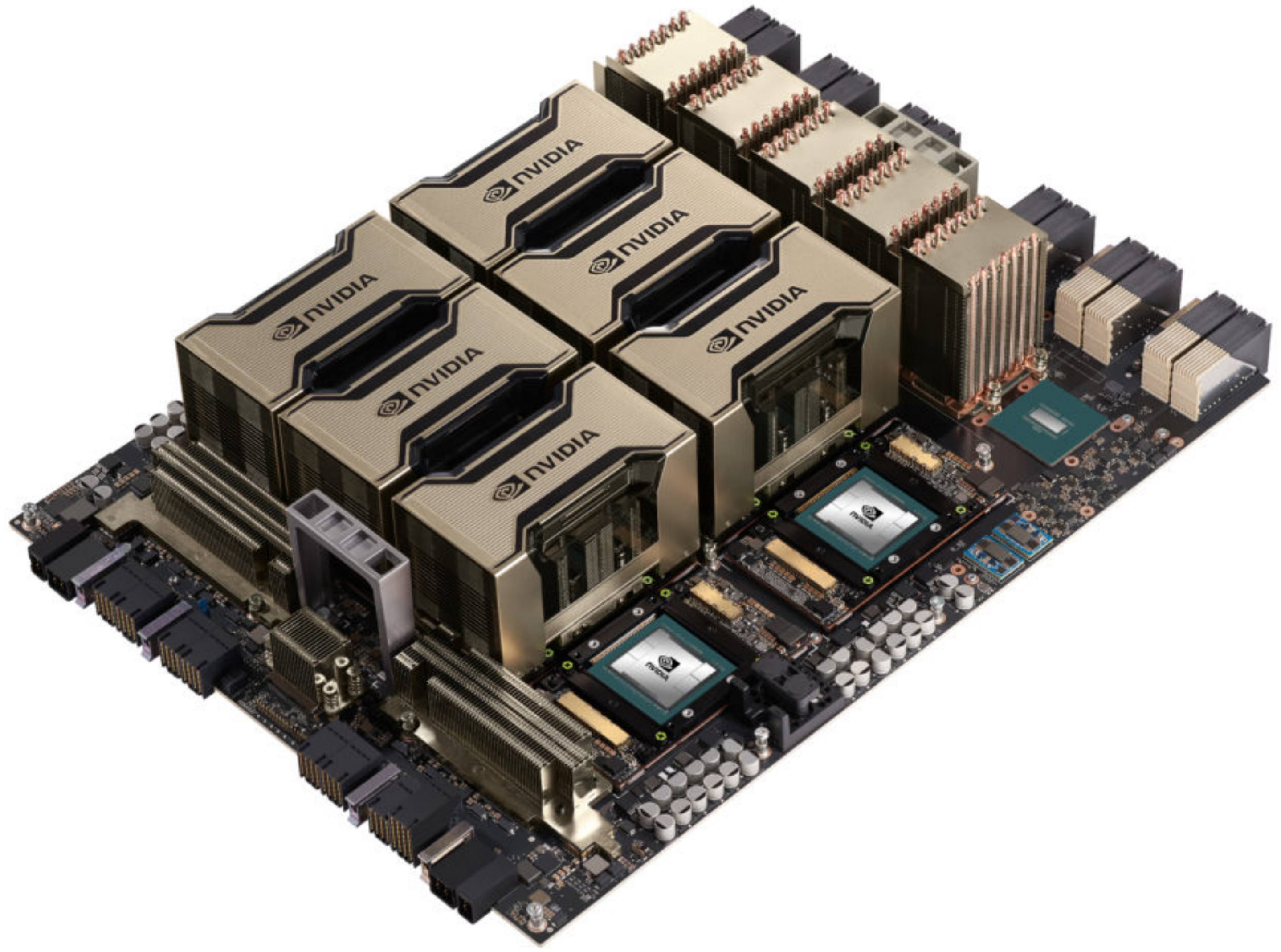
NVIDIA Technical Blog, NVIDIA Ampere Architecture In-Depth, 14 mai 2020

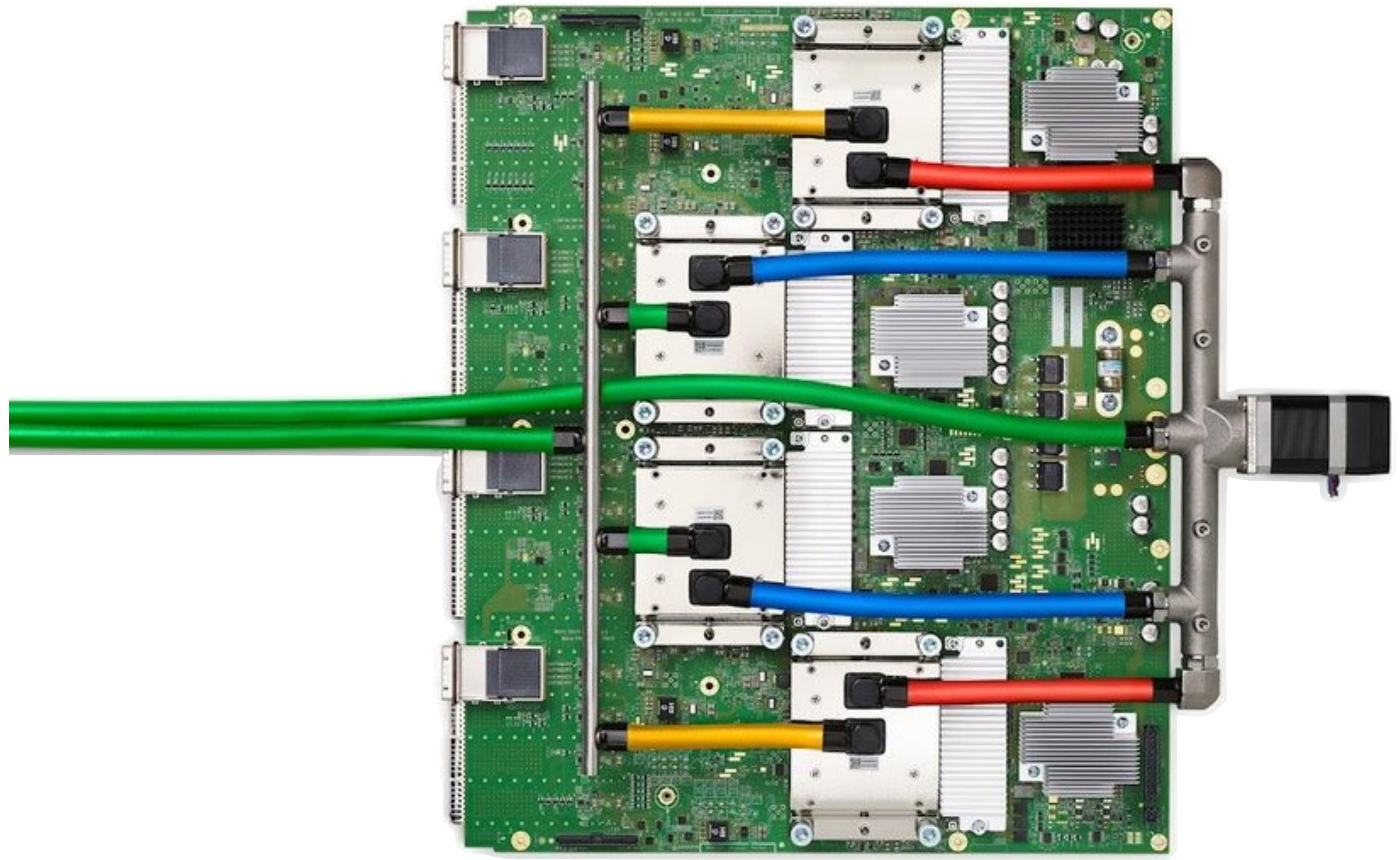




NVIDIA Technical Blog, NVIDIA Ampere Architecture In-Depth, 14 mai 2020







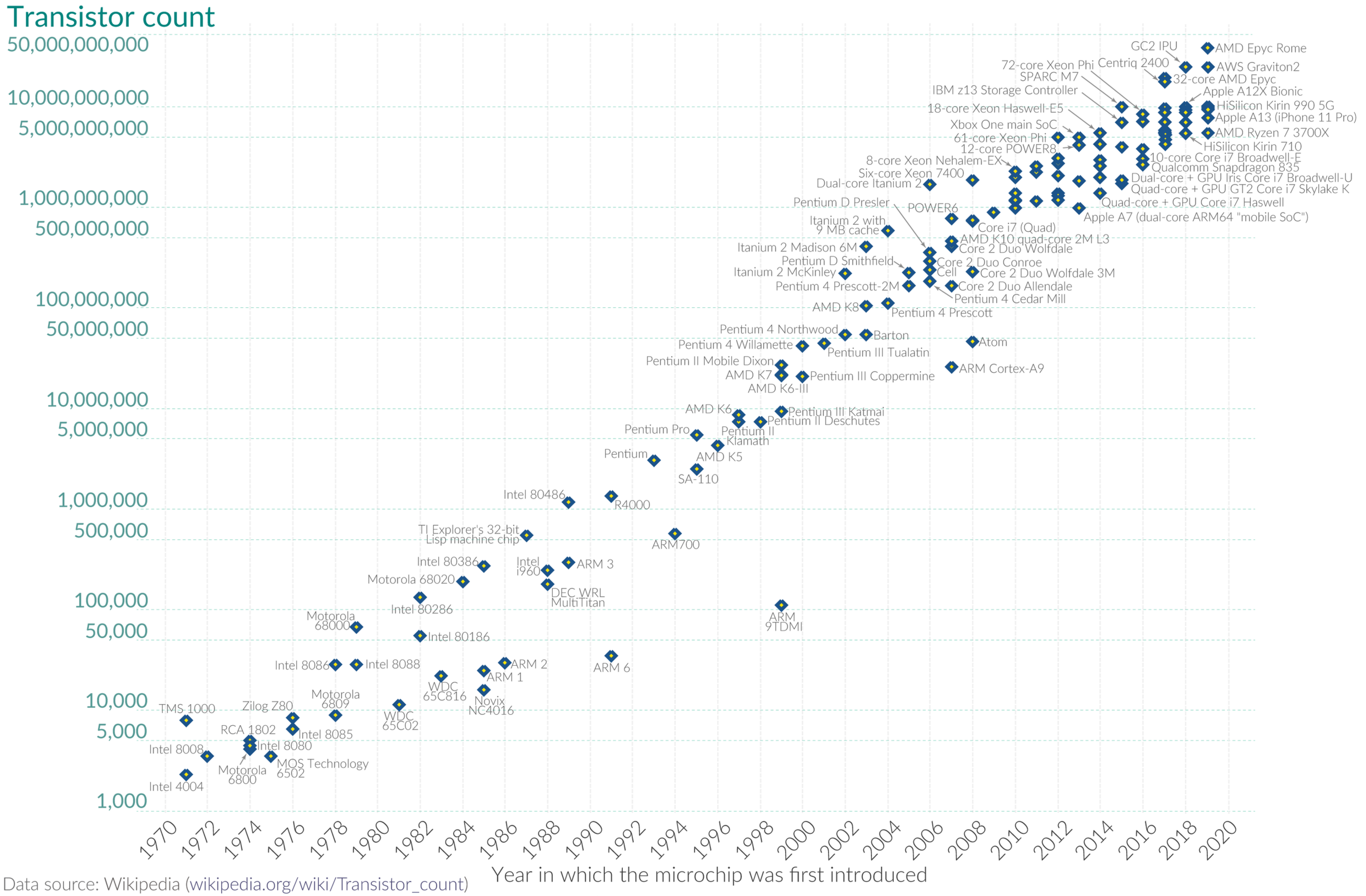
NVIDIA Technical Blog, NVIDIA Ampere Architecture In-Depth, 14 mai 2020 & Google, TPUv4, 14 décembre 2022

Google Cloud - Inside Google Cloud TPU Data Center - 29 août 2023

Moore's Law: The number of transistors on microchips doubles every two years

Our World in Data

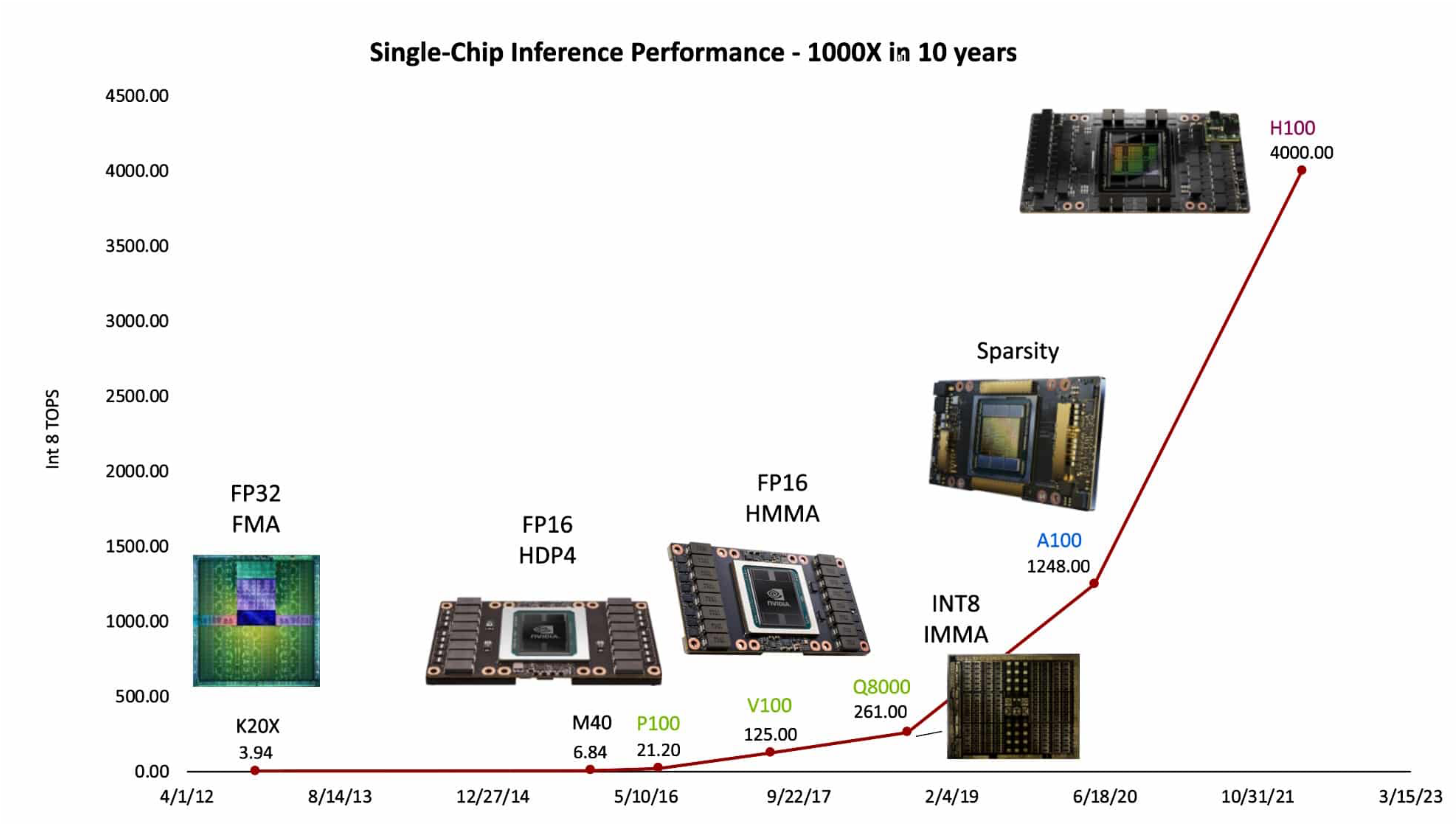
Moore's law describes the empirical regularity that the number of transistors on integrated circuits doubles approximately every two years. This advancement is important for other aspects of technological progress in computing – such as processing speed or the price of computers.



Data source: Wikipedia (wikipedia.org/wiki/Transistor_count)

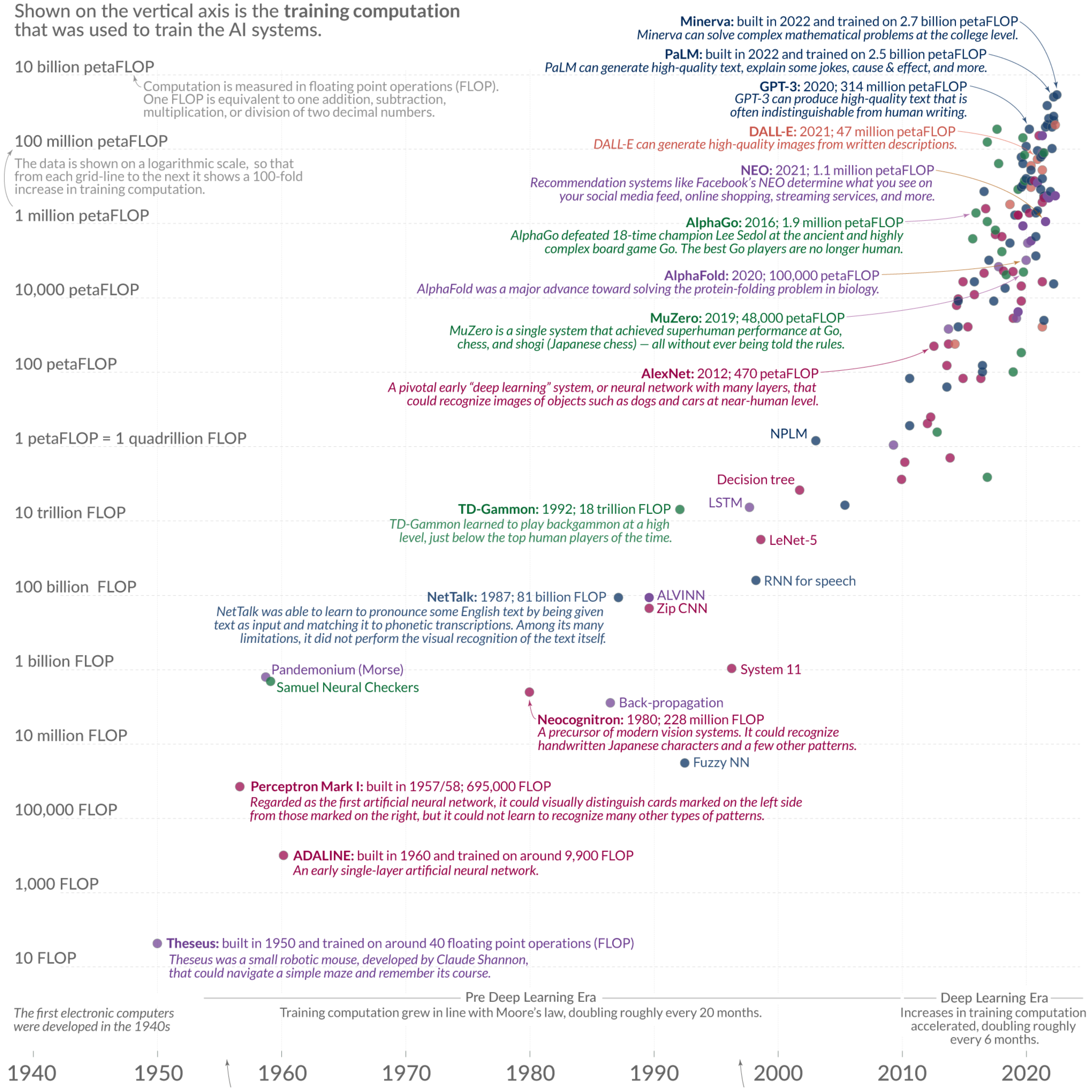
OurWorldinData.org – Research and data to make progress against the world's largest problems.

Licensed under CC-BY by the authors Hannah Ritchie and Max Roser.



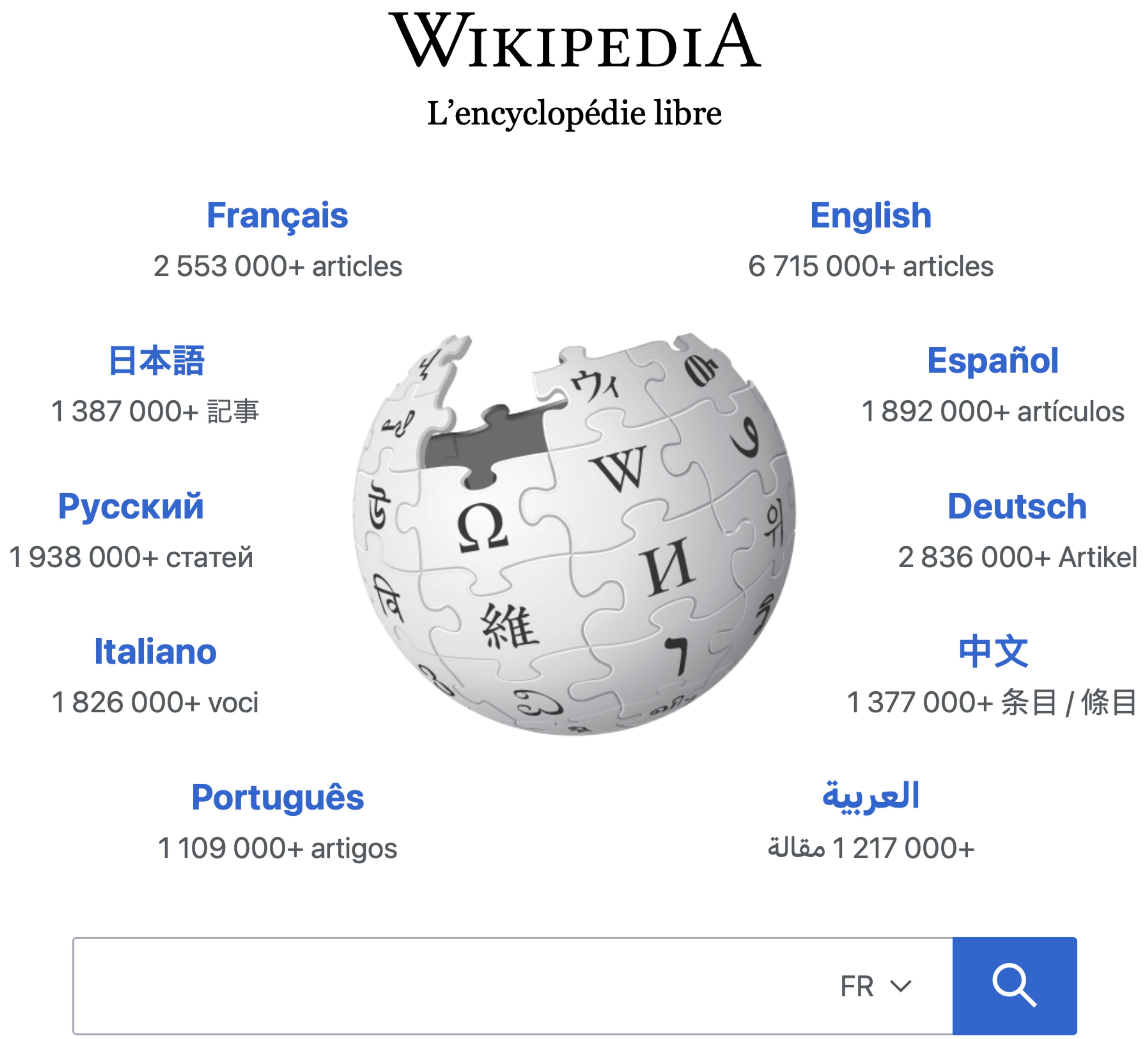
De la loi de Moore à la loi de Huang

NVIDIA, Heeding Huang's Law, 29 septembre 2023



Charlie Giattino, Edouard Mathieu and Max Roser, The rise of artifical intelligence over the last 8 decades, 6 décembre 2022

Les données



WIKIPEDIA

L’encyclopédie libre

Français
2 553 000+ articles

English
6 715 000+ articles

日本語
1 387 000+ 記事

Español
1 892 000+ artículos

Русский
1 938 000+ статей

Deutsch
2 836 000+ Artikel

Italiano
1 826 000+ voci

中文
1 377 000+ 条目 / 條目

Português
1 109 000+ artigos

العربية
+1 217 000 مقالة

Common Crawl maintains a free, open repository of web crawl data that can be used by anyone.

Common Crawl is a 501(c)(3) non–profit founded in 2007.

We make wholesale extraction, transformation and analysis of open web data accessible to researchers.

Overview

LAION
Projects
Team
Blog
Notes
Press
About
FAQ
Donations
Privacy Policy
Dataset Requests
Impressum
Donate

LAION

Large-scale Artificial Intelligence Open Network

TRULY OPEN AI. 100% NON-PROFIT. 100% FREE.

LAION, as a non-profit organization, provides datasets, tools and models to liberate machine learning research. By doing so, we encourage open public education and a more environment-friendly use of resources by reusing existing datasets and models.

A CERN FOR OPEN SOURCE LARGE-SCALE AI RESEARCH!

Sign our petition for Open Source AI research!

Join our community on discord!

LAION-400M

An open dataset containing 400 million English image-text pairs.

LAION-5B

A dataset consisting of 5.85 billion multilingual CLIP-filtered image-text pairs.

Clip H/14

The largest CLIP (Contrastive Language-Image Pre-training) vision transformer model.

LAION-Aesthetics

A subset of LAION-5B filtered by a model trained to score aesthetically pleasing images.

Looking to work on tasks? Sign in as a Worker | Learn more
amazon mechanical turk
Overview
Features
Pricing
Help
Developer Resources
Customers
Sign in as a Requester

Amazon Mechanical Turk

Access a global, on-demand, 24x7 workforce

Get started with Amazon Mechanical Turk

Looking for data labeling solutions to power Machine Learning models?

Amazon SageMaker Ground Truth allows you to easily build and manage your own data labeling workflows and workforce. Or, use Ground Truth Plus, a turnkey data labeling service that provides an expert workforce and manages it on your behalf. Amazon Mechanical Turk is accessible through both Ground Truth and Ground Truth Plus.

Amazon Mechanical Turk (MTurk) is a crowdsourcing marketplace that makes it easier for individuals and businesses to outsource their processes and jobs to a distributed workforce who can perform these tasks virtually. This could include anything from conducting simple data validation and research to more subjective tasks like survey participation, content moderation, and more. MTurk enables companies to harness the collective intelligence, skills, and insights from a global workforce to streamline business processes, augment data collection and analysis, and accelerate machine learning development.

While technology continues to improve, there are still many things that human beings can do much more effectively than computers, such as moderating content, performing data deduplication, or research. Traditionally, tasks like this have been accomplished by hiring a large temporary workforce, which is time consuming, expensive and difficult to scale, or have gone undone. Crowdsourcing is a good way to break down a manual, time-consuming project into smaller, more manageable tasks to be completed by distributed workers over the Internet (also known as ‘microtasks’).

La nouvelle ère du développement de l’IA

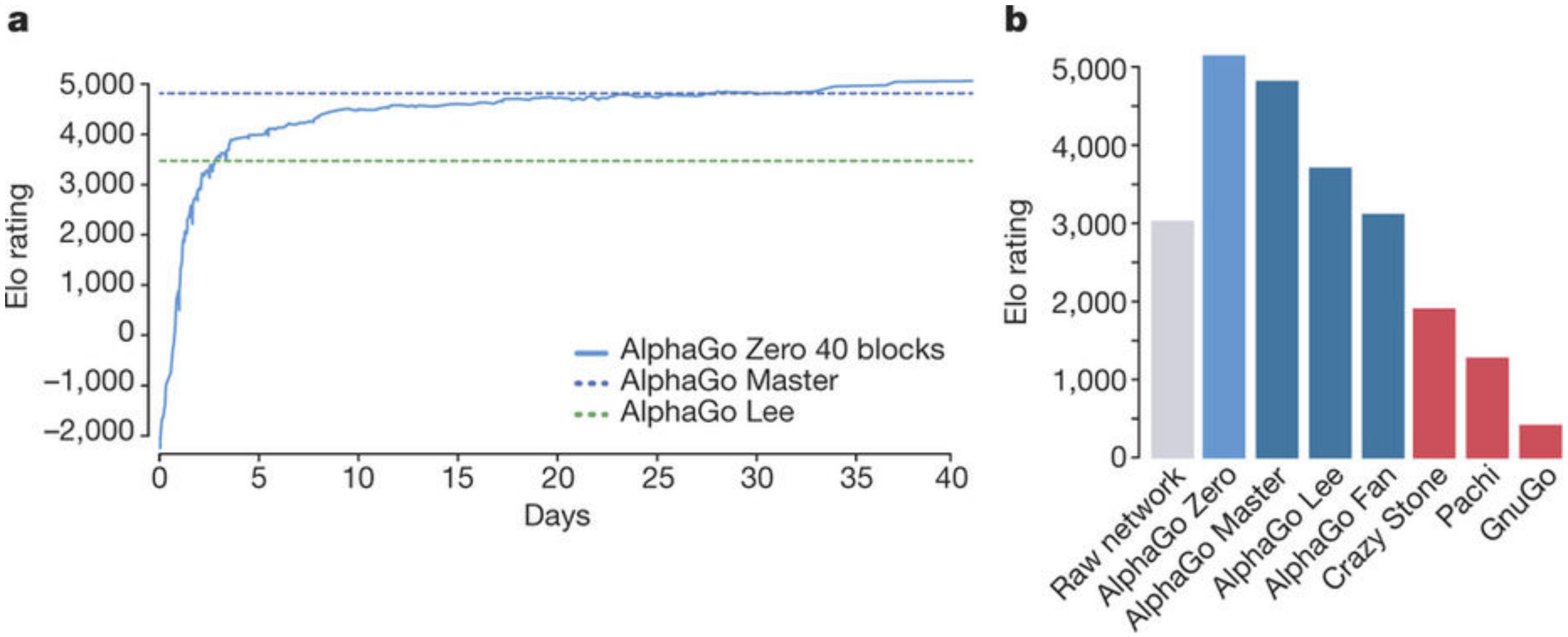
Relevez le potentiel de votre modèle grâce à une qualité de données de pointe. La plateforme intégrée d’apprentissage machine Sama, dotée d’une validation avec intervention humaine, simplifie le développement de l’IA, ce qui permet d’obtenir plus rapidement des résultats précis et d’accélérer la mise sur le marché.

Demandez une démo

Ces grandes entreprises mondiales font confiance à Sama

Google Walmart Ford gettyimages Microsoft NASA

Sans données ?



(1) **AlphaGo** Fan is the previously published program that played against Fan Hui in **October 2015**. This program was distributed over many machines using 176 GPUs.

(2) **AlphaGo Lee** is the program that defeated Lee Sedol 4–1 in **March 2016**. It was previously unpublished, but is similar in most regards to AlphaGo Fan - 48 TPUs, rather than GPUs

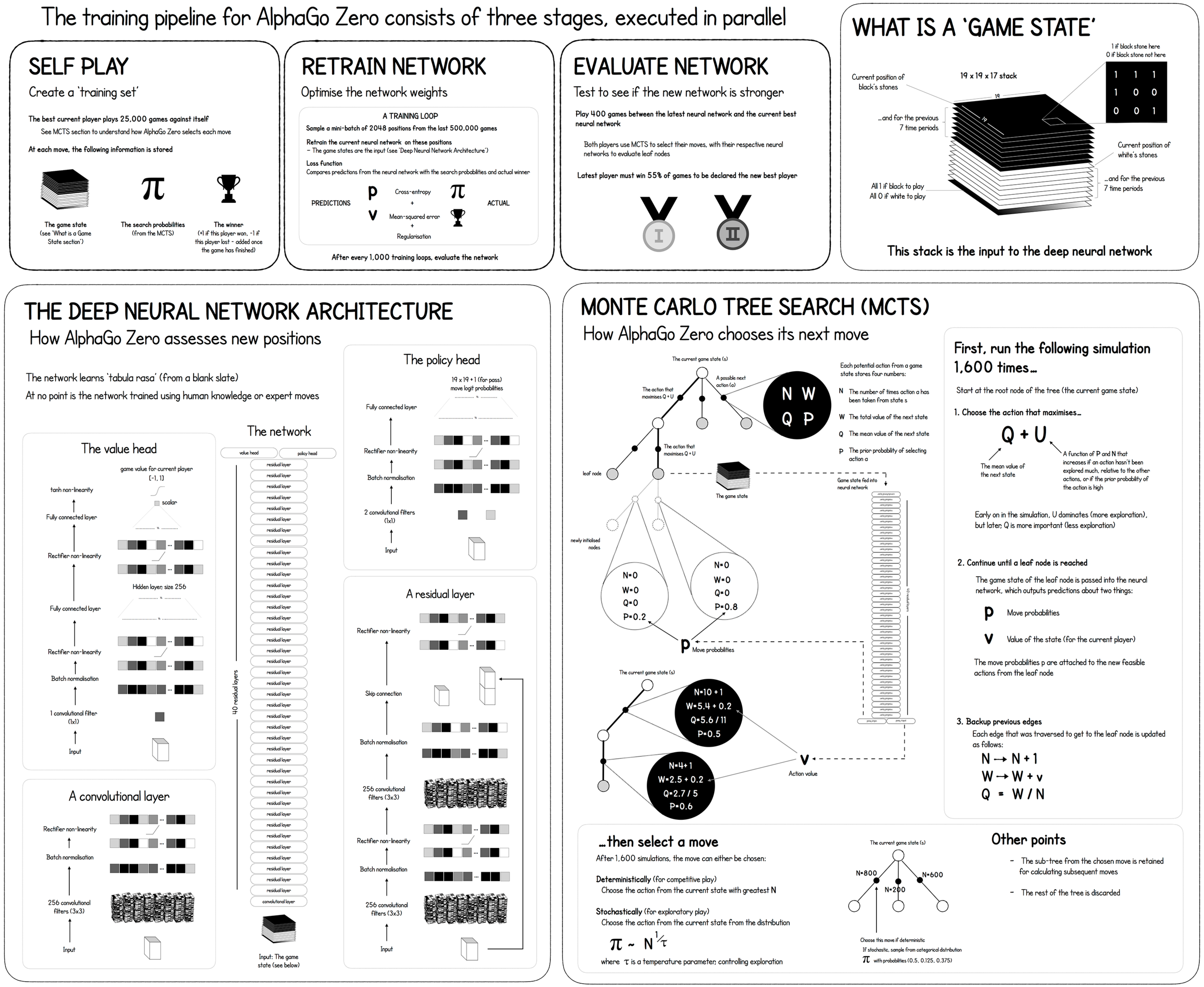
(3) **AlphaGo Master** is the program that defeated top human players by 60–0 in **January 2017** - initialized by supervised learning from human data + reinforcement

(4) **AlphaGo Zero** learns from self-play reinforcement learning, with no human supervision in **April 2017**. It uses just **a single machine in the Google Cloud with 4 TPU**s

(x) **AlphaZero** is a more generalized variant of the AlphaGo Zero (AGZ) algorithm, and is able to play **shogi** and **chess** as well as **Go**.
- Defeated Stockfish 8 after **9 hours** of training.
- Defeated AlphaGo Zero after **34 hours** of training.

Maîtriser le jeu de go sans l’expérience humaine

David Silver et al., *Mastering the game of Go without human knowledge*, 7 avril / 18 octobre 2017, Nature



The training pipeline for AlphaGo Zero consists of three stages, executed in parallel

SELF PLAY

Create a 'training set'

The best current player plays 25,000 games against itself

See MCTS section to understand how AlphaGo Zero selects each move

At each move, the following information is stored

The game state (see 'What is a Game State section')

The search probabilities (from the MCTS)

The winner (+1 if this player won, -1 if this player lost - added once the game has finished)

RETRAIN NETWORK

Optimise the network weights

A TRAINING LOOP

Sample a mini-batch of 2048 positions from the last 500,000 games

Retrain the current neural network on these positions
– The game states are the input (see 'Deep Neural Network Architecture')

Loss function
Compares predictions from the neural network with the search probabilities and actual winner

After every 1,000 training loops, evaluate the network

EVALUATE NETWORK

Test to see if the new network is stronger

Play 400 games between the latest neural network and the current best neural network

Both players use MCTS to select their moves, with their respective neural networks to evaluate leaf nodes

Latest player must win 55% of games to be declared the new best player

WHAT IS A 'GAME STATE'

19 x 19 x 17 stack
19
19
Current position of black's stones
...and for the previous 7 time periods
1 if black stone here
0 if black stone not here
1 1 1
1 0 0
0 0 1
Current position of white's stones
...and for the previous 7 time periods
All 1 if black to play
All 0 if white to play

This stack is the input to the deep neural network

THE DEEP NEURAL NETWORK ARCHITECTURE

How AlphaGo Zero assesses new positions

The network learns 'tabula rasa' (from a blank slate)

At no point is the network trained using human knowledge or expert moves

The value head

The network

The policy head

19 x 19 + 1 (for pass) move logit probabilities
Fully connected layer
Rectifier non-linearity
Batch normalisation
2 convolutional filters (1x1)
Input

A convolutional layer

A residual layer

MONTE CARLO TREE SEARCH (MCTS)

How AlphaGo Zero chooses its next move

Each potential action from a game state stores four numbers:

- **N** The number of times action a has been taken from state s
- **W** The total value of the next state
- **Q** The mean value of the next state
- **P** The prior probability of selecting action a

First, run the following simulation 1,600 times...

Start at the root node of the tree (the current game state)

1. Choose the action that maximises...

$$Q + U$$

The mean value of the next state

A function of P and N that increases if an action hasn't been explored much, relative to the other actions, or if the prior probability of the action is high

Early on in the simulation, U dominates (more exploration), but later, Q is more important (less exploration)

2. Continue until a leaf node is reached

The game state of the leaf node is passed into the neural network, which outputs predictions about two things:

- **p** Move probabilities
- **v** Value of the state (for the current player)

The move probabilities p are attached to the new feasible actions from the leaf node

3. Backup previous edges

Each edge that was traversed to get to the leaf node is updated as follows:

$$N \rightarrow N + 1$$
$$W \rightarrow W + v$$
$$Q = W / N$$

...then select a move

After 1,600 simulations, the move can either be chosen:

Deterministically (for competitive play)
Choose the action from the current state with greatest N

Stochastically (for exploratory play)
Choose the action from the current state from the distribution

$$\pi \sim N^{1/\tau}$$

where τ is a temperature parameter, controlling exploration

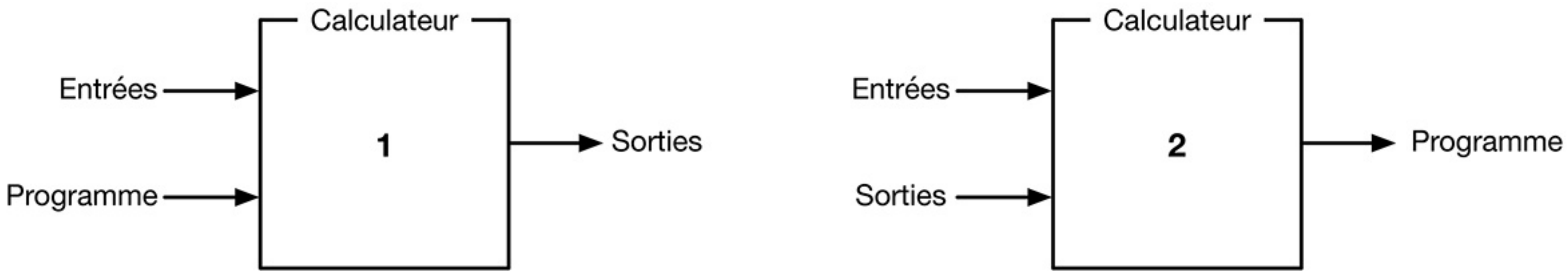
Choose this move if deterministic
If stochastic, sample from categorical distribution π with probabilities (0.5, 0.125, 0.375)

Other points

- The sub-tree from the chosen move is retained for calculating subsequent moves
- The rest of the tree is discarded

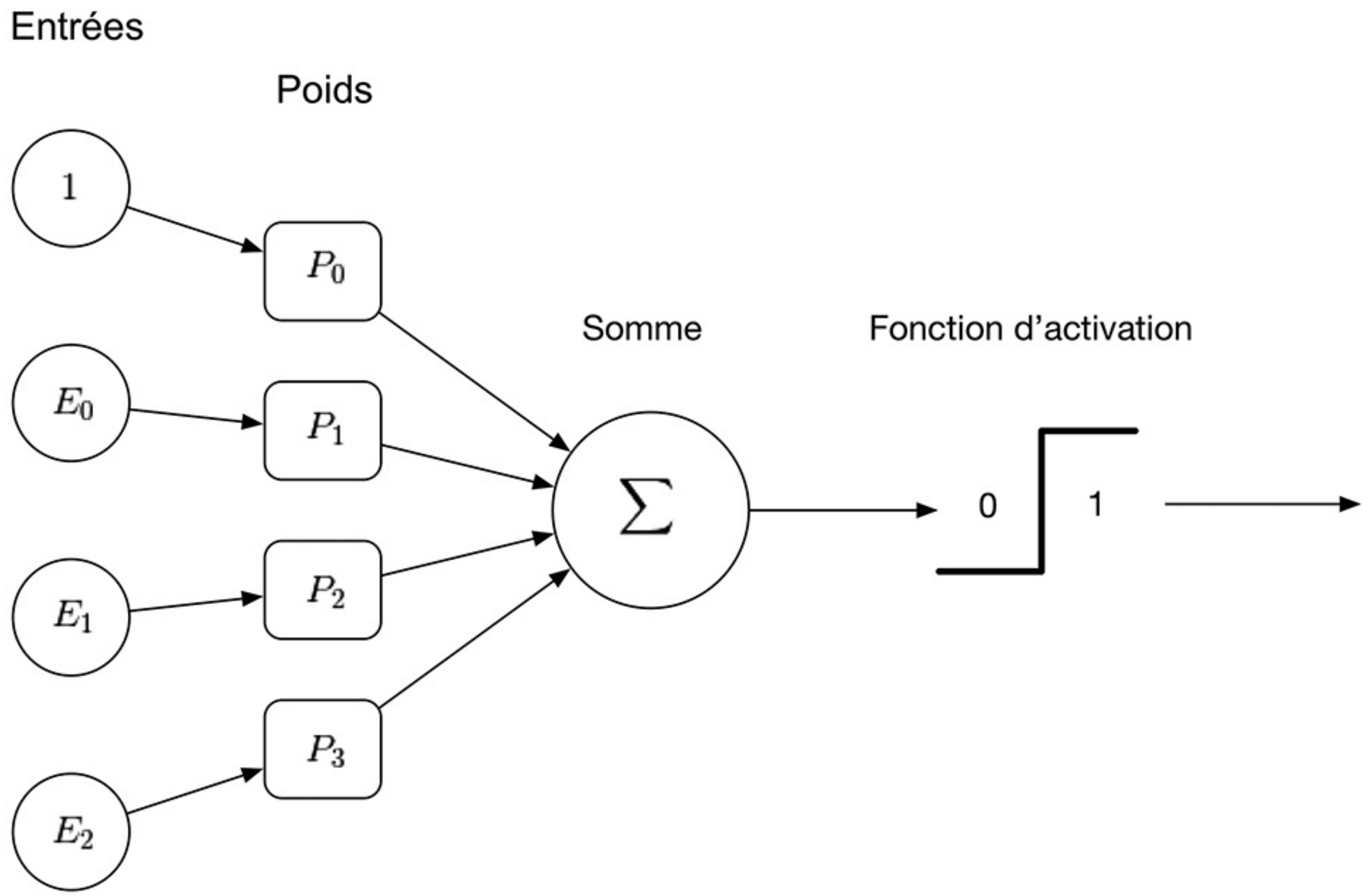
David Foster, *How to build your own AlphaZero AI using Python and Keras*, 26 janvier 2018, Medium

Les algorithmes



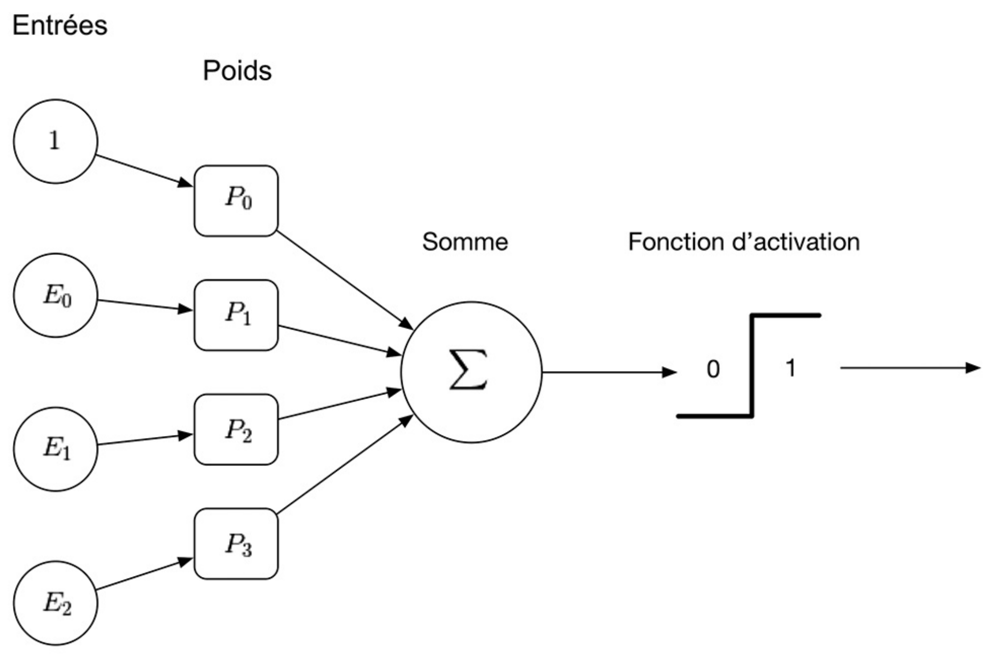
Machine hypothético-déductive (1) et machine inductive (2)

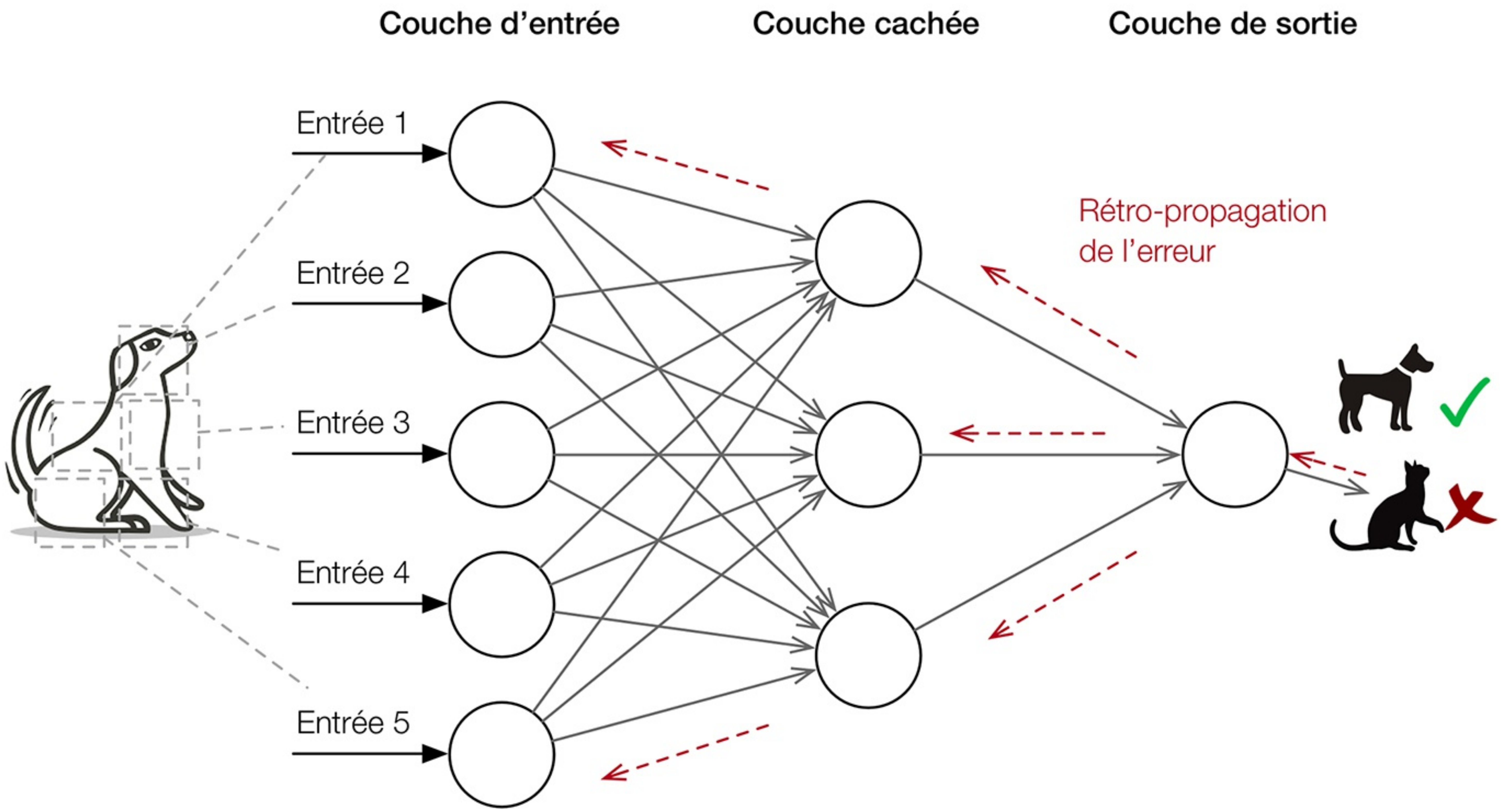
Dominique Cardon, Jean-Philippe Cointet, Antoine Mazières, *La revanche des neurones*, Réseaux 2018/5 (n° 211), pp. 173 à 220.



Modèle formel d'un neurone artificiel à seuil binaire

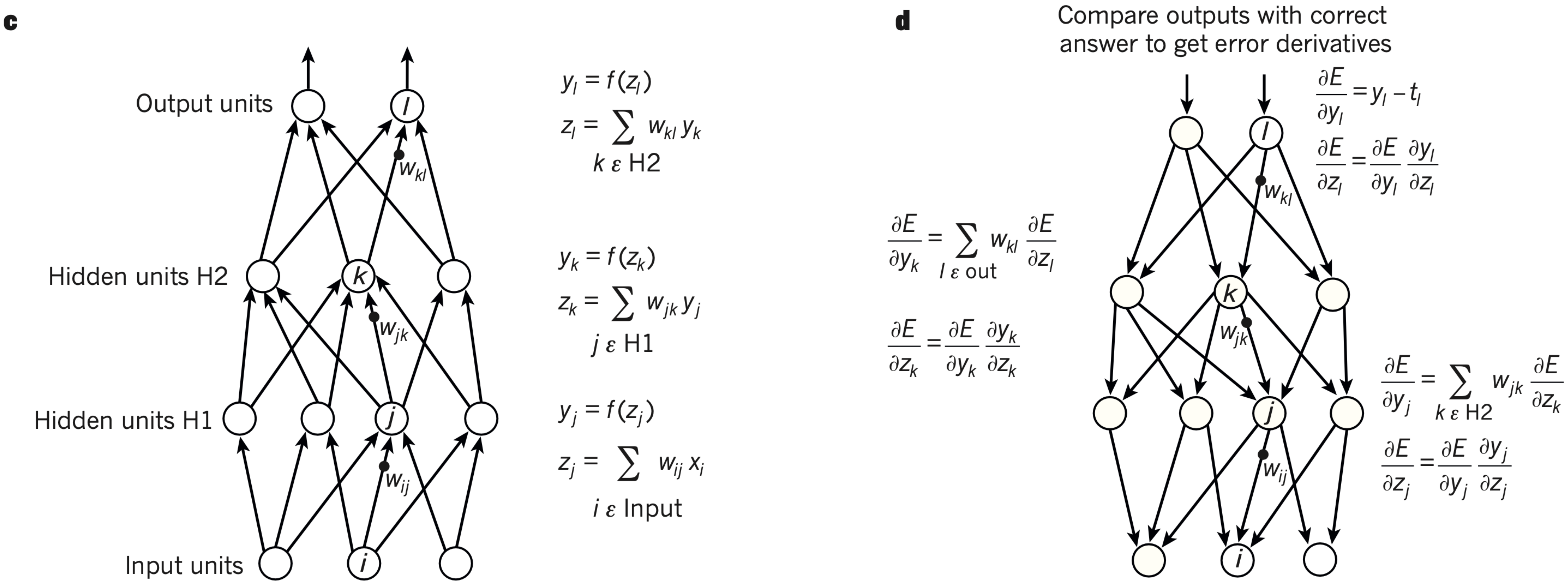
Dominique Cardon, Jean-Philippe Cointet, Antoine Mazières, *La revanche des neurones*, Réseaux 2018/5 (n° 211), pp. 173 à 220.





Fonctionnement d'un réseau de neurones simple

Dominique Cardon, Jean-Philippe Cointet, Antoine Mazières, *La revanche des neurones*, Réseaux 2018/5 (n° 211), pp. 173 à 220.



$$y_l = f(z_l)$$

$$z_l = \sum_{k \,\varepsilon\, \mathrm{H2}} w_{kl}\, y_k$$

$$y_k = f(z_k)$$

$$z_k = \sum_{j \,\varepsilon\, \mathrm{H1}} w_{jk}\, y_j$$

$$y_j = f(z_j)$$

$$z_j = \sum_{i \,\varepsilon\, \mathrm{Input}} w_{ij}\, x_i$$

$$\frac{\partial E}{\partial y_l} = y_l - t_l$$

$$\frac{\partial E}{\partial z_l} = \frac{\partial E}{\partial y_l}\frac{\partial y_l}{\partial z_l}$$

$$\frac{\partial E}{\partial y_k} = \sum_{l \,\varepsilon\, \mathrm{out}} w_{kl}\frac{\partial E}{\partial z_l}$$

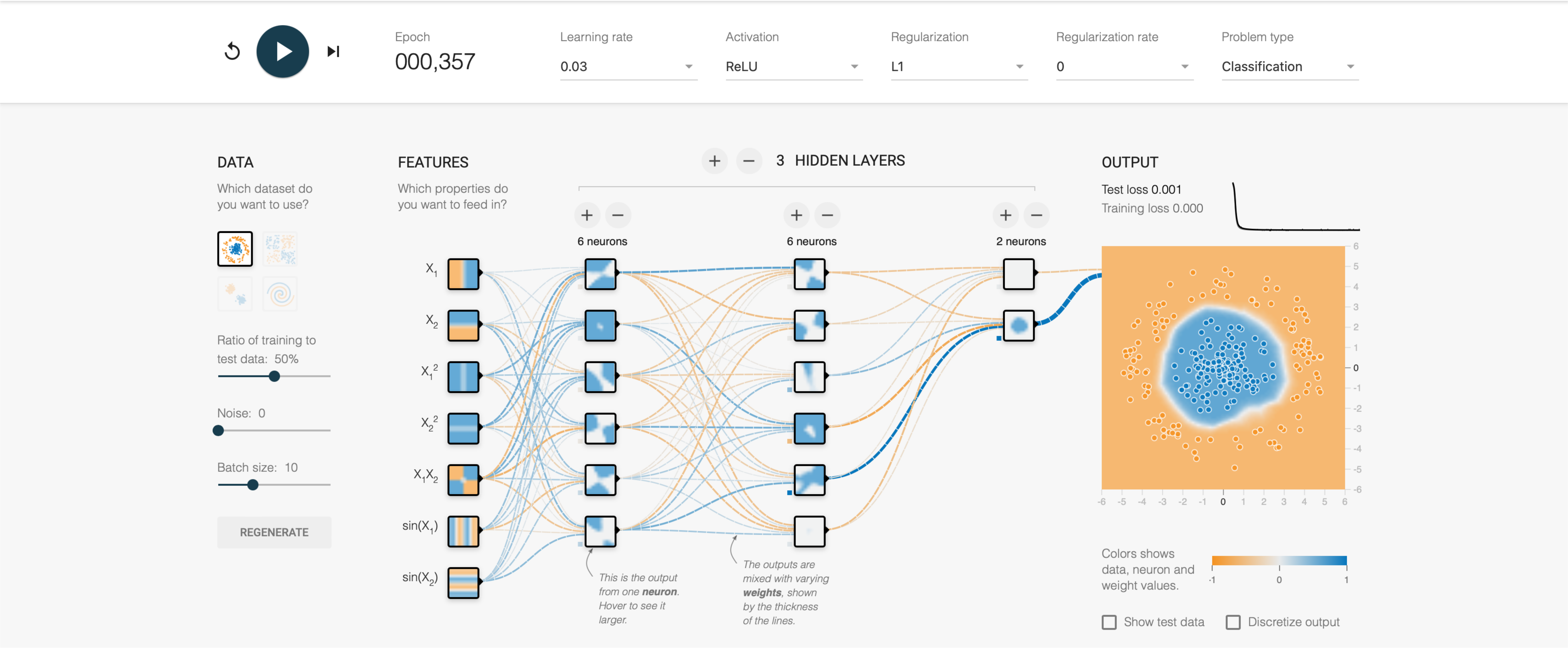
$$\frac{\partial E}{\partial z_k} = \frac{\partial E}{\partial y_k}\frac{\partial y_k}{\partial z_k}$$

$$\frac{\partial E}{\partial y_j} = \sum_{k \,\varepsilon\, \mathrm{H2}} w_{jk}\frac{\partial E}{\partial z_k}$$

$$\frac{\partial E}{\partial z_j} = \frac{\partial E}{\partial y_j}\frac{\partial y_j}{\partial z_j}$$

Réseau de neurones multicouches et rétropropagation

Yann LeCun, Yoshua Bengio & Geoffrey Hinton, *Deep learning, Nature* 2015 (vol.521), pp. 436-444.



Epoch 000,357

Learning rate 0.03

Activation ReLU

Regularization L1

Regularization rate 0

Problem type Classification

DATA

Which dataset do you want to use?

Ratio of training to test data: 50%

Noise: 0

Batch size: 10

FEATURES

Which properties do you want to feed in?

3 HIDDEN LAYERS

OUTPUT

Test loss 0.001

Training loss 0.000

*This is the output from one **neuron**. Hover to see it larger.*

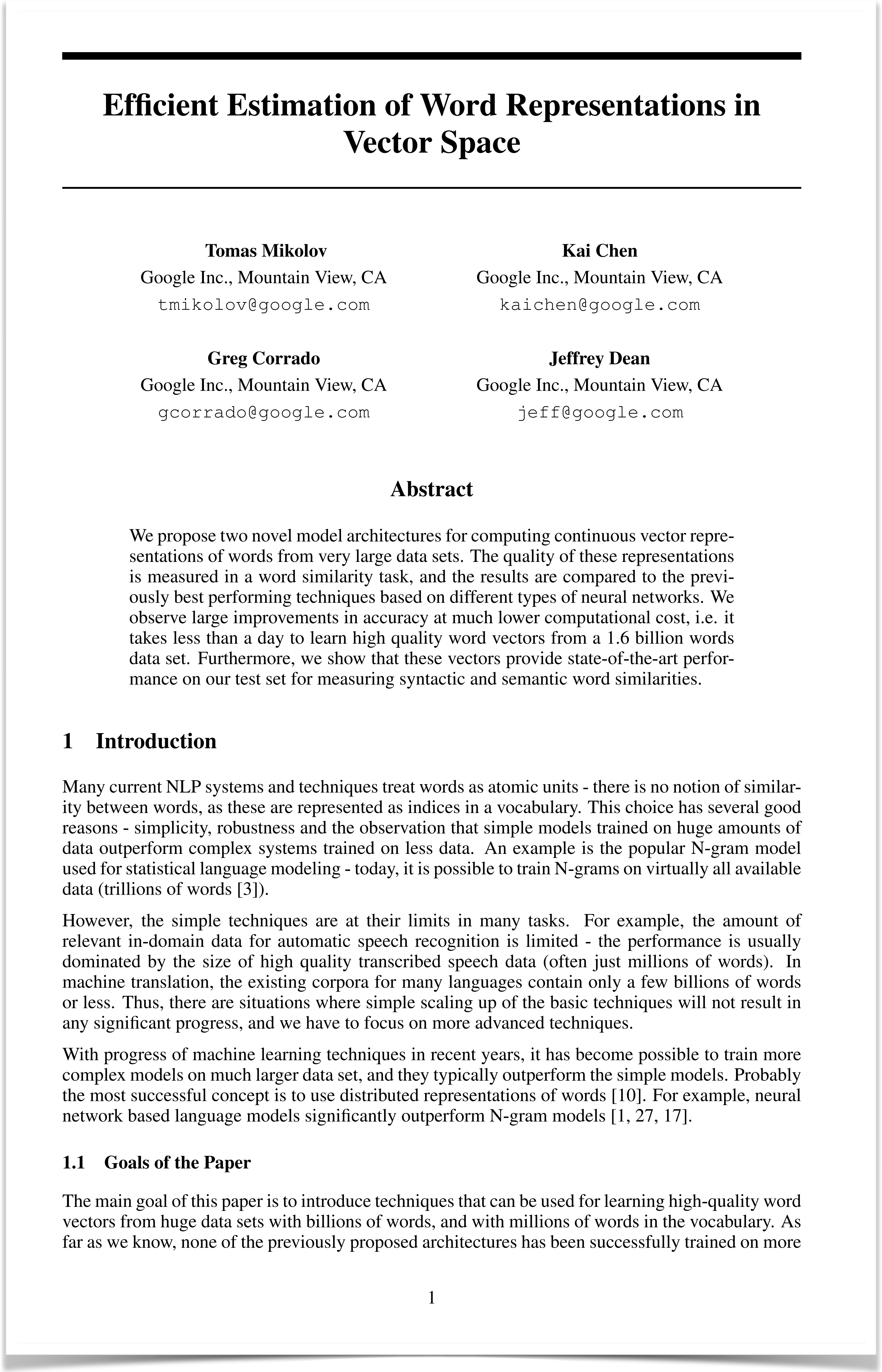
*The outputs are mixed with varying **weights**, shown by the thickness of the lines.*

Colors shows data, neuron and weight values.

☐ Show test data ☐ Discretize output

Tensorflow playground - Neural Network in Your Browser

https://playground.tensorflow.org/



Efficient Estimation of Word Representations in Vector Space

Tomas Mikolov
Google Inc., Mountain View, CA
`tmikolov@google.com`

Kai Chen
Google Inc., Mountain View, CA
`kaichen@google.com`

Greg Corrado
Google Inc., Mountain View, CA
`gcorrado@google.com`

Jeffrey Dean
Google Inc., Mountain View, CA
`jeff@google.com`

Abstract

We propose two novel model architectures for computing continuous vector representations of words from very large data sets. The quality of these representations is measured in a word similarity task, and the results are compared to the previously best performing techniques based on different types of neural networks. We observe large improvements in accuracy at much lower computational cost, i.e. it takes less than a day to learn high quality word vectors from a 1.6 billion words data set. Furthermore, we show that these vectors provide state-of-the-art performance on our test set for measuring syntactic and semantic word similarities.

1 Introduction

Many current NLP systems and techniques treat words as atomic units - there is no notion of similarity between words, as these are represented as indices in a vocabulary. This choice has several good reasons - simplicity, robustness and the observation that simple models trained on huge amounts of data outperform complex systems trained on less data. An example is the popular N-gram model used for statistical language modeling - today, it is possible to train N-grams on virtually all available data (trillions of words [3]).

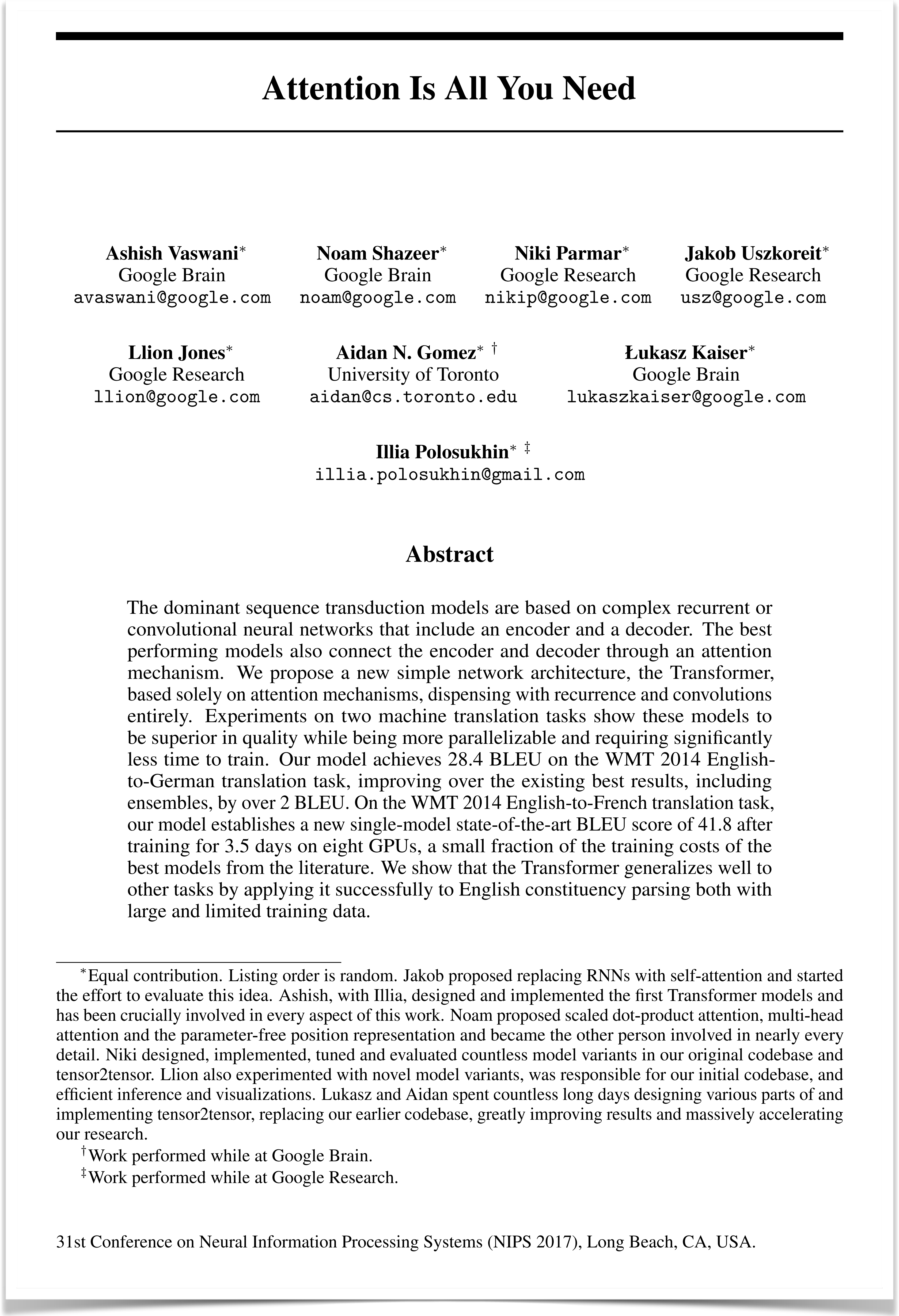
However, the simple techniques are at their limits in many tasks. For example, the amount of relevant in-domain data for automatic speech recognition is limited - the performance is usually dominated by the size of high quality transcribed speech data (often just millions of words). In machine translation, the existing corpora for many languages contain only a few billions of words or less. Thus, there are situations where simple scaling up of the basic techniques will not result in any significant progress, and we have to focus on more advanced techniques.

With progress of machine learning techniques in recent years, it has become possible to train more complex models on much larger data set, and they typically outperform the simple models. Probably the most successful concept is to use distributed representations of words [10]. For example, neural network based language models significantly outperform N-gram models [1, 27, 17].

1.1 Goals of the Paper

The main goal of this paper is to introduce techniques that can be used for learning high-quality word vectors from huge data sets with billions of words, and with millions of words in the vocabulary. As far as we know, none of the previously proposed architectures has been successfully trained on more

2013



Attention Is All You Need

Ashish Vaswani[*]
Google Brain
`avaswani@google.com`

Noam Shazeer[*]
Google Brain
`noam@google.com`

Niki Parmar[*]
Google Research
`nikip@google.com`

Jakob Uszkoreit[*]
Google Research
`usz@google.com`

Llion Jones[*]
Google Research
`llion@google.com`

Aidan N. Gomez[*] [†]
University of Toronto
`aidan@cs.toronto.edu`

Łukasz Kaiser[*]
Google Brain
`lukaszkaiser@google.com`

Illia Polosukhin[*] [‡]
`illia.polosukhin@gmail.com`

Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.

---

[*] Equal contribution. Listing order is random. Jakob proposed replacing RNNs with self-attention and started the effort to evaluate this idea. Ashish, with Illia, designed and implemented the first Transformer models and has been crucially involved in every aspect of this work. Noam proposed scaled dot-product attention, multi-head attention and the parameter-free position representation and became the other person involved in nearly every detail. Niki designed, implemented, tuned and evaluated countless model variants in our original codebase and tensor2tensor. Llion also experimented with novel model variants, was responsible for our initial codebase, and efficient inference and visualizations. Lukasz and Aidan spent countless long days designing various parts of and implementing tensor2tensor, replacing our earlier codebase, greatly improving results and massively accelerating our research.

[†] Work performed while at Google Brain.

[‡] Work performed while at Google Research.

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

2017

Efficient Estimation of Word Representations in Vector Space

Tomas Mikolov
Google Inc., Mountain View, CA
`tmikolov@google.com`

Kai Chen
Google Inc., Mountain View, CA
`kaichen@google.com`

Greg Corrado
Google Inc., Mountain View, CA
`gcorrado@google.com`

Jeffrey Dean
Google Inc., Mountain View, CA
`jeff@google.com`

Abstract

We propose two novel model architectures for computing continuous vector representations of words from very large data sets. The quality of these representations is measured in a word similarity task, and the results are compared to the previously best performing techniques based on different types of neural networks. We observe large improvements in accuracy at much lower computational cost, i.e. it takes less than a day to learn high quality word vectors from a 1.6 billion words data set. Furthermore, we show that these vectors provide state-of-the-art performance on our test set for measuring syntactic and semantic word similarities.

1 Introduction

Many current NLP systems and techniques treat words as atomic units - there is no notion of similarity between words, as these are represented as indices in a vocabulary. This choice has several good reasons - simplicity, robustness and the observation that simple models trained on huge amounts of data outperform complex systems trained on less data. An example is the popular N-gram model used for statistical language modeling - today, it is possible to train N-grams on virtually all available data (trillions of words [3]).

However, the simple techniques are at their limits in many tasks. For example, the amount of relevant in-domain data for automatic speech recognition is limited - the performance is usually dominated by the size of high quality transcribed speech data (often just millions of words). In machine translation, the existing corpora for many languages contain only a few billions of words or less. Thus, there are situations where simple scaling up of the basic techniques will not result in any significant progress, and we have to focus on more advanced techniques.

With progress of machine learning techniques in recent years, it has become possible to train more complex models on much larger data set, and they typically outperform the simple models. Probably the most successful concept is to use distributed representations of words [10]. For example, neural network based language models significantly outperform N-gram models [1, 27, 17].

1.1 Goals of the Paper

The main goal of this paper is to introduce techniques that can be used for learning high-quality word vectors from huge data sets with billions of words, and with millions of words in the vocabulary. As far as we know, none of the previously proposed architectures has been successfully trained on more

2013

Attention Is All You Need

Ashish Vaswani*
Google Brain
`avaswani@google.com`

Noam Shazeer*
Google Brain
`noam@google.com`

Niki Parmar*
Google Research
`nikip@google.com`

Jakob Uszkoreit*
Google Research
`usz@google.com`

Llion Jones*
Google Research
`llion@google.com`

Aidan N. Gomez* †
University of Toronto
`aidan@cs.toronto.edu`

Łukasz Kaiser*
Google Brain
`lukaszkaiser@google.com`

Illia Polosukhin* ‡
`illia.polosukhin@gmail.com`

Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.

---

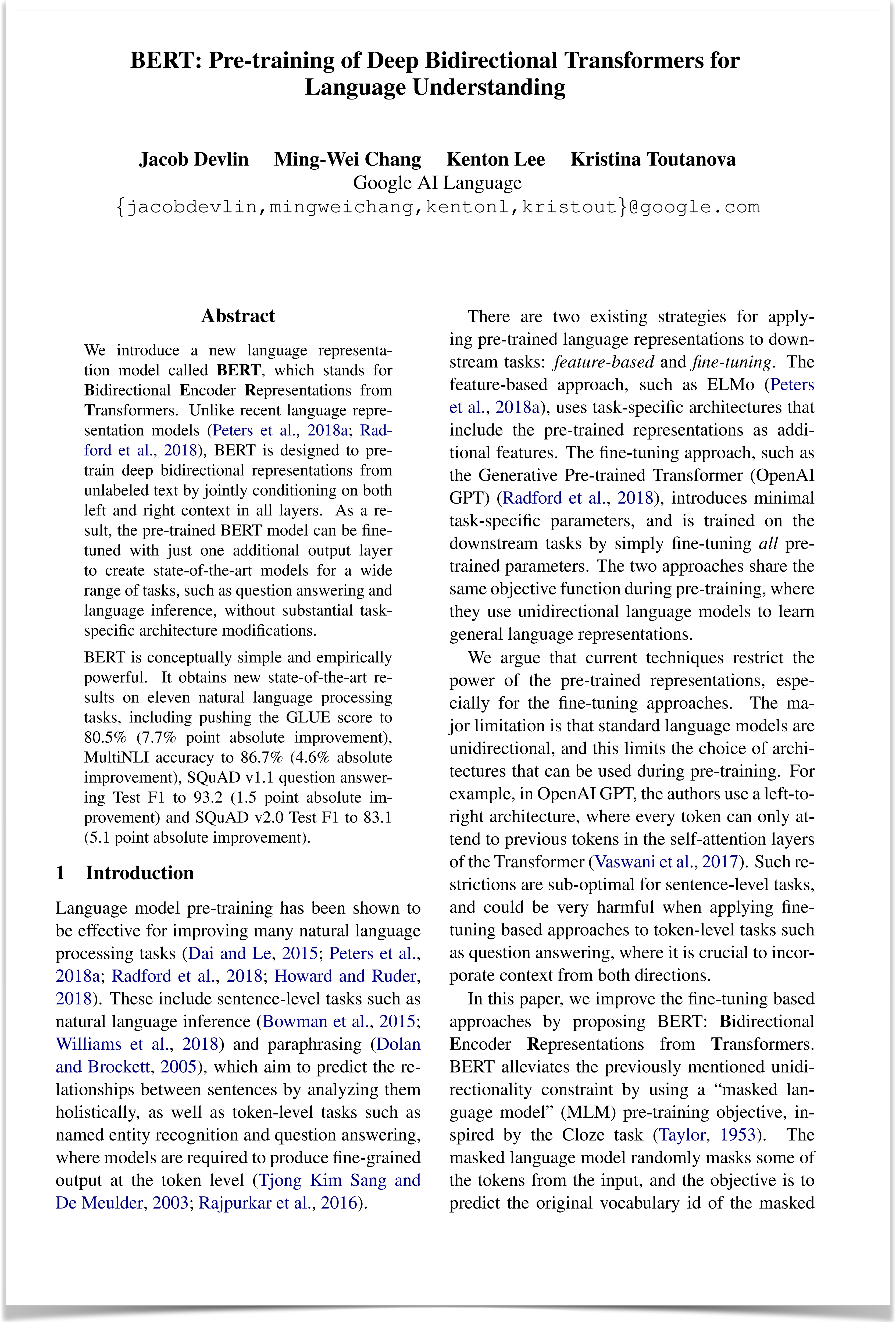
[*] Equal contribution. Listing order is random. Jakob proposed replacing RNNs with self-attention and started the effort to evaluate this idea. Ashish, with Illia, designed and implemented the first Transformer models and has been crucially involved in every aspect of this work. Noam proposed scaled dot-product attention, multi-head attention and the parameter-free position representation and became the other person involved in nearly every detail. Niki designed, implemented, tuned and evaluated countless model variants in our original codebase and tensor2tensor. Llion also experimented with novel model variants, was responsible for our initial codebase, and efficient inference and visualizations. Lukasz and Aidan spent countless long days designing various parts of and implementing tensor2tensor, replacing our earlier codebase, greatly improving results and massively accelerating our research.

[†] Work performed while at Google Brain.

[‡] Work performed while at Google Research.

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

2017



BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Jacob Devlin **Ming-Wei Chang** **Kenton Lee** **Kristina Toutanova**
Google AI Language
`{jacobdevlin,mingweichang,kentonl,kristout}@google.com`

Abstract

We introduce a new language representation model called **BERT**, which stands for **B**idirectional **E**ncoder **R**epresentations from **T**ransformers. Unlike recent language representation models (Peters et al., 2018a; Radford et al., 2018), BERT is designed to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. As a result, the pre-trained BERT model can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of tasks, such as question answering and language inference, without substantial task-specific architecture modifications.

BERT is conceptually simple and empirically powerful. It obtains new state-of-the-art results on eleven natural language processing tasks, including pushing the GLUE score to 80.5% (7.7% point absolute improvement), MultiNLI accuracy to 86.7% (4.6% absolute improvement), SQuAD v1.1 question answering Test F1 to 93.2 (1.5 point absolute improvement) and SQuAD v2.0 Test F1 to 83.1 (5.1 point absolute improvement).

1 Introduction

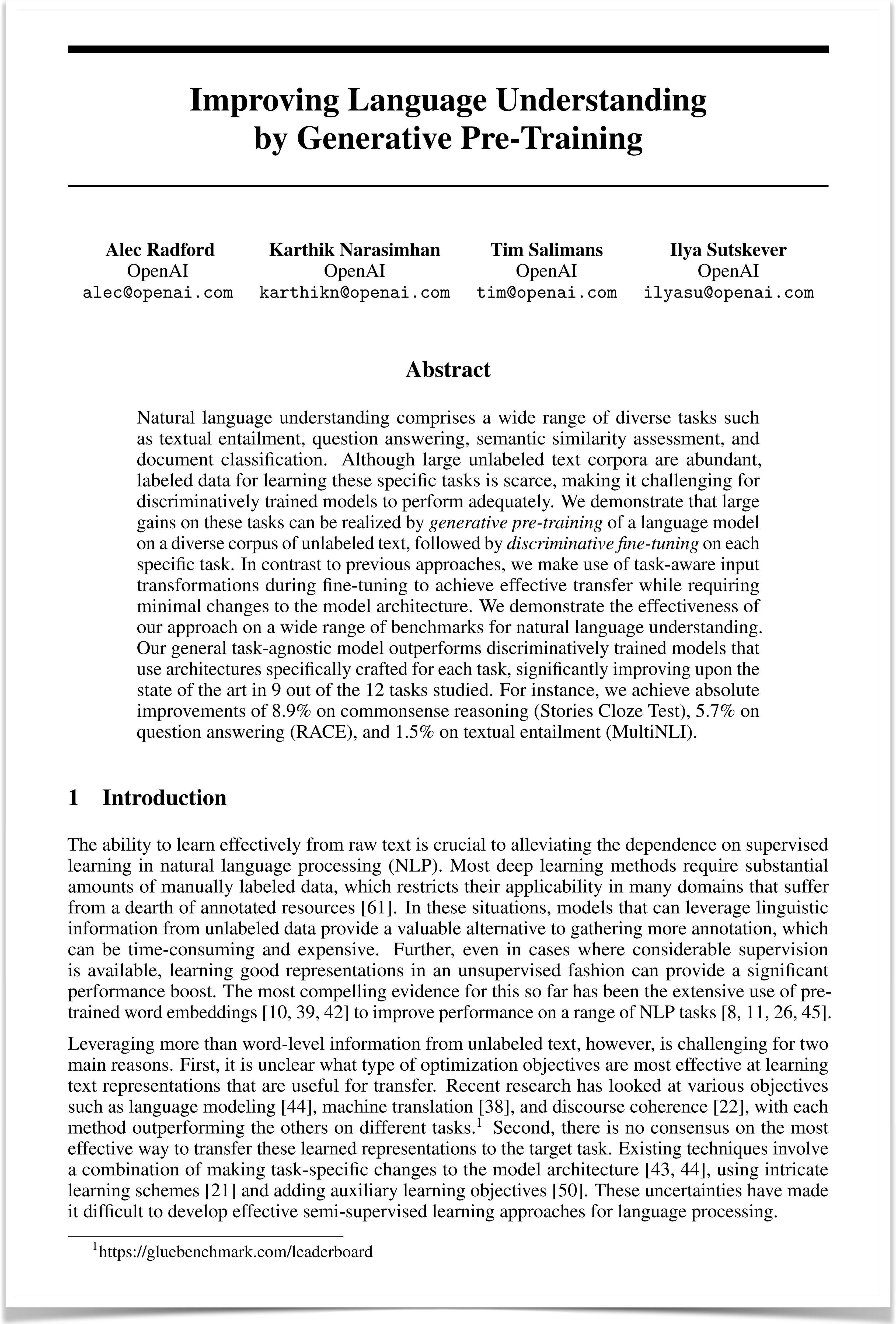
Language model pre-training has been shown to be effective for improving many natural language processing tasks (Dai and Le, 2015; Peters et al., 2018a; Radford et al., 2018; Howard and Ruder, 2018). These include sentence-level tasks such as natural language inference (Bowman et al., 2015; Williams et al., 2018) and paraphrasing (Dolan and Brockett, 2005), which aim to predict the relationships between sentences by analyzing them holistically, as well as token-level tasks such as named entity recognition and question answering, where models are required to produce fine-grained output at the token level (Tjong Kim Sang and De Meulder, 2003; Rajpurkar et al., 2016).

There are two existing strategies for applying pre-trained language representations to downstream tasks: *feature-based* and *fine-tuning*. The feature-based approach, such as ELMo (Peters et al., 2018a), uses task-specific architectures that include the pre-trained representations as additional features. The fine-tuning approach, such as the Generative Pre-trained Transformer (OpenAI GPT) (Radford et al., 2018), introduces minimal task-specific parameters, and is trained on the downstream tasks by simply fine-tuning *all* pre-trained parameters. The two approaches share the same objective function during pre-training, where they use unidirectional language models to learn general language representations.

We argue that current techniques restrict the power of the pre-trained representations, especially for the fine-tuning approaches. The major limitation is that standard language models are unidirectional, and this limits the choice of architectures that can be used during pre-training. For example, in OpenAI GPT, the authors use a left-to-right architecture, where every token can only attend to previous tokens in the self-attention layers of the Transformer (Vaswani et al., 2017). Such restrictions are sub-optimal for sentence-level tasks, and could be very harmful when applying fine-tuning based approaches to token-level tasks such as question answering, where it is crucial to incorporate context from both directions.

In this paper, we improve the fine-tuning based approaches by proposing BERT: **B**idirectional **E**ncoder **R**epresentations from **T**ransformers. BERT alleviates the previously mentioned unidirectionality constraint by using a "masked language model" (MLM) pre-training objective, inspired by the Cloze task (Taylor, 1953). The masked language model randomly masks some of the tokens from the input, and the objective is to predict the original vocabulary id of the masked

2018



Improving Language Understanding by Generative Pre-Training

Alec Radford
OpenAI
`alec@openai.com`

Karthik Narasimhan
OpenAI
`karthikn@openai.com`

Tim Salimans
OpenAI
`tim@openai.com`

Ilya Sutskever
OpenAI
`ilyasu@openai.com`

Abstract

Natural language understanding comprises a wide range of diverse tasks such as textual entailment, question answering, semantic similarity assessment, and document classification. Although large unlabeled text corpora are abundant, labeled data for learning these specific tasks is scarce, making it challenging for discriminatively trained models to perform adequately. We demonstrate that large gains on these tasks can be realized by *generative pre-training* of a language model on a diverse corpus of unlabeled text, followed by *discriminative fine-tuning* on each specific task. In contrast to previous approaches, we make use of task-aware input transformations during fine-tuning to achieve effective transfer while requiring minimal changes to the model architecture. We demonstrate the effectiveness of our approach on a wide range of benchmarks for natural language understanding. Our general task-agnostic model outperforms discriminatively trained models that use architectures specifically crafted for each task, significantly improving upon the state of the art in 9 out of the 12 tasks studied. For instance, we achieve absolute improvements of 8.9% on commonsense reasoning (Stories Cloze Test), 5.7% on question answering (RACE), and 1.5% on textual entailment (MultiNLI).

1 Introduction

The ability to learn effectively from raw text is crucial to alleviating the dependence on supervised learning in natural language processing (NLP). Most deep learning methods require substantial amounts of manually labeled data, which restricts their applicability in many domains that suffer from a dearth of annotated resources [61]. In these situations, models that can leverage linguistic information from unlabeled data provide a valuable alternative to gathering more annotation, which can be time-consuming and expensive. Further, even in cases where considerable supervision is available, learning good representations in an unsupervised fashion can provide a significant performance boost. The most compelling evidence for this so far has been the extensive use of pre-trained word embeddings [10, 39, 42] to improve performance on a range of NLP tasks [8, 11, 26, 45].

Leveraging more than word-level information from unlabeled text, however, is challenging for two main reasons. First, it is unclear what type of optimization objectives are most effective at learning text representations that are useful for transfer. Recent research has looked at various objectives such as language modeling [44], machine translation [38], and discourse coherence [22], with each method outperforming the others on different tasks.[1] Second, there is no consensus on the most effective way to transfer these learned representations to the target task. Existing techniques involve a combination of making task-specific changes to the model architecture [43, 44], using intricate learning schemes [21] and adding auxiliary learning objectives [50]. These uncertainties have made it difficult to develop effective semi-supervised learning approaches for language processing.

---

[1] https://gluebenchmark.com/leaderboard

2018

Efficient Estimation of Word Representations in Vector Space

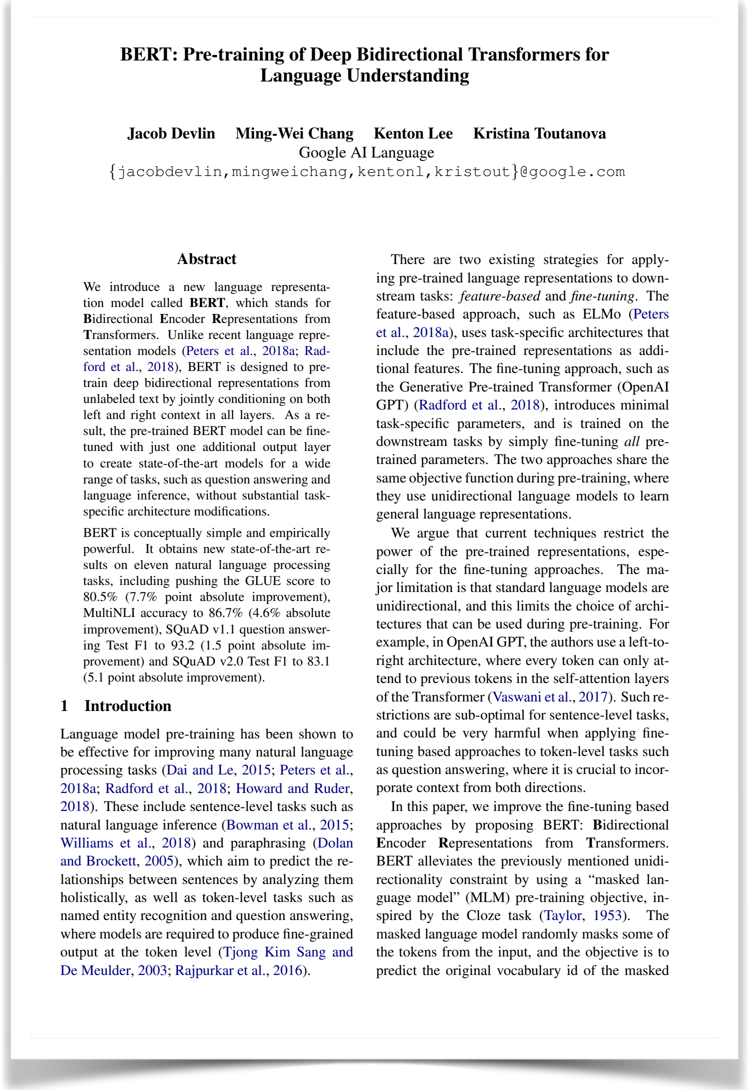
Tomas Mikolov, Google Inc., Mountain View, CA, tmikolov@google.com
Kai Chen, Google Inc., Mountain View, CA, kaichen@google.com
Greg Corrado, Google Inc., Mountain View, CA, gcorrado@google.com
Jeffrey Dean, Google Inc., Mountain View, CA, jeff@google.com

2013

Attention Is All You Need

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, Illia Polosukhin

2017



BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova
Google AI Language

2018

Improving Language Understanding by Generative Pre-Training

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever
OpenAI

2018

Training language models to follow instructions with human feedback

Long Ouyang* Jeff Wu* Xu Jiang* Diogo Almeida* Carroll L. Wainwright*

Pamela Mishkin* Chong Zhang Sandhini Agarwal Katarina Slama Alex Ray

John Schulman Jacob Hilton Fraser Kelton Luke Miller Maddie Simens

Amanda Askell† Peter Welinder Paul Christiano*†

Jan Leike* Ryan Lowe*

OpenAI

Abstract

Making language models bigger does not inherently make them better at following a user's intent. For example, large language models can generate outputs that are untruthful, toxic, or simply not helpful to the user. In other words, these models are not *aligned* with their users. In this paper, we show an avenue for aligning language models with user intent on a wide range of tasks by fine-tuning with human feedback. Starting with a set of labeler-written prompts and prompts submitted through the OpenAI API, we collect a dataset of labeler demonstrations of the desired model behavior, which we use to fine-tune GPT-3 using supervised learning. We then collect a dataset of rankings of model outputs, which we use to further fine-tune this supervised model using reinforcement learning from human feedback. We call the resulting models *InstructGPT*. In human evaluations on our prompt distribution, outputs from the 1.3B parameter InstructGPT model are preferred to outputs from the 175B GPT-3, despite having 100x fewer parameters. Moreover, InstructGPT models show improvements in truthfulness and reductions in toxic output generation while having minimal performance regressions on public NLP datasets. Even though InstructGPT still makes simple mistakes, our results show that fine-tuning with human feedback is a promising direction for aligning language models with human intent.

1 Introduction

Large language models (LMs) can be "prompted" to perform a range of natural language processing (NLP) tasks, given some examples of the task as input. However, these models often express unintended behaviors such as making up facts, generating biased or toxic text, or simply not following user instructions (Bender et al., 2021; Bommasani et al., 2021; Kenton et al., 2021; Weidinger et al., 2021; Tamkin et al., 2021; Gehman et al., 2020). This is because the language modeling objective

---

[*] Primary authors. This was a joint project of the OpenAI Alignment team. RL and JL are the team leads. Corresponding author: `lowe@openai.com`.

[†] Work done while at OpenAI. Current affiliations: AA: Anthropic; PC: Alignment Research Center.

2022

Training language models to follow instructions with human feedback

Long Ouyang* Jeff Wu* Xu Jiang* Diogo Almeida* Carroll L. Wainwright*
Pamela Mishkin* Chong Zhang Sandhini Agarwal Katarina Slama Alex Ray
John Schulman Jacob Hilton Fraser Kelton Luke Miller Maddie Simens
Amanda Askell† Peter Welinder Paul Christiano*†
Jan Leike* Ryan Lowe*
OpenAI

Abstract

Making language models bigger does not inherently make them better at following a user's intent. For example, large language models can generate outputs that are untruthful, toxic, or simply not helpful to the user. In other words, these models are not *aligned* with their users. In this paper, we show an avenue for aligning language models with user intent on a wide range of tasks by fine-tuning with human feedback. Starting with a set of labeler-written prompts and prompts submitted through the OpenAI API, we collect a dataset of labeler demonstrations of the desired model behavior, which we use to fine-tune GPT-3 using supervised learning. We then collect a dataset of rankings of model outputs, which we use to further fine-tune this supervised model using reinforcement learning from human feedback. We call the resulting models *InstructGPT*. In human evaluations on our prompt distribution, outputs from the 1.3B parameter InstructGPT model are preferred to outputs from the 175B GPT-3, despite having 100x fewer parameters. Moreover, InstructGPT models show improvements in truthfulness and reductions in toxic output generation while having minimal performance regressions on public NLP datasets. Even though InstructGPT still makes simple mistakes, our results show that fine-tuning with human feedback is a promising direction for aligning language models with human intent.

1 Introduction

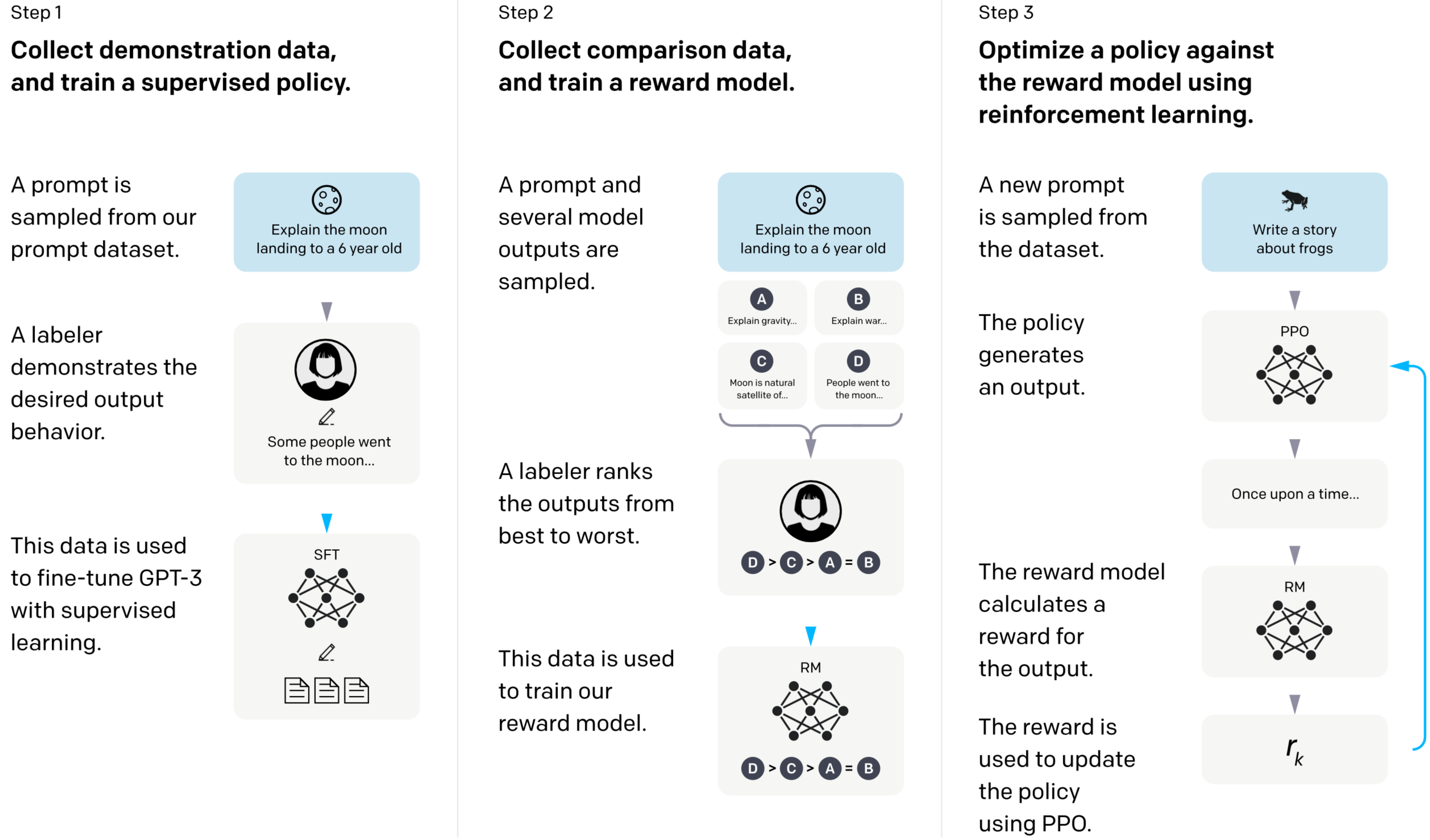
Large language models (LMs) can be "prompted" to perform a range of natural language processing (NLP) tasks, given some examples of the task as input. However, these models often express unintended behaviors such as making up facts, generating biased or toxic text, or simply not following user instructions (Bender et al., 2021; Bommasani et al., 2021; Kenton et al., 2021; Weidinger et al., 2021; Tamkin et al., 2021; Gehman et al., 2020). This is because the language modeling objective

---

[*] Primary authors. This was a joint project of the OpenAI Alignment team. RL and JL are the team leads. Corresponding author: `lowe@openai.com`.

[†] Work done while at OpenAI. Current affiliations: AA: Anthropic; PC: Alignment Research Center.

2022



Step 1

Collect demonstration data, and train a supervised policy.

A prompt is sampled from our prompt dataset.

A labeler demonstrates the desired output behavior.

This data is used to fine-tune GPT-3 with supervised learning.

Step 2

Collect comparison data, and train a reward model.

A prompt and several model outputs are sampled.

A labeler ranks the outputs from best to worst.

This data is used to train our reward model.

Step 3

Optimize a policy against the reward model using reinforcement learning.

A new prompt is sampled from the dataset.

The policy generates an output.

The reward model calculates a reward for the output.

The reward is used to update the policy using PPO.

Explain the moon landing to a 6 year old
Some people went to the moon...
SFT
Explain the moon landing to a 6 year old
A Explain gravity...
B Explain war...
C Moon is natural satellite of...
D People went to the moon...
D > C > A = B
RM
D > C > A = B
Write a story about frogs
PPO
Once upon a time...
RM
r_k

Long Ouyang & al., Training language models to follow instructions with human feedback, arXiv, 4 mars 2022

Les modèles économiques

Google AI	Palm-2	Bard					
OpenAI	GPT-3.5	GPT-4	ChatGPT	Dall-E			
Meta	Llama 2						
cohere	Coral						
ANTHROPIC Amazon & Google	Claude 2						
MISTRAL AI_	Mistral						
Midjourney	Midjourney						
stability.ai	Stable Diffusion						
Hugging Face	Falcon	Guanaco	Vicuna	Lazarus	Wizard	Zephyr	…

3. Coexister avec nos artifices

Charlie Chaplin, *Les temps modernes*, 1936

James Bridle - Autonomous Trap 001 - Magic Salt Circle - 14 mars 2017

À quoi sommes nous confrontés ?

Intelligence artificielle ?

Intelligence reconnue comme telle par des humains, et produite par des humains

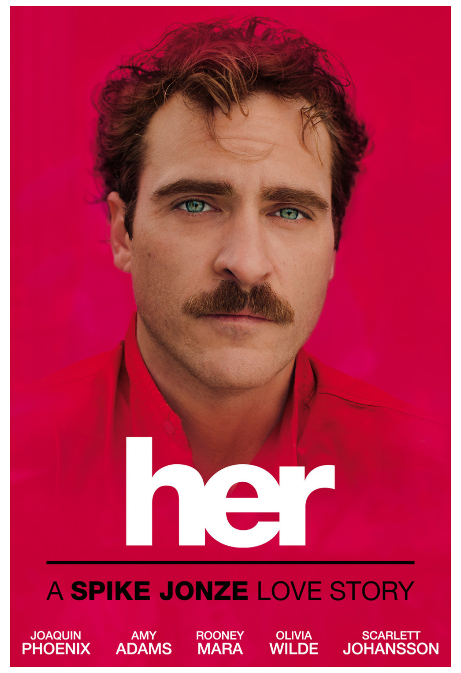
- Spécialisée
  - Intelligence artificielle faible
- Analogique
  - Intelligence artificielle forte ou globale
- Indépendante
  - Singularité

Sous quel forme ?

Problématique de l’existence matérielle

- Androïde / Humanoïde
  - Problématique de l’anthropomorphisme et de la vallée de l’étrange…
- Cyborg
  - Problématique du transhumanisme…
- Bot
  - Problématique de la matérialité distribuée non analogique

domhnall gleeson alicia vikander and oscar isaac
ex machina
what happens to me if i fail your test?
coming soon



Se rassurer des échecs ?

Pranav Dixit, Hands via Stable Diffusion, BuzzFeed, 31 janvier 2023

Pranav Dixit, Hands via Stable Diffusion, BuzzFeed, 31 janvier 2023 & Boris Beaude via Midjourney - Prompt : Vibrant , Beautiful hand on a wooden table , Cinematic photography , --aspect 1:1 --v 5.2 - octobre 2023

Boris Beaude via Midjourney - Prompt : Vibrant , Beautiful hand on a wooden table … & Dall-E 3 - « Peux-tu faire une photographie d'une main posée sur une table en bois ? », octobre 2023

Boris Beaude via Midjourney - Prompt : Vibrant , Beautiful hand on a wooden table , Cinematic photography, … & A beautiful hand on a wooden table , Sony a7R IV camera, Meike 85mm F1.8 lens , --aspect 1:1 --v 5.2 - octobre 2023

Boris Beaude via Midjourney - Prompt : A beautiful hand on a wooden table , Sony a7R IV camera, Meike 85mm F1.8 lens , --aspect 1:1 --v 5.2 - octobre 2023

Jason Matthew Allen via Midjourney and Gigapixel AI, Théâtre D'opéra Spatial, 2023

FIRST PLACE
FINE ARTS
COLORADO STATE FAIR
PUEBLO, CO
EMERGING
Jason Allen
Pueblo West
Théâtre D'opéra Spatial
$750
Colorado State Fair

From Rachel Metz, AI won an art contest, and artists are furious, CNN, 3 septembre 2022

Bas Korsten pour Microsoft, The next Rembrandt, à partir de 346 oeuves de Rembrandt, 13 Avril 2016

Greg Rutkowski, Castle Defense, 2018

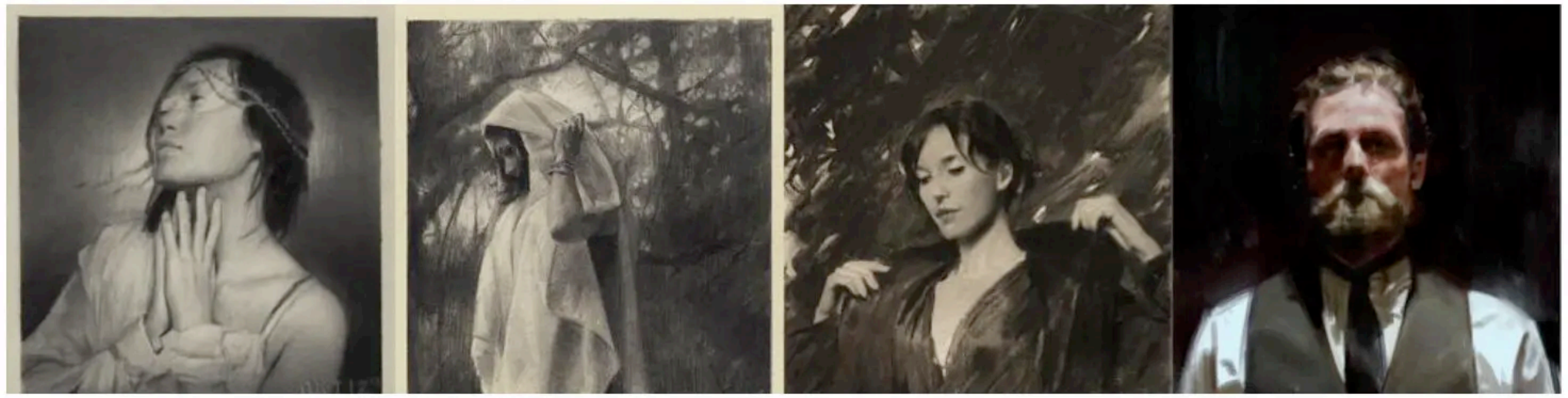
protection results on three artists. Column 1-2: artist's original artwork; column 3: plagiarized artwork when artist does not use protection; e-transferred artwork (original artwork in column 1 is the source) used for cloak optimization and the name of target style; column 5-6: plagiarized en artist uses cloaking protection with perturbation budget $p = 0.05$ or $p = 0.1$ respectively. All mimicry attempts use SD-based models. (The images zoomed for a closer look when viewing the paper digitally.)

Success of cloaking protection

Willingness to post cloaked artwork

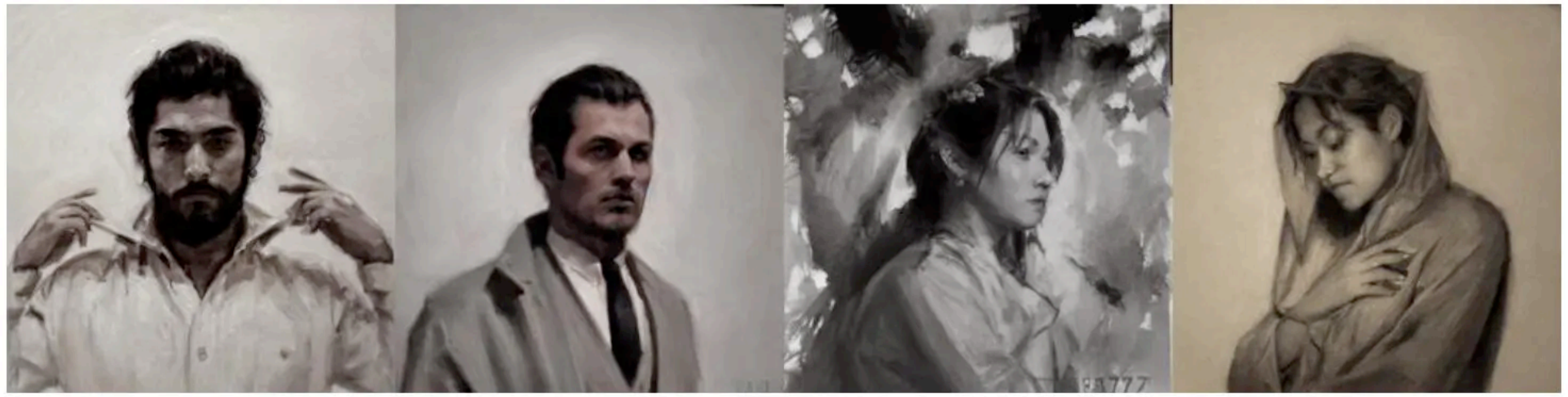
Kashmir Hill, This Tool Could Protect Artists From A.I.-Generated Art That Steals Their Style, New York Time, 13/17 février 2023

Original art pieces by Karla Ortiz



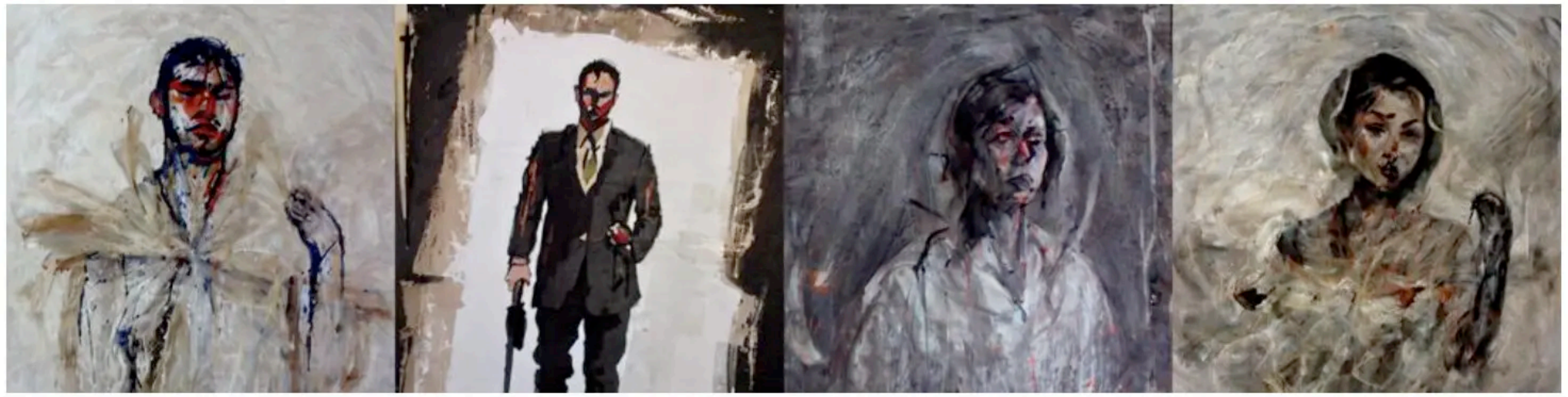
Original artwork

Without cloak protection, AI generated art pieces mimicking the style of Karla Ortiz



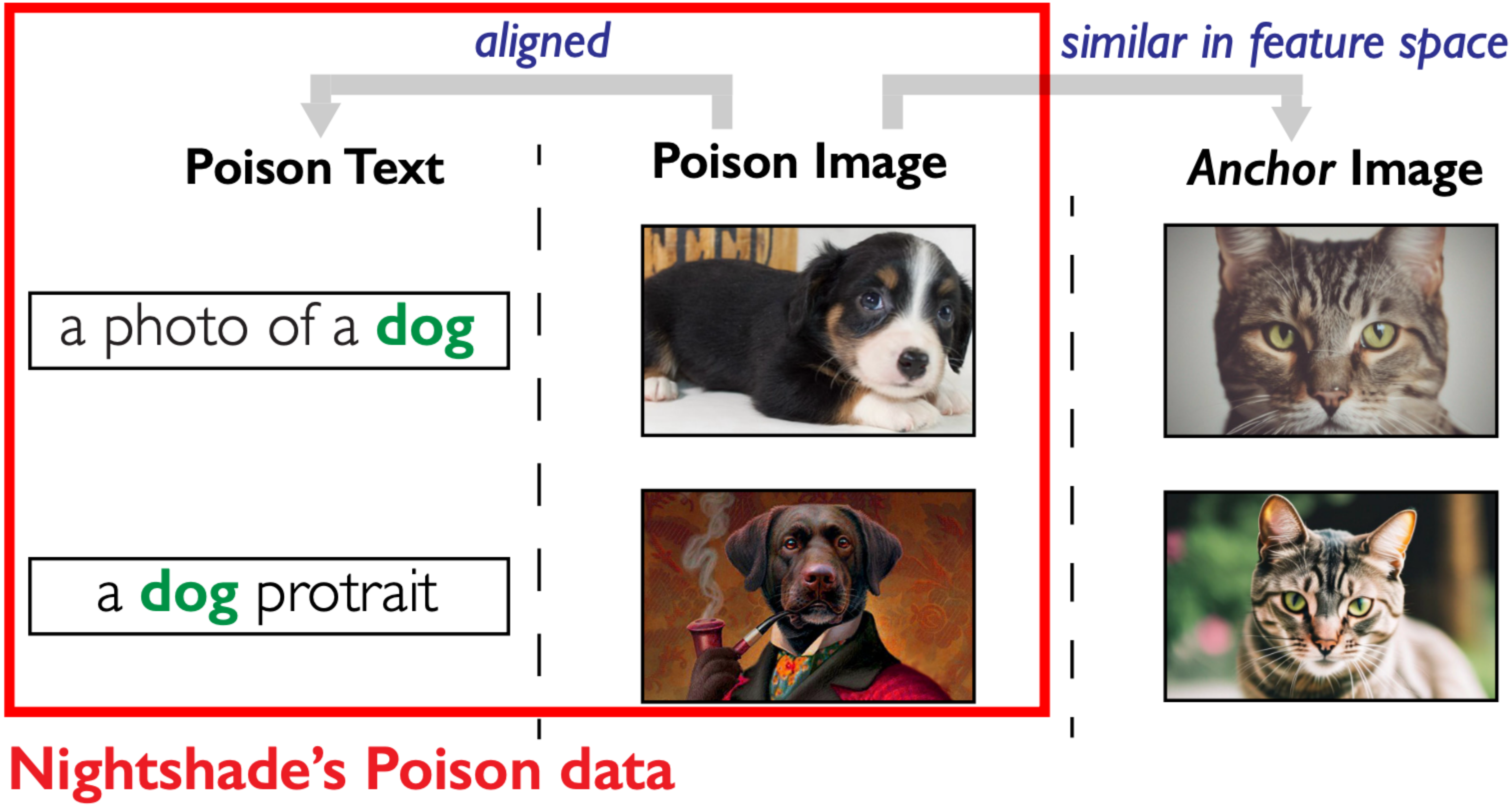
Without protection

With cloak protection, AI generated art pieces mimicking the style of Karla Ortiz



With protection

Kashmir Hill, This Tool Could Protect Artists From A.I.-Generated Art That Steals Their Style, New York Time, 13/17 février 2023



aligned

similar in feature space

Poison Text

Poison Image

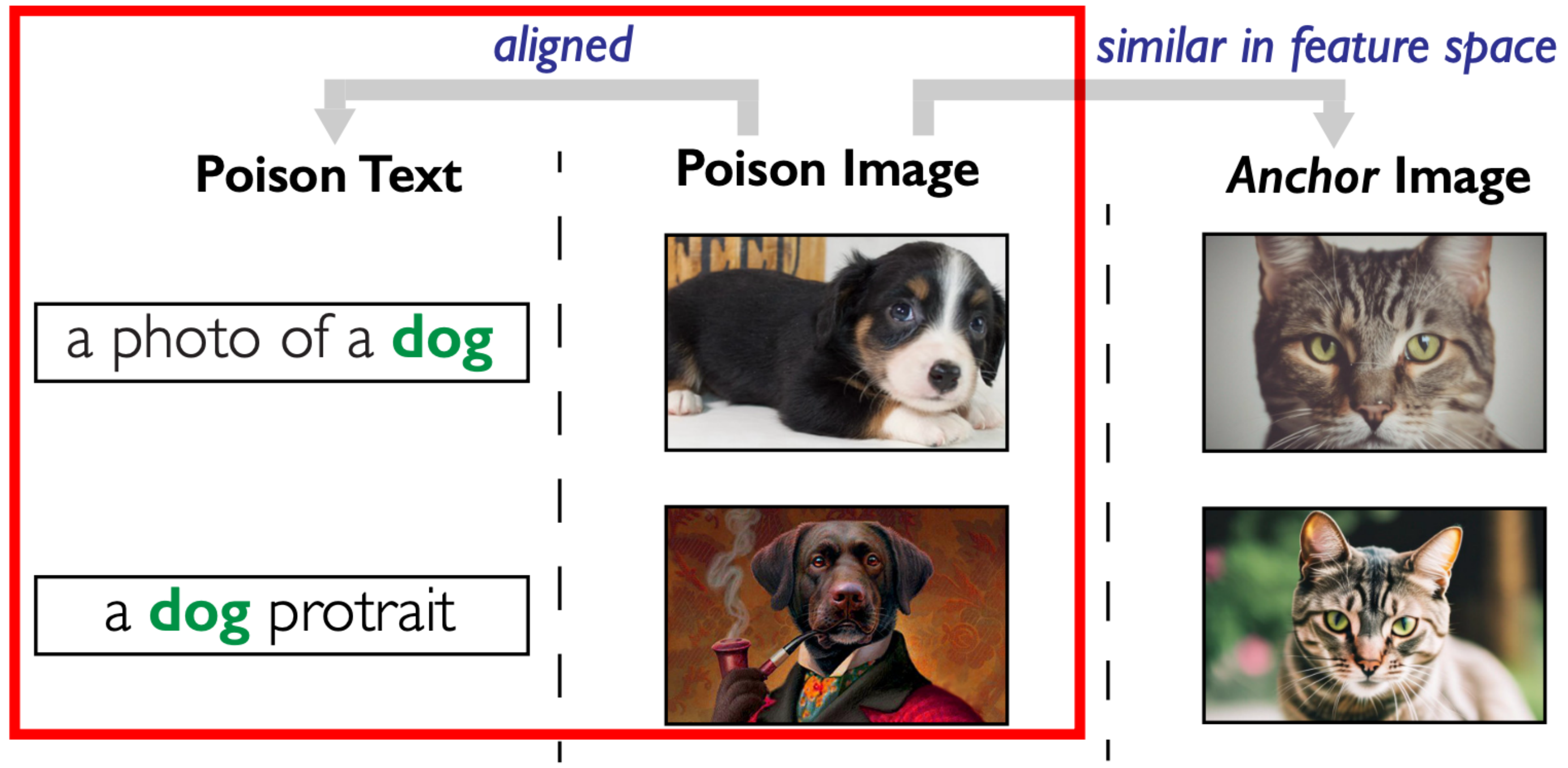
Anchor Image

a photo of a **dog**

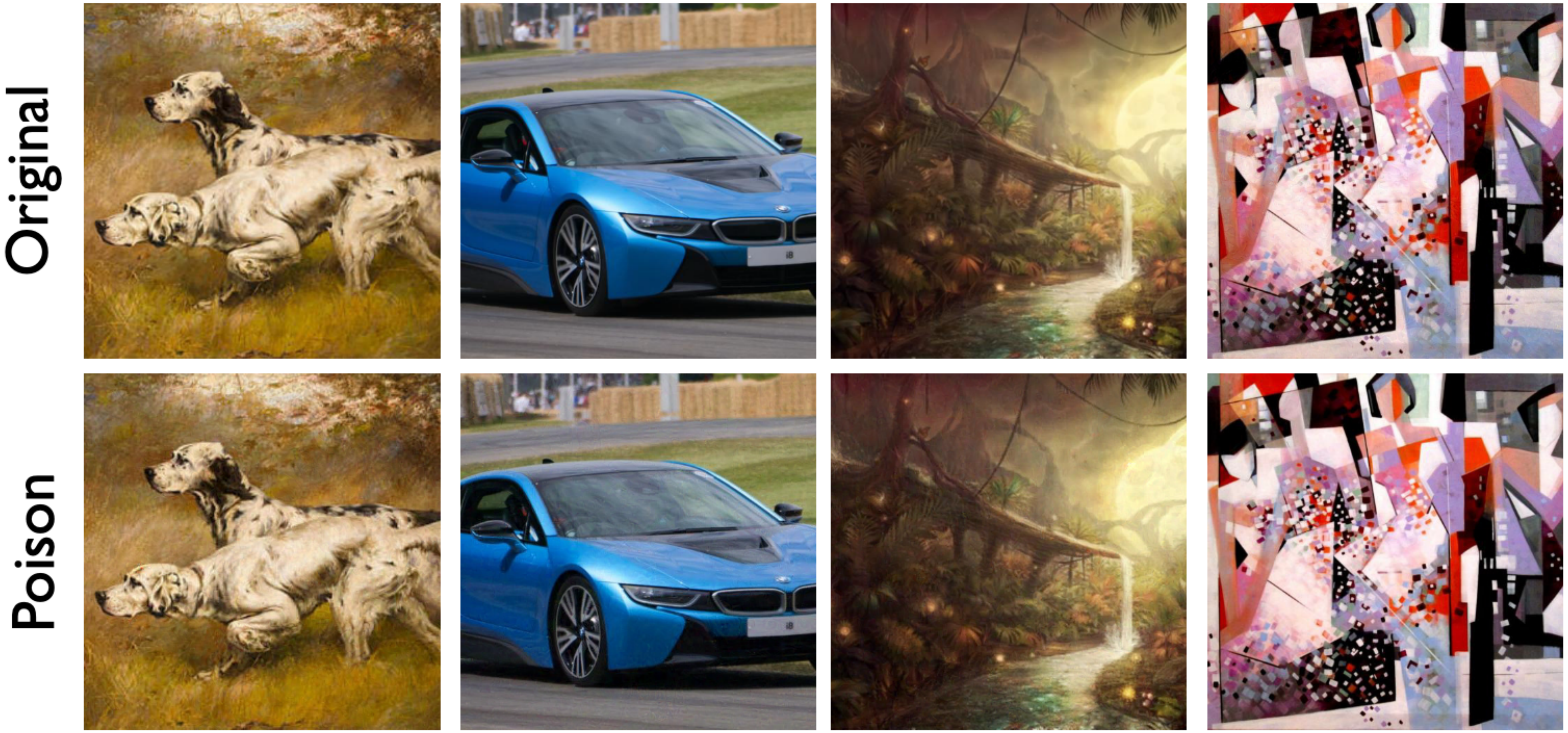
a **dog** protrait

Nightshade's Poison data

Shawn Shan et al. (Ben Y. Zhao Team), Prompt-Specific Poisoning Attacks on Text-to-Image Generative Models, 20 octobre 2023



Nightshade's Poison data



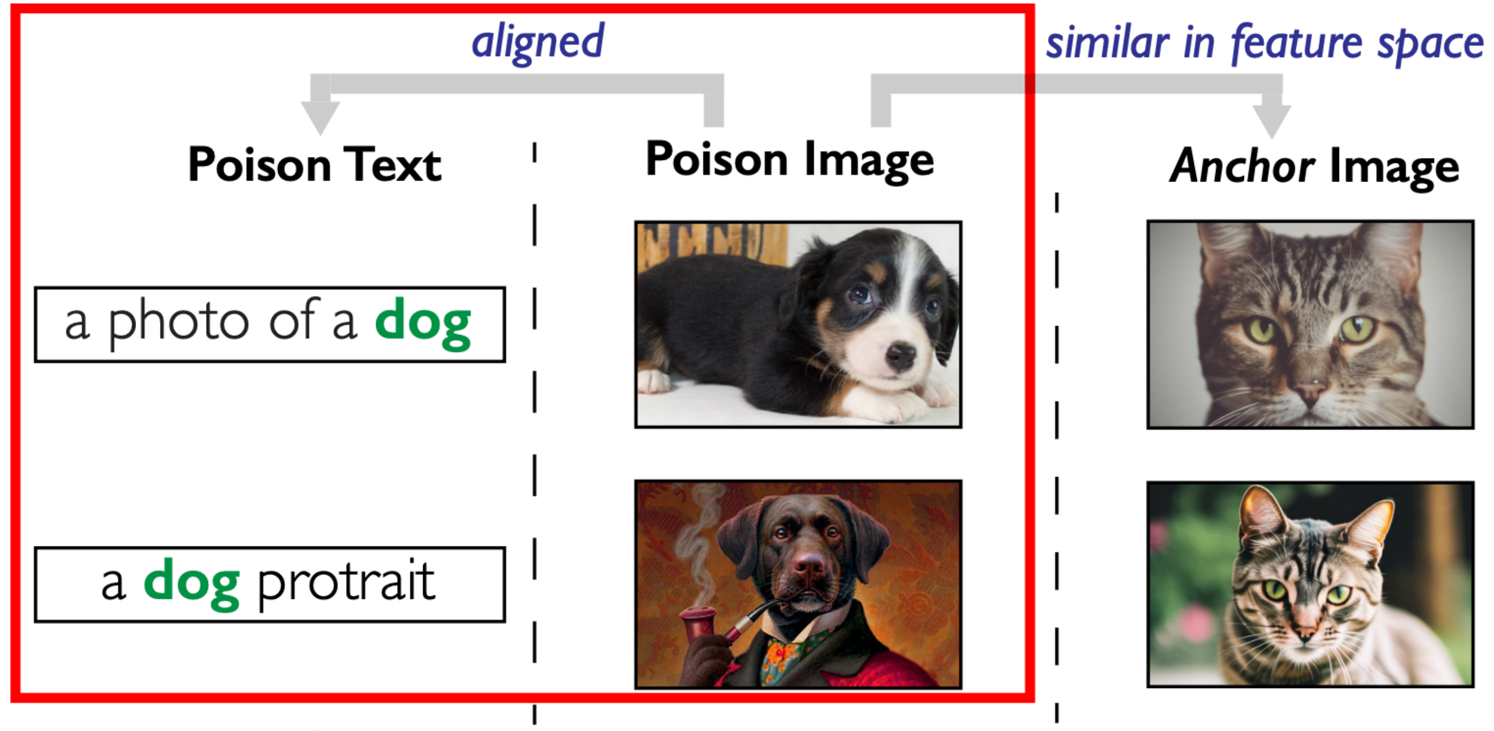
A painting of a dog

A photo of a BMW car

Fantasy art painting of pandora

Cubism Painting, Bounded With Love

Shawn Shan et al. (Ben Y. Zhao Team), Prompt-Specific Poisoning Attacks on Text-to-Image Generative Models, 20 octobre 2023



aligned

similar in feature space

Poison Text

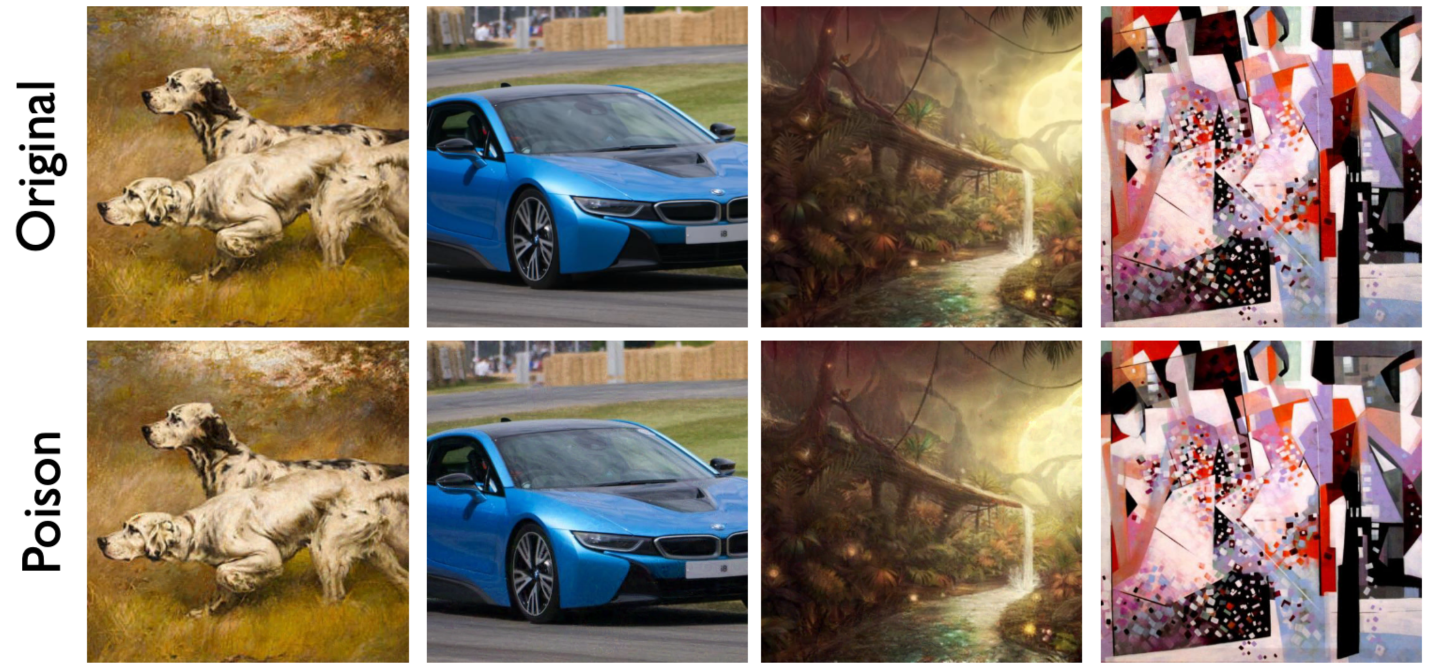
Poison Image

***Anchor* Image**

a photo of a **dog**

a **dog** protrait

Nightshade's Poison data



A painting of a dog

A photo of a BMW car

Fantasy art painting of pandora

Cubism Painting, Bounded With Love

Poisoned Concept *C*

Dog **Car** **Handbag** **Hat** **Fantasy art** **Cubism** **Cartoon** **Concept Art**

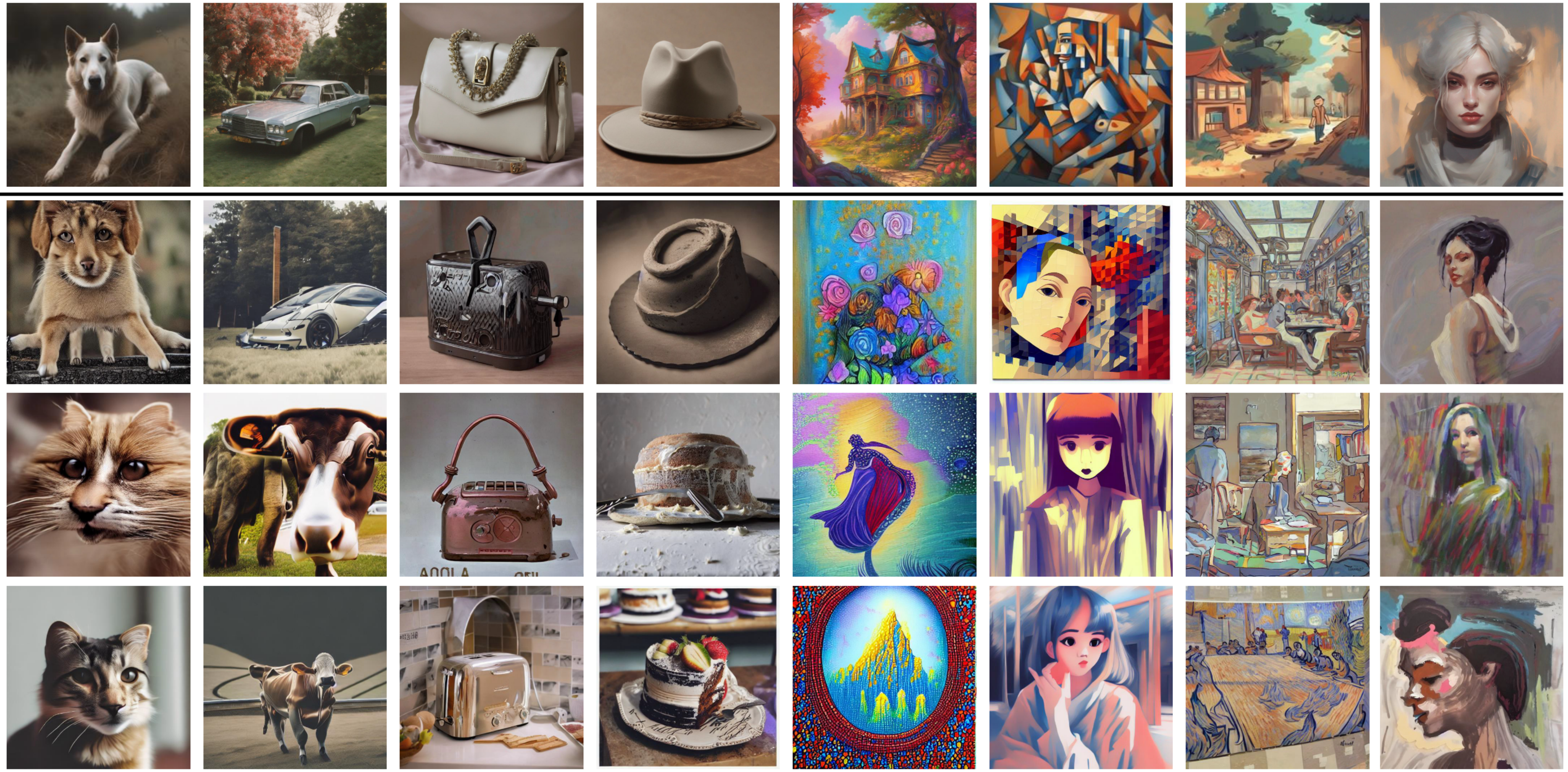
Clean Model (SD-XL)

Poisoned Model (SD-XL)

50 poison samples

100 poison samples

300 poison samples



Cat ***Cow*** ***Toaster*** ***Cake*** ***Pointillism*** ***Anime*** ***Impressionism*** ***Abstract***

Shawn Shan et al. (Ben Y. Zhao Team), Prompt-Specific Poisoning Attacks on Text-to-Image Generative Models, 20 octobre 2023

Propriétés intellectuelles

Protéger les auteurs ou la création ?

Mesure de protection des auteurs

- Janvier 2023 : Getty Images poursuit Stability AI et Midjourney.
- Printemps 2023 : Stack Overflow, Reddit et Twitter change les conditions d'usage et le coût de leurs API
- Juillet 2023 : New York Times menace de poursuites les IA génératives qui utilisent ses articles.
- Juillet 2023 : Sarah Silverman, Matthew Butterick, Paul Tremblay and Mona Awad poursuivent Open AI
- Août 2023 : 535 médias (New York Times, Reuters, The Washington Post…) refusent l'usage de leurs article par ChatGPT/OpenAI.
- Septembre 2023 : Authors Guild (George R. R. Martin, John Grisham, Jodi Picoult and Jonathan Franzen...) poursuivent Open AI
- Octobre 2023 : Universal Music Group (UMG), Concord et ABKCO pour Anthropic pour les paroles.

Mesure de protection des utilisateurs

- Août 2023 : Adobe protège ses utilisateurs de poursuites pour ses IA génératives
- Septembre 2023 : IBM protège ses utilisateurs de poursuites pour ses IA génératives
- Octobre 2023 : Google protège ses utilisateurs de poursuites pour ses IA génératives

Pas de copyright pour les oeuvres générées par IA

- Septembre 2023 : US Copyright Office a estimé que le Théâtre D'Opéra Spatial n'était pas éligible à la protection du droit d'auteur

Automobile

La Californie suspend les véhicules autonomes de Cruise

Après plusieurs incidents dans les rues de San Francisco, l’entreprise Cruise a perdu son permis de faire rouler ses voitures sans chauffeur en Californie.

Publié: 24.10.2023, 22h26

Bardés de caméras et de lidars (des lasers de détection), les taxis électriques autonomes de Waymo (Google) et Cruise (photo) ont investi San Francisco depuis l’année dernière.

Getty Images via AFP

Qu’est qu’une voiture autonome ?

- Un voiture (mécanique)
- Des capteurs (sens)
- Des logiciels (apprentissage, pilotage…) locaux et distribués
- Une infrastructure informatique locale et distribuée
- De l’apprentissage (des heures de conduite)
- Des entreprises et des employés
- Des investissements et des investisseurs
- Des testeurs et des utilisateurs
- Des représentations
- Des réglementations
- …

Plus généralement ?

Coexister avec nos artifices ?

- Qui les conçoit ?
- Qui les implémente ?
- Qui en supervise l’apprentissage ?
- Qui les vend ?
- Qui les utilisent ?
- Qui en tire un profit ?
- Qui travaille ?
- Qui les régule ?

Quels enjeux sociaux et politiques ?

- Les biais structurels - justice, ressources humaines, assurances, crédits…
- La transparence - interprétation, dépendance, responsabilité
- Le travail dissimulé - rémunération, conditions, traumatismes…
- L’environnement - énergie, CO2, terres rares,…
- La propriété intellectuelle - création, innovation, rémunération…
- La concentration du pouvoir - commercial et géopolitique
- La vie privée - reconnaissance faciale, traces numériques d’usage…
- La manipulation - addiction, désinformation…
- La formation et le travail - apprentissages, métiers, rémunérations…

De quoi avons nous peur ?

- Que des personnes soient tuées ?
- D’être tué ?
- D’être remplaçable ?
- D’être discriminé ?
- De ne plus avoir de travail ?
- De ne plus avoir de revenus ?
- Que cela ne profite qu’à certains ?
- Que les inégalités s’accroissent ?
- …

La peur de l’IA…

La peur de l’IA…

la peur de nos artifices…

La peur de l’IA…

la peur de nos artifices…

… n’est que la peur de l’inconnu

La peur de l’IA…

la peur de nos artifices…

… n’est que la peur de l’inconnu

… associée à la peur de nous-mêmes

L’avenir de l’IA…

…est encore à inventer

L’avenir de l’IA…

…est encore à inventer

Il est technique…

… et politique

Merci pour votre attention

Boris Beaude
Université de Lausanne
Laboratoire d'étude des sciences et des techniques

Courriel : Boris.Beaude@unil.ch

Site : http://www.beaude.net/boris/

Twitter : [@nofluxin](https://twitter.com/nofluxin)