IA

& Sociétés

Fondements
et implications sociétales

Pr. Boris Beaude · Formation continue UNIL-EPFL

Unil.



IA

Usages et implications sociétales

Fondements, pratiques, enjeux et perspectives critiques

Pr. Boris Beaude · Formation continue UNIL-EPFL

Unil.



Image of a goldfish inside a "Cola-Coca" bottle on a wooden table.

Fév. 2026

IA

Usages et implications sociétales

Fondements, pratiques, enjeux et perspectives critiques

Pr. Boris Beaude · Formation continue UNIL-EPFL

Unil.

Boris Beaude - 15 septembre 2026



Déc. 2024

IA

Usages et implications sociétales

Fondements, pratiques, enjeux et perspectives critiques

Pr. Boris Beaude · Formation continue UNIL-EPFL

Unil.



Fév. 2026

IA

Usages et implications sociétales

Fondements, pratiques, enjeux et perspectives critiques

Pr. Boris Beaude · Formation continue UNIL-EPFL

Unil.

 CC BY NC SA

RESEARCH ARTICLE SUMMARY

SYNTHETIC BIOLOGY

Generative design of bacteriophages with genome language models

Samuel H. King, Claudia L. Driscoll, David B. Li, Daniel Guo, Aditi T. Merchant, Garyk Brixi, Max E. Wilkinson, Brian L. Hie*

Full article and list of author affiliations: https://doi.org/10.1126/science.aec2657

INTRODUCTION: Evolution continuously forges new biological innovations written in genomes. Navigating this vast design space could access functions that would transform biotechnology, but even the simplest genomes are highly complex and can be rendered nonviable by a single mutation. Accordingly, most progress in biological design has been made at the scale of individual genes and gene circuits, whereas design at the scale of whole genomes has remained largely beyond reach.

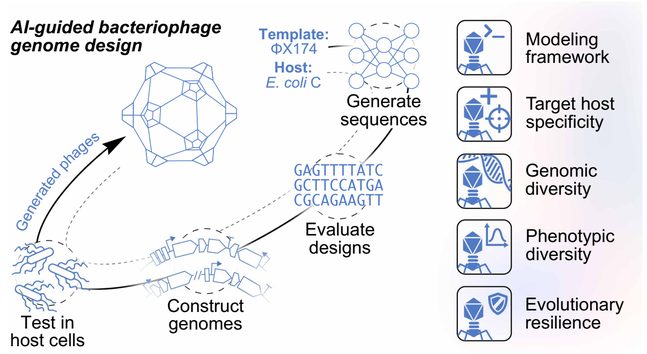
RATIONALE: Genome language models are artificial intelligence (AI) algorithms that have shown promise in designing biological systems. Much like how other language models are trained on large corpora of text, genome language models are trained on large corpora of DNA comprising millions of genomes from all domains of life. This enables these models to learn the evolutionary constraints that shape DNA sequences in nature. However, the ability of genome language models to generate entire functional genomes has not been tested. Bacteriophages, viruses that infect bacteria, are specifically well suited for this task, as they are relatively small, experimentally tractable, and have broad applications in molecular biology, microbial engineering, and therapeutics.

RESULTS: In this work, we leveraged genome language models, Evo 1 and Evo 2, to generate complete phage genomes with realistic genetic architectures and specificity for a bacterial host, *Escherichia coli* C. Using the natural phage ΦX174 as a design template, we established a framework for generating and evaluating thousands of AI-generated genomes, nearly 300 of which we chemically synthesized and tested in laboratory conditions, yielding 16 viable phages. The viable generated phages showed strong host specificity and diverse fitness profiles, including competitive infection kinetics. The generated phages were different from any known natural phages, exhibiting de novo mutations, divergent genes and regulatory elements, and variable genome lengths. One of the phages utilized a DNA packaging protein from an evolutionarily distant phage in its capsid structure. We also tested whether the generated phages could overcome bacterial resistance, a central challenge in developing phage-based antimicrobial therapies, and found that a mixture of designed phages rapidly overcame ΦX174-resistant *E. coli* strains, whereas a comparable mixture of naturally sourced ΦX174-like phages could not.

CONCLUSION: Our results demonstrate that generative models capture evolutionary constraints in DNA sequences with enough fidelity to produce complete bacteriophage genomes divergent from those observed in nature and with prespecified traits. Our approach expands what synthetic genomics can achieve alongside methods such as directed evolution and rational engineering, lays out a path for generating adaptive and resilient phage therapies against rapidly evolving pathogens, and establishes a foundation for the generative design of larger, more complex genomes. Genome design can augment the broader toolkit of genome sequencing, synthesis, and editing, enabling the composition of biological systems at the genome scale. □

*Corresponding author. Email: brianhie@stanford.edu Cite this article as S. H. King *et al.*, *Science* **393**, eaec2657 (2026). DOI: 10.1126/science.aec2657

Downloaded from https://www.science.org at University of Lausanne on August 11, 2026



A framework for AI-guided bacteriophage genome design. Language models trained on millions of genomes enable the generation of complete phages. Sequences are generated and computationally evaluated with design criteria inspired from the phage ΦX174 and its host *E. coli* C. The most promising designs are chemically synthesized and tested in host cells, yielding viable generated phage genomes divergent from known natural phages, while retaining host specificity.

Préalable

Nous n'avons pas attendu l'IA, pour…

- nous entretuer massivement,
- déshumaniser nos semblables,
- détruire notre environnement,
- maltraiter les enfants.
- découvrir les antibiotiques,
- inventer les droits humains,
- réparer la couche d'ozone,
- rendre l'école obligatoire,

et manipuler les imaginaires ou créer des sushis

Unil.

Mais l'IA change l'échelle de nos actions …

Idées reçues ?

Les IA générative :

... ne sont que des statistiques

... ne sont que des probabilités

... sont des perroquets stochastiques

... sont des boîtes noires

... ne sont pas intelligentes

... sont incapables de créer

... ne sont pas artificielles

... sont biaisées

... discriminent

... sont « woke »

... pensent en anglais et traduisent

... sont moins performantes en français

... uniformisent le langage

... consomment 10x plus qu'une recherche sur Google

... sont une bulle spéculative

... sont de la hype

... ne sont pas les plus intéressantes

... ne sont pas dangereuses

... ne fonctionne pas sans travail humain caché

... n'augmentent pas la productivité

... produit une dette cognitive...

Programme

0. Vous n’avez pas rêvé

1. Que s’est-il passé ?

2. Etudier l’IA

3. Et si ça marche…

0. Vous n'avez pas rêvé

C'était mauvais. Et c'était survendu …







Boris Beaude - 15 septembre 2026





Unil.

La critique de l'IA était juste, mais l'IA a changé !



Unil.

La critique de l'IA était juste, mais l'IA a changé !

Boris Beaude - 15 septembre 2026 CC BY NC SA



Un chat errant dans une ruelle new-yorkaise, au lever du jour, à l'affût d'une souris.

Une bouche d'égout laisse s'échapper de la vapeur.

Midjourney - Style personnel - 27 décembre 2025

Unil.

La critique de l'IA était juste, mais l'IA a changé !

Boris Beaude - 15 septembre 2026 CC BY NC SA

Les échecs récents ne se révisent pas facilement.

Résoudre cette dissonance exige de prendre du recul …

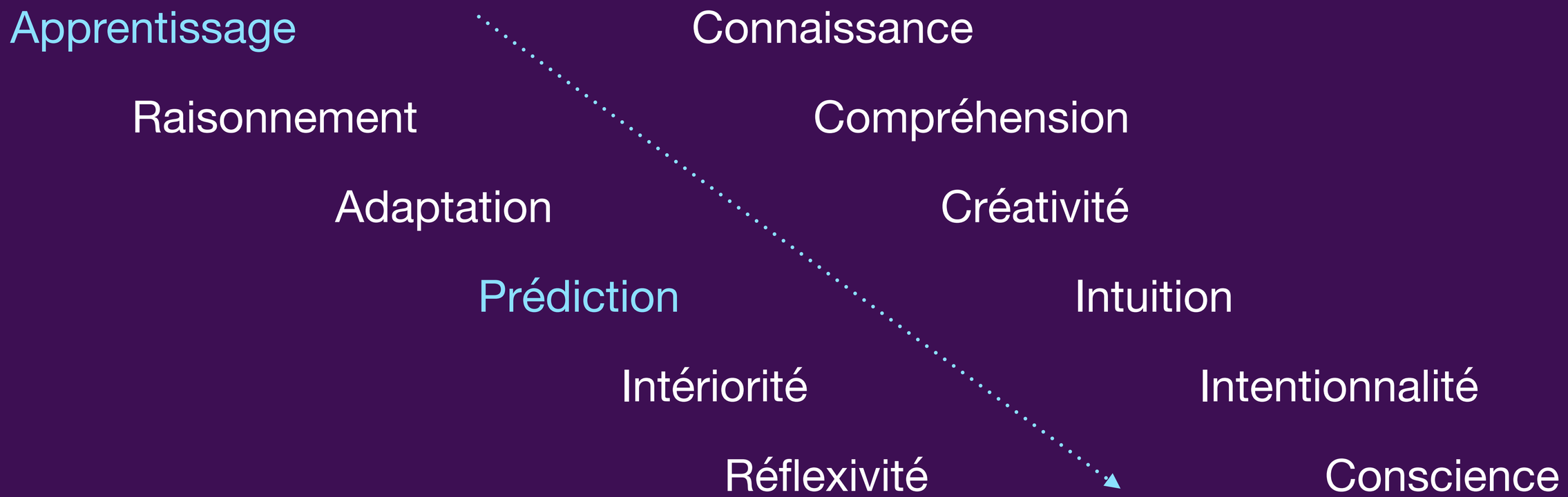
1. Que s'est-il passé ?

Fondements de la dynamique actuelle

A. Intelligences

Un concept fuyant aux acceptions multiples

Définitions problématiques





L'*AI effect*

Reconnaissance de caractères (Lecun - Tri postal - 1989 / Chèques - 1993)

Échecs (IBM Deep Blue - Kasparov - 1997)

Érudition - Jeopardy (IBM Watson - Ken Jennings - 2011)

Go (DeepMind AlphaGo - Lee Sedol — 2016)

Traduction ? (DeepL - 2017)

Compréhension d'un texte ? (Alibaba - 2018)

Art visuel ? (Dall-E - 2021, Midjourney - 2022)

Conduite automobile ? (Waymo - 2022)

Dialogue généraliste (ChatGPT - 2022)

Désambiguïsation lexicale (GPT 3.5 - 2022)

Littérature ? (Claude Sonnet 3.5 - Hervé Le Tellier - 2025)



Garry Kasparov



Ken Jennings



Lee Sedol



Hervé Le Tellier

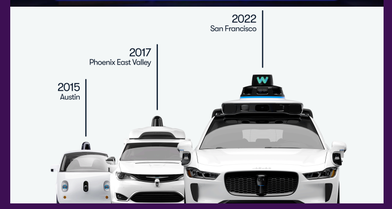
1 • QUE S'EST-IL PASSÉ ?











1990

Reconnaissance de caractères (Lecun - Tri postal - 1989 / Chèques - 1993)

Échecs (IBM Deep Blue - Kasparov - 1997)



Garry Kasparov

2000



Ken Jennings

2010

Érudition - Jeopardy (IBM Watson - Ken Jennings - 2011)



Lee Sedol

2020

Go (DeepMind AlphaGo - Lee Sedol — 2016)

Traduction ? (DeepL - 2017)

Compréhension d'un texte ? (Alibaba - 2018)

Art visuel ? (Dall-E - 2021, Midjourney - 2022)

Conduite automobile ? (Waymo - 2022)

Dialogue généraliste (ChatGPT - 2022)

Désambiguïsation lexicale (GPT 3.5 - 2022)



Hervé Le Tellier

Littérature ? (Claude Sonnet 3.5 - Hervé Le Tellier - 2025)

Traduction

Compréhension

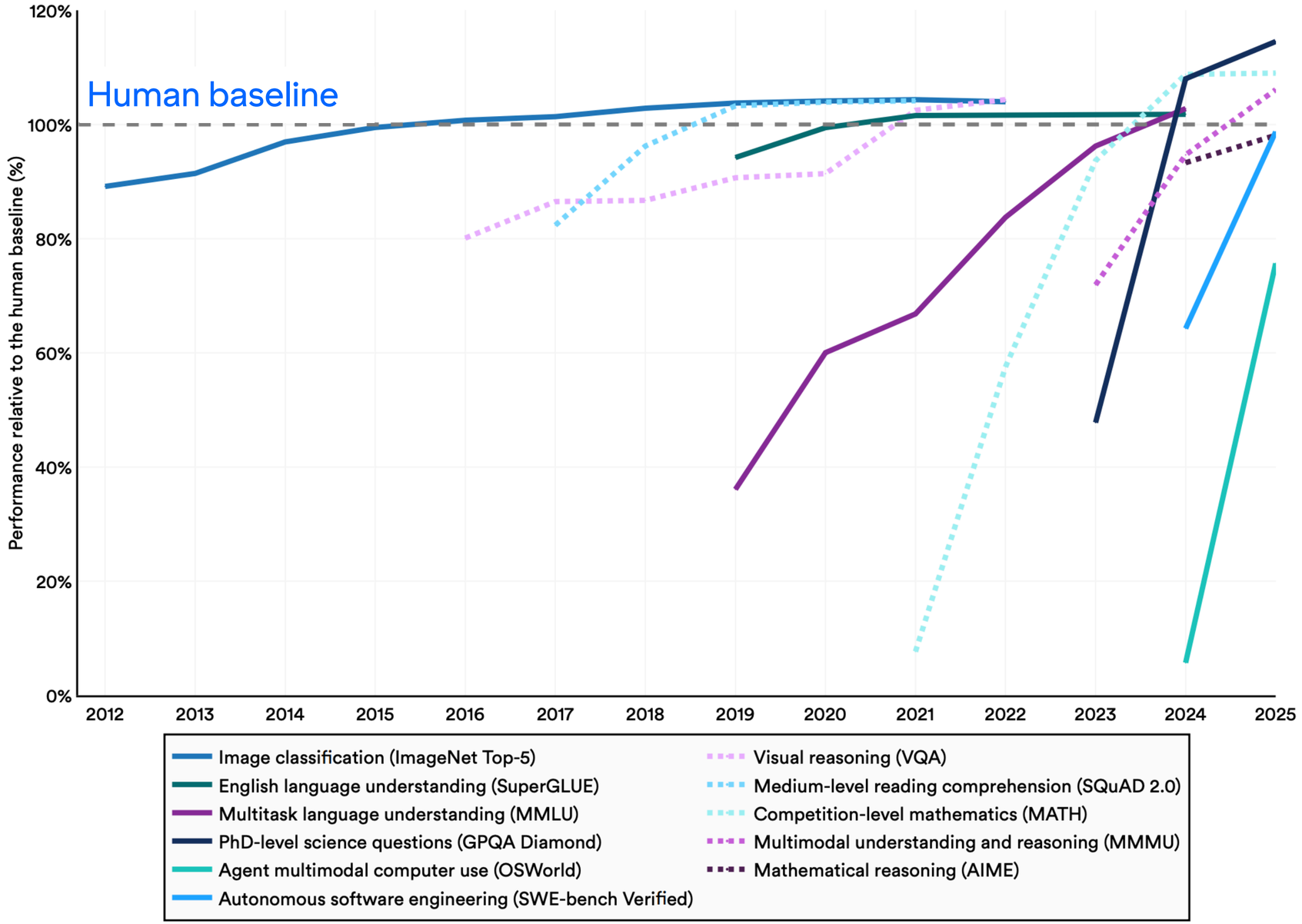
Connaissance

Programmation

Vision

Mathématiques

...



Unil.

Traduction

Compréhension

Connaissance

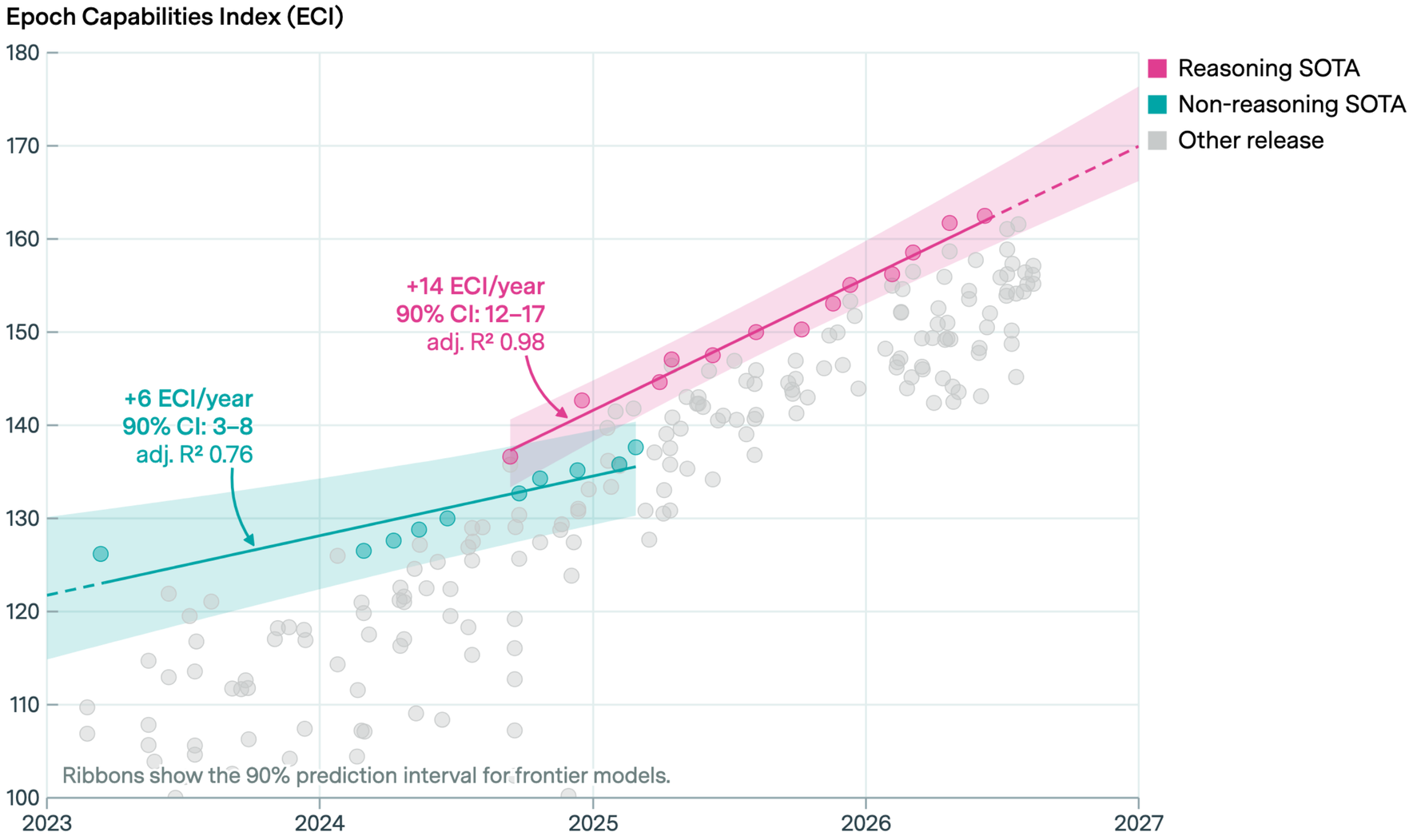
Programmation

Vision

Mathématiques

...

The ECI frontier has advanced by 14 points per year since the introduction of reasoning models

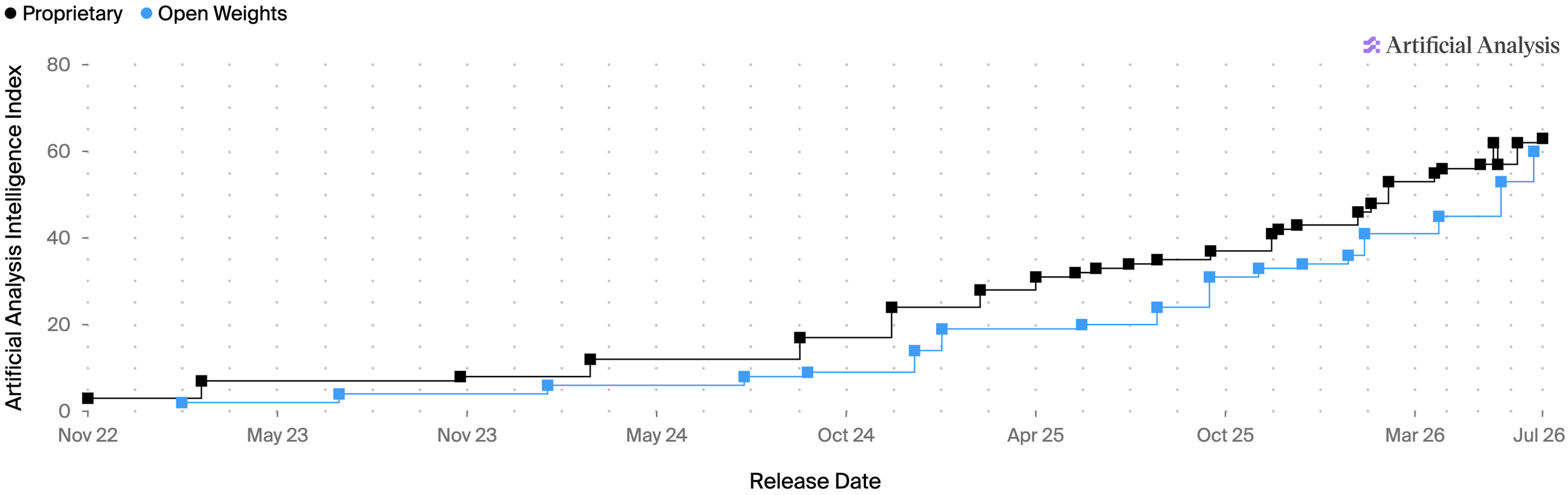


EPOCH AI | CC-BY

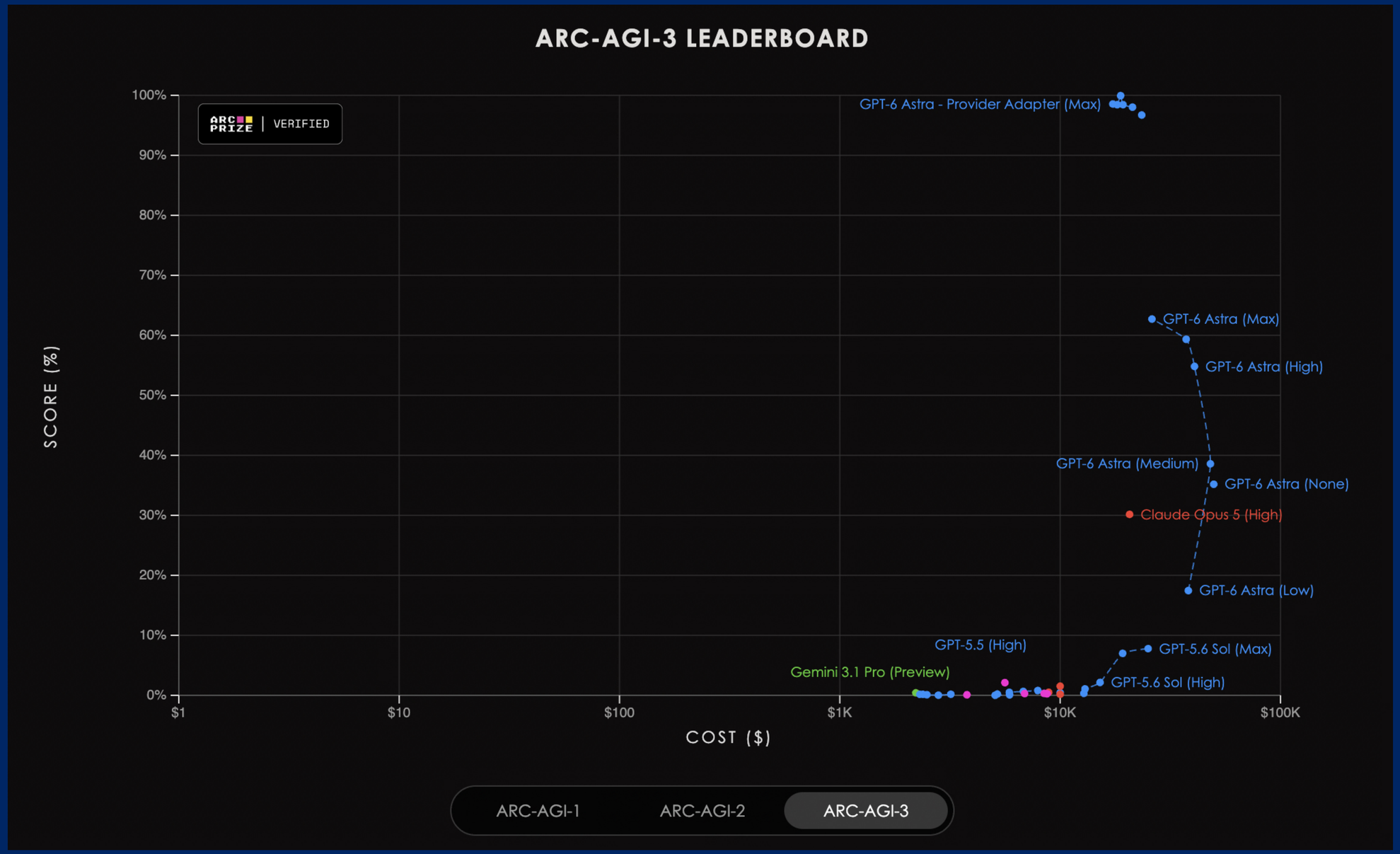
epoch.ai

Progress in Open Weights vs. Proprietary Intelligence

Artificial Analysis Intelligence Index v4.1.1 incorporates 9 evaluations: GDPval-AA v2, τ^3-Banking, Terminal-Bench v2.1, SciCode, Humanity's Last Exam, GPQA Diamond, CritPt, AA-Omniscience, AA-LCR



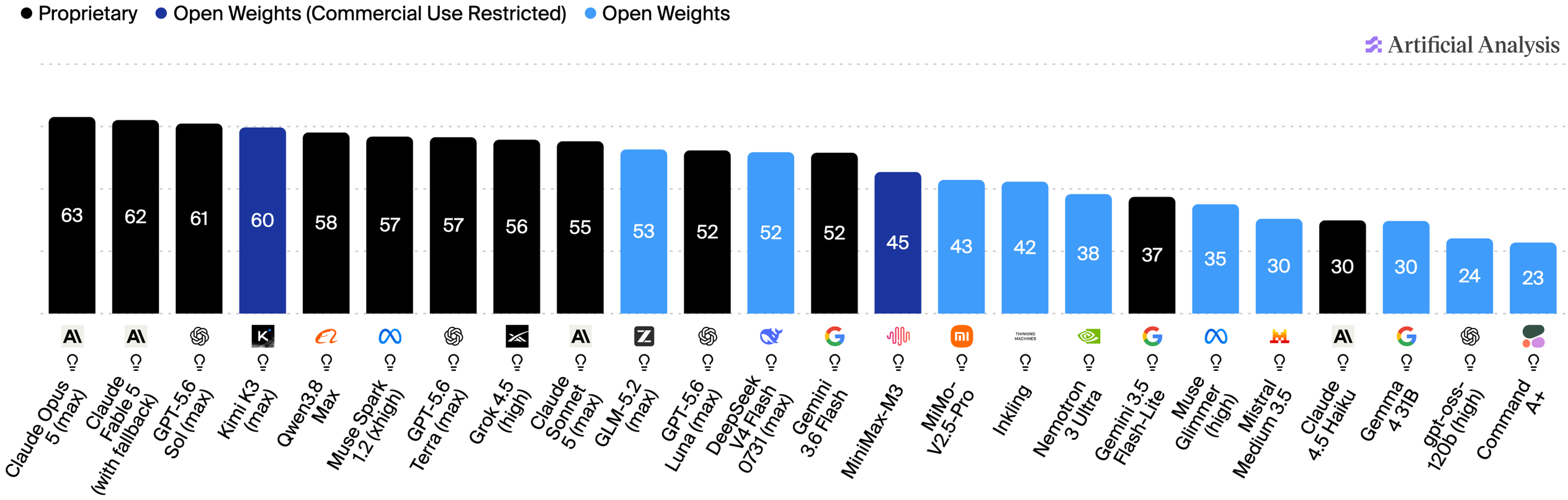
Reasoning models are indicated by a lightbulb icon


ARC-AGI-3 LEADERBOARD
ARC PRIZE | VERIFIED
SCORE (%)
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
$1
$10
$100
$1K
$10K
$100K
COST ($)
GPT-6 Astra - Provider Adapter (Max)
GPT-6 Astra (Max)
GPT-6 Astra (High)
GPT-6 Astra (Medium)
GPT-6 Astra (None)
Claude Opus 5 (High)
GPT-6 Astra (Low)
GPT-5.5 (High)
GPT-5.6 Sol (Max)
Gemini 3.1 Pro (Preview)
GPT-5.6 Sol (High)
ARC-AGI-1
ARC-AGI-2
ARC-AGI-3

Unil.

Artificial Analysis Intelligence Index by Open Weights / Proprietary

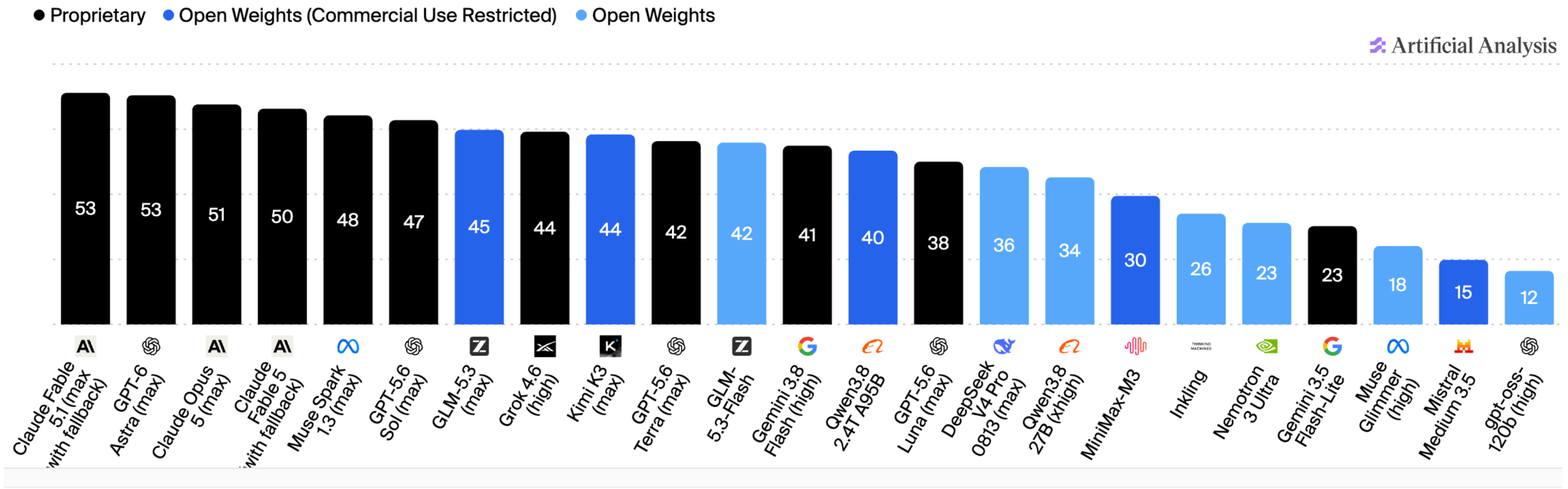
Artificial Analysis Intelligence Index v4.1.1 incorporates 9 evaluations: GDPval-AA v2, τ^3-Banking, Terminal-Bench v2.1, SciCode, Humanity's Last Exam, GPQA Diamond, CritPt, AA-Omniscience, AA-LCR



Reasoning models are indicated by a lightbulb icon

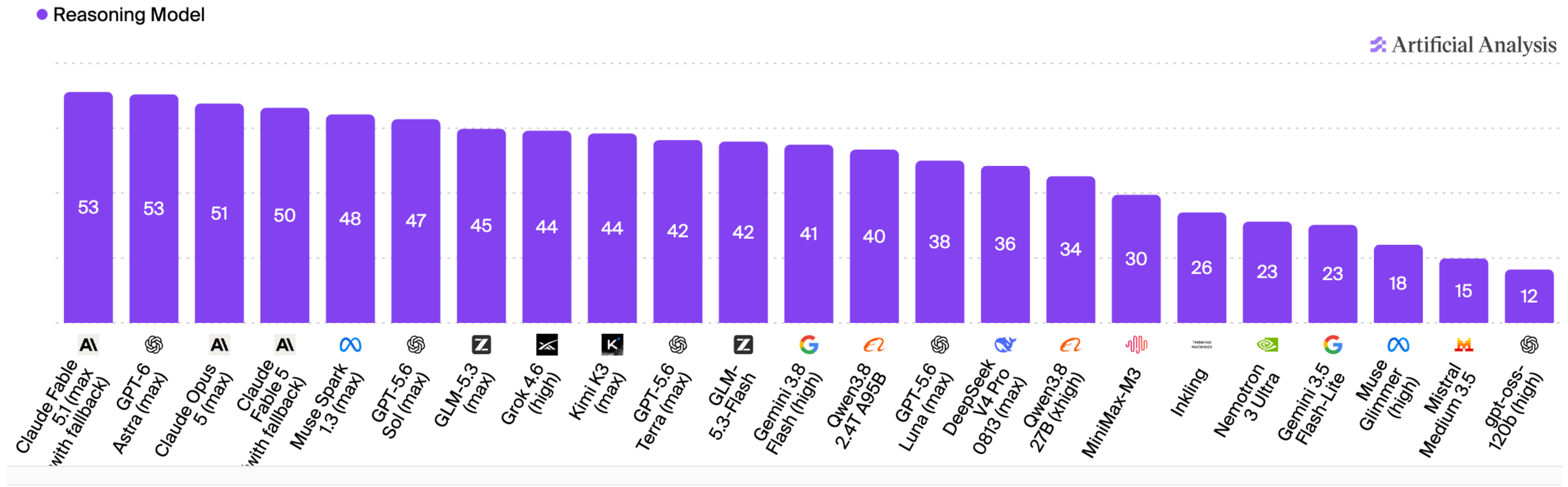
Artificial Analysis Intelligence Index by Open Weights / Proprietary

Artificial Analysis Intelligence Index v4.3 incorporates 10 evaluations: AA-Briefcase, GDPval-AA v2, AutomationBench-AA, Terminal-Bench v4.0, SciCode, Humanity's Last Exam, GDP.pdf, CritPt, AA-Omniscience, AA-LCR v1.1



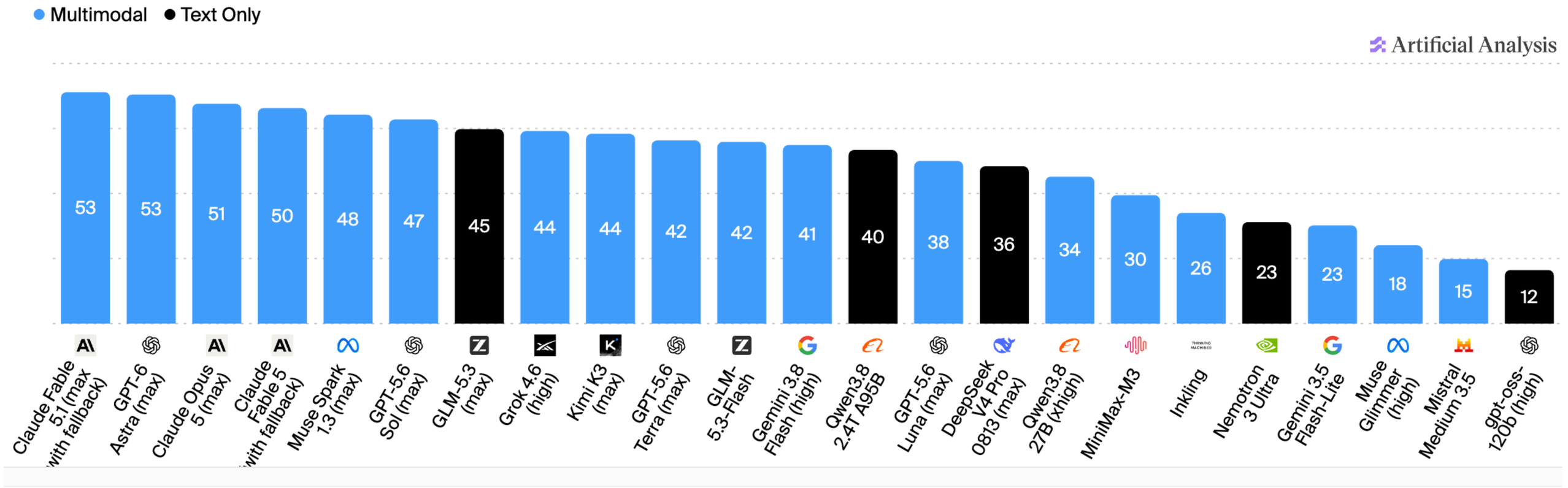
Artificial Analysis Intelligence Index by Model Type

Artificial Analysis Intelligence Index v4.3 incorporates 10 evaluations: AA-Briefcase, GDPval-AA v2, AutomationBench-AA, Terminal-Bench v4.0, SciCode, Humanity's Last Exam, GDP.pdf, CritPt, AA-Omniscience, AA-LCR v1.1



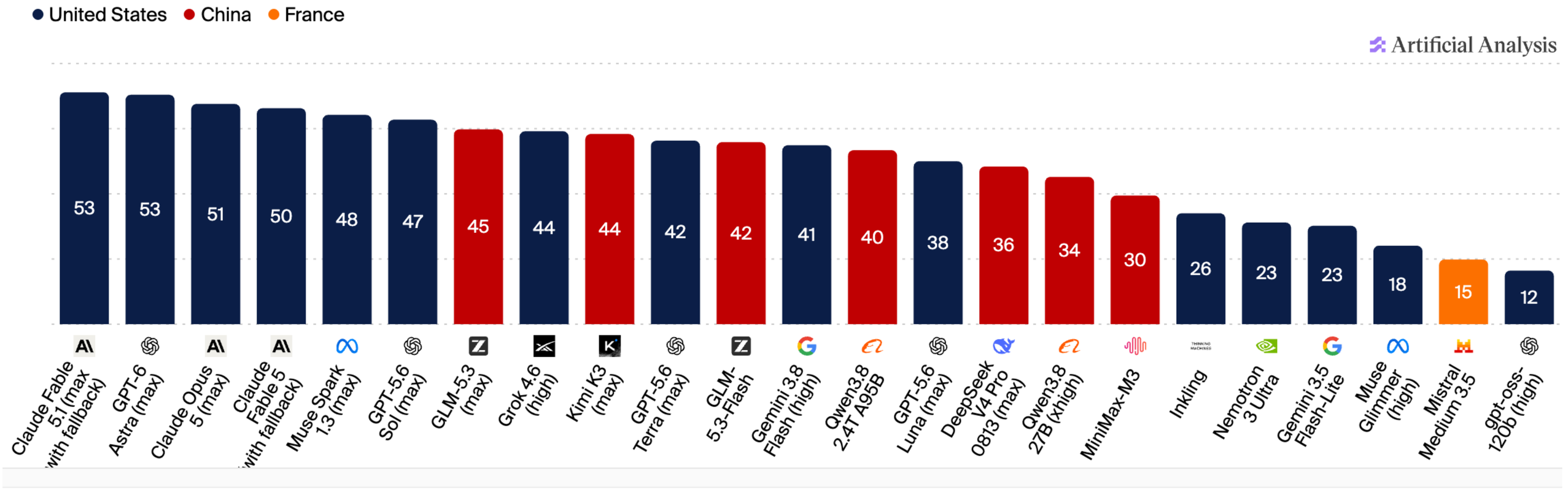
Artificial Analysis Intelligence Index by Input Modality

Artificial Analysis Intelligence Index v4.3 incorporates 10 evaluations: AA-Briefcase, GDPval-AA v2, AutomationBench-AA, Terminal-Bench v4.0, SciCode, Humanity's Last Exam, GDP.pdf, CritPt, AA-Omniscience, AA-LCR v1.1



Leading Models by Country

Artificial Analysis Intelligence Index · Leading models

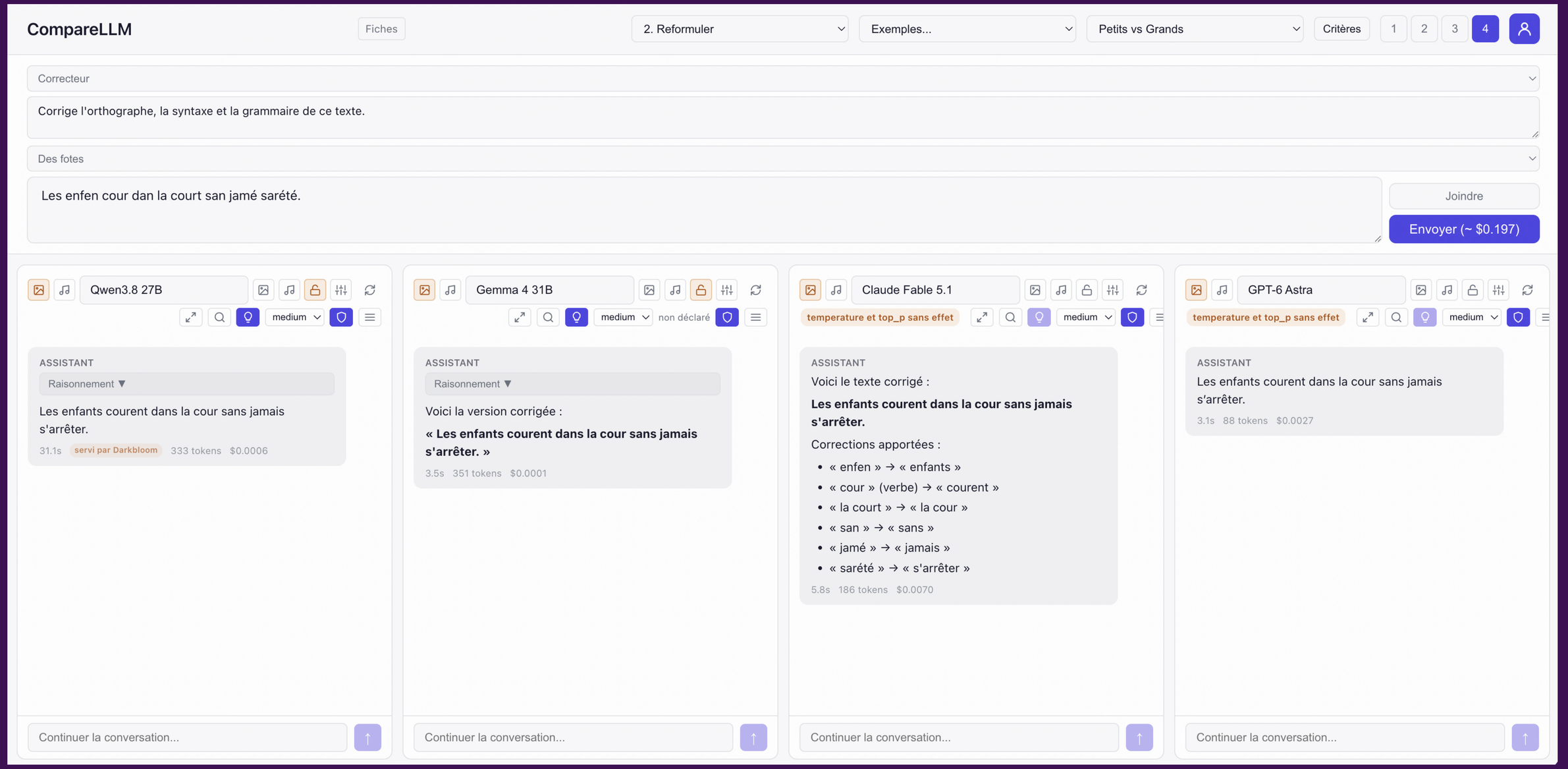


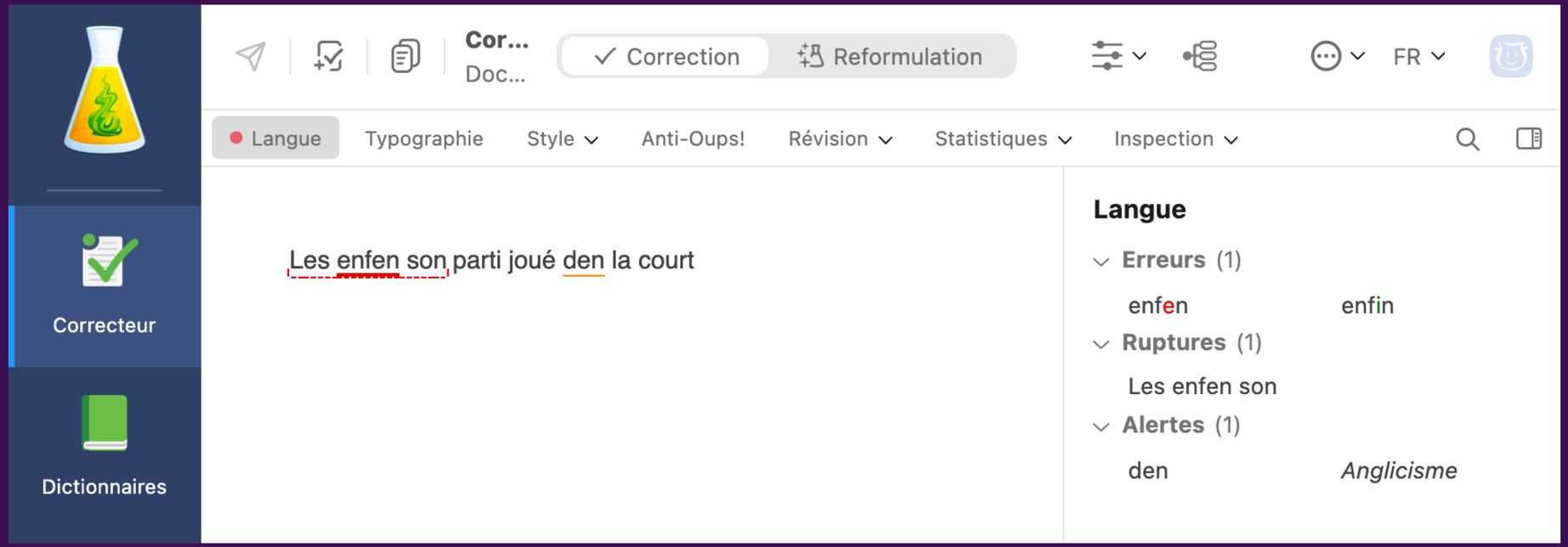
B. Exemple simple

Contextes et modèles de monde

Corrige cette phrase : Les enfen son parti joué den la court

1 • QUE S'EST-IL PASSÉ ?




Correcteur
Dictionnaires
Cor...
Doc...
Correction
Reformulation
FR
Langue
Typographie
Style
Anti-Oups!
Révision
Statistiques
Inspection
Les enfen son parti joué den la court
Langue
Erreurs (1)
enfen
enfin
Ruptures (1)
Les enfen son
Alertes (1)
den
Anglicisme

C. L'accélération

Un siècle d'innovations, une décennie hors normes

L'accélération

Près d'un siècle d'innovations

Données
Abondantes

Processeurs
Plus puissants

Algorithmes
Plus adaptés

Fin 2022, ChatGPT héritait de recherches accumulées depuis la Seconde Guerre mondiale

La rupture actuelle tient à des innovations récentes

Des données abondantes

Croissance exponentielle des données

Corpus de pré-entraînement

Dataset	Année	Taille	Thématique / Contenu	Ouvert	Utilisé par
Common Crawl	2008–	Pétaoctets (brut)	Web généraliste (pages HTML brutes)	Oui	Source nombreux corpus
C4 (Colossal Clean Crawled Corpus)	2019	~156B tokens	Web anglophone filtré (1 snapshot CC)	Oui	T5, UL2, LLaMA 1
The Pile	2020	~300B tokens (825 Go)	Multidomain : web, livres, code, articles scientifiques, droit, brevets, e-mails, forums	Oui (restreint)	GPT-NeoX, Pythia
ROOTS	2022	~341B tokens (1,6 To)	Multilingue (59 langues) : web, livres, articles, Wikipédia, code	Accès conditionnel	BLOOM
RedPajama V1	2023	1,2T tokens	Web, livres, code, Wikipédia, ArXiv, StackExchange	Oui	RedPajama-INCITE
RedPajama V2	2023	100T+ tokens (brut)	Web (84 snapshots CC) avec signaux qualité	Oui	Recherche, curation personnalisée
RefinedWeb	2023	5T tokens (600B public)	Web exclusivement, déduplication stricte	Partiel	Falcon 1 & 2
SlimPajama	2023	627B tokens	Web, livres, code, Wikipédia (dédup. globale de RedPajama)	Oui	Cerebras, recherche
The Stack v1 / v2	2022–24	~900B tokens (v2)	Code source : 619 langages de programmation	Oui	StarCoder, StarCoder2
Dolma (v1 → v3)	2024–25	3T → 9,3T tokens	Web, articles scientifiques, code, livres, Wikipédia, Reddit, PDFs académiques	Oui	OLMo, OLMo 2, OLMo 3
FineWeb	2024	15T tokens (~44 To)	Web nettoyé (96 dumps CC, 2013–2024)	Oui	SmolLM, recherche
FineWeb-Edu	2024	~1,3T tokens	Sous-ensemble éducatif de FineWeb (filtré par classifieur)	Oui	SmolLM, recherche
FineWeb-2	2024	Multi-To	Web multilingue (1 000+ langues)	Oui	Recherche multilingue
DCLM (DataComp-LM)	2024	240T tokens (pool)	Web (benchmark pour comparer les stratégies de curation)	Oui	DCLM-Baseline
Common Pile	2025	~8 To	Textes exclusivement sous licence ouverte / domaine public	Oui	En cours d'adoption
Common Corpus	2025	Multi-To	Textes multilingues sous licence ouverte (initiative Pleias)	Oui	Recherche
MassiveText	2021	~2,4T tokens	Web (interne), livres, actualités, code, Wikipédia	Non	Gopher, Chinchilla
Corpus GPT-3	2020	~300B tokens	Web filtré (CC), WebText2, livres, Wikipédia	Non	GPT-3
Corpus LLaMA 1	2023	1,4T tokens	Web (CCNet), C4, GitHub, livres, ArXiv, Wikipédia, StackExchange	Non	LLaMA 1
Corpus LLaMA 3	2024	15T+ tokens	Web, code (×4), multilingue (30+ langues)	Non	LLaMA 3 / 3.1

Corpus d'instructions (SFT — Supervised Fine-Tuning)

Dataset	Année	Taille	Thématique / Contenu	Source	Ouvert	Utilisé par
FLAN Collection	2023	1 800+ tâches	NLP académique : classification, QA, résumé, traduction, raisonnement	Tâches NLP + templates	Oui	**Flan-T5, Flan-PaLM**
Super-NaturalInstructions	2022	1 616 tâches, 5M+ instances	NLP multi-tâches : extraction, génération, classification, dialogue	Annotation experte	Oui	**Tk-Instruct**
Stanford Alpaca	2023	52K exemples	Instructions généralistes (Q&A, rédaction, brainstorming, code)	Synthétique (text-davinci-003)	Oui	**Alpaca 7B**
Dolly-15k	2023	15K exemples	Q&A ouverte/fermée, résumé, brainstorming, rédaction, extraction	100 % humain	Oui (CC-BY-SA)	**Dolly 2.0**
OpenAssistant (OASST1/2)	2023	161K messages, 35 langues	Conversations multilingues d'assistant : aide, code, créativité, factuel	Crowdsourcing (13 500 volontaires)	Oui	**Open-Assistant**
ShareGPT	2023	~90K conversations	Conversations réelles utilisateur–ChatGPT (tous sujets)	Scraping de conversations partagées	Gris (non licencié)	**Vicuna**
UltraChat	2023	1,5M dialogues	Conversations : monde, création de contenu, aide sur matériaux existants	Synthétique (ChatGPT ↔ ChatGPT)	Oui	**Zephyr, UltraLM**
WizardLM / Evol-Instruct	2023	70K+ instructions	Instructions complexifiées (raisonnement multi-étape, code avancé)	Synthétique (évolution par LLM)	Oui	**WizardLM**
Orca-style data	2023	5M + 1M exemples	Instructions avec traces d'explication détaillées (chain-of-thought)	Synthétique (GPT-3.5 / GPT-4)	Partiel	**Orca 1 & 2**
OpenHermes 2.5	2023	~1M exemples	Agrégation multi-source : code, raisonnement, rôle-play, Q&A	Synthétique (multiples pipelines)	Oui	**Hermes-Mistral**
Tülu 3 SFT Mix	2024	~939K exemples	Mélange : math, code, sécurité, instruction following, chat	43 % synthétique, 57 % public	Oui	**Tülu 3 (OLMo)**
Magpie	2024	300K–4M exemples	Instructions auto-générées par LLM (sans prompt humain initial)	Synthétique (auto-extraction LLaMA 3)	Oui	**Recherche**

Corpus d'alignement et de préférences (RLHF / DPO)

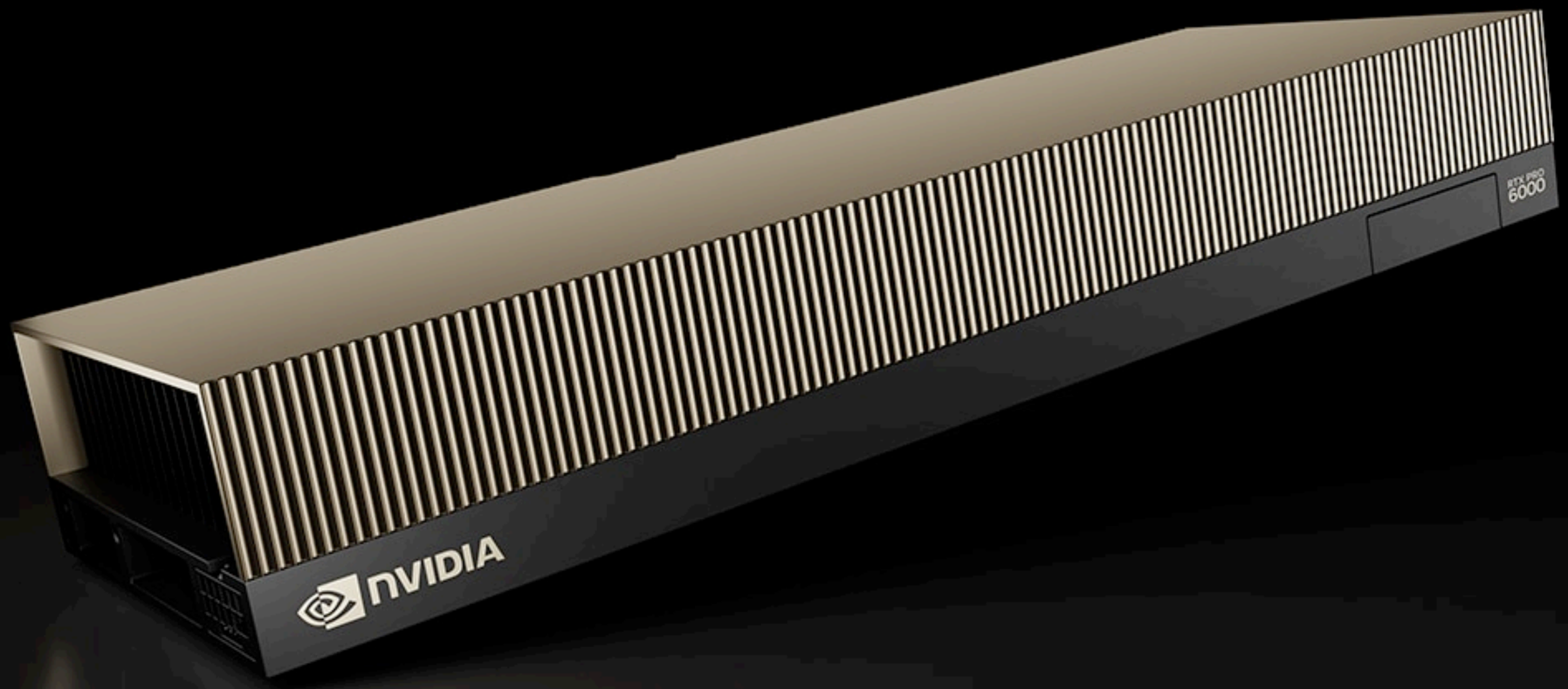
Dataset	Année	Taille	Thématique / Contenu	Source	Ouvert	Utilisé par
Anthropic HH-RLHF	2022	~161K paires	Helpfulness & harmlessness : paires (réponse choisie vs rejetée)	Annotation humaine	Oui	**Claude (early), recherche**
InstructGPT data	2022	~33K comparaisons	Comparaisons de réponses classées (prompts API OpenAI)	Annotation humaine (40 contractuels)	Non	**InstructGPT, GPT-3.5**
UltraFeedback	2023	64K prompts × 4 réponses	Évaluations multi-critères (utilité, honnêteté, profondeur, harmlessness)	Annotation GPT-4	Oui	**Zephyr-β, Tülu 2, Notus**
HelpSteer2	2024	~10K–21K paires	Multi-attributs : helpfulness, correctness, coherence, complexity, verbosity	100 % humain (Scale AI)	Oui (CC-BY-4.0)	**Nemotron, reward models**
Chatbot Arena	2023–	1M+ conversations	Duels humains entre modèles anonymes (préférences organiques)	Utilisateurs réels	Partiel	**Benchmark Elo de référence**
Nectar	2023	183K exemples	Préférences classées (7 réponses/prompt, ranking GPT-4)	Synthétique	Oui	**Starling-LM**
PKU-SafeRLHF	2023	330K+ annotations	Sécurité / harmlessness spécifiquement	Annotation humaine	Oui	**Beaver, recherche safety**
Tülu 3 preference data	2024	~1M paires	Préférences on-policy (générées par le modèle en entraînement)	Synthétique + vérifiable	Oui	**Tülu 3**
UltraInteract	2024	86K exemples	Préférences sur le raisonnement (math, code, logique) avec arbres de solutions	Synthétique	Oui	**Eurus**

L’extractivisme, les partenariats et l’opacité sont devenus la norme

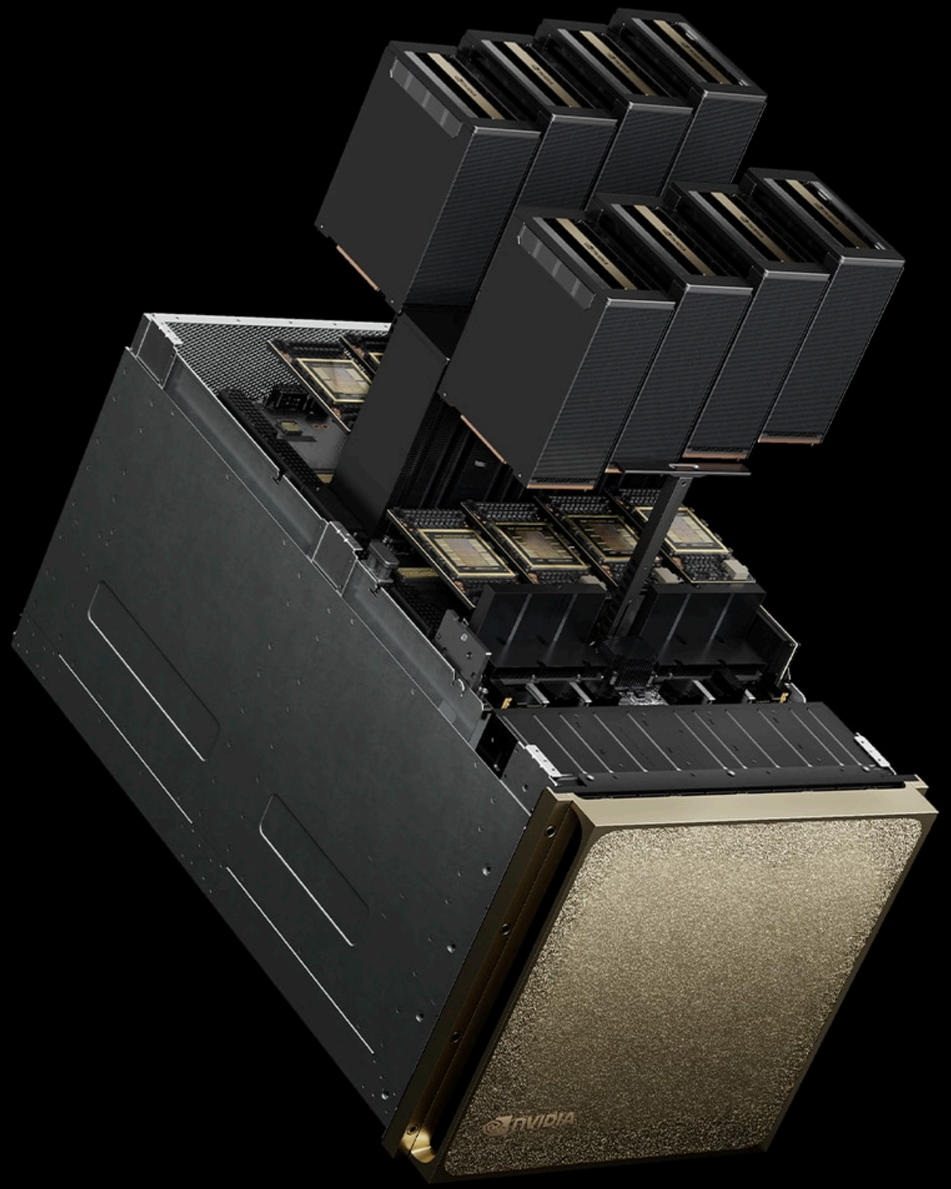
OpenAI, Google, Anthropic, Mistral, DeepSeek,...

Des infrastructures dédiées

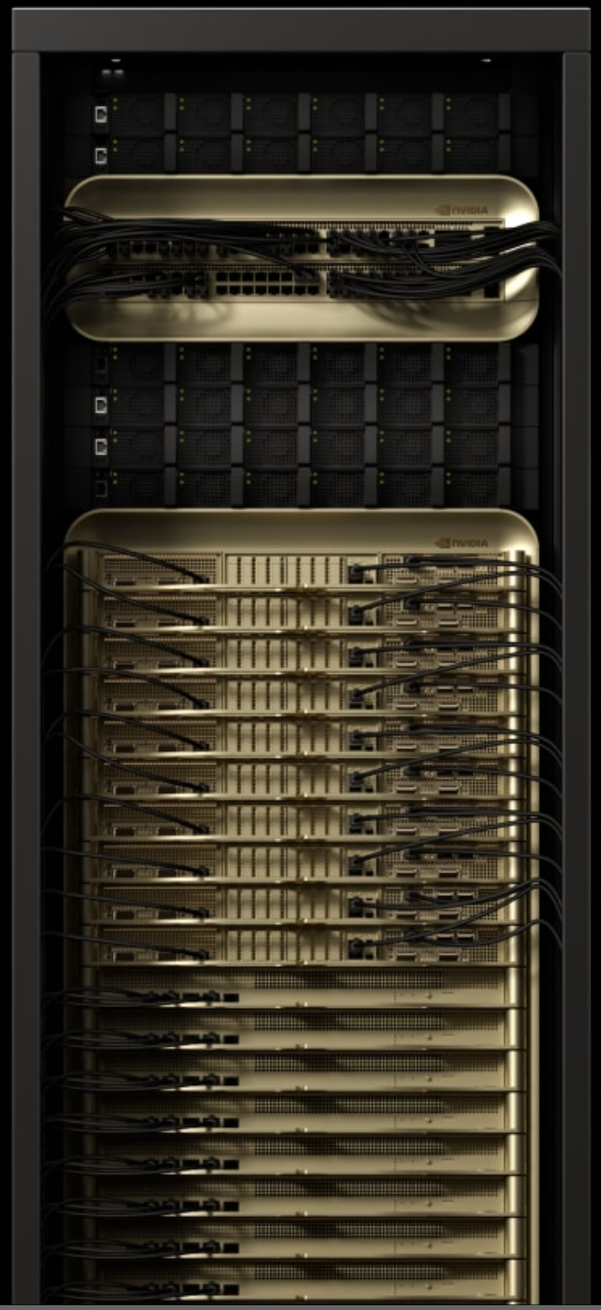
Croissance exponentielle des capacités de traitement



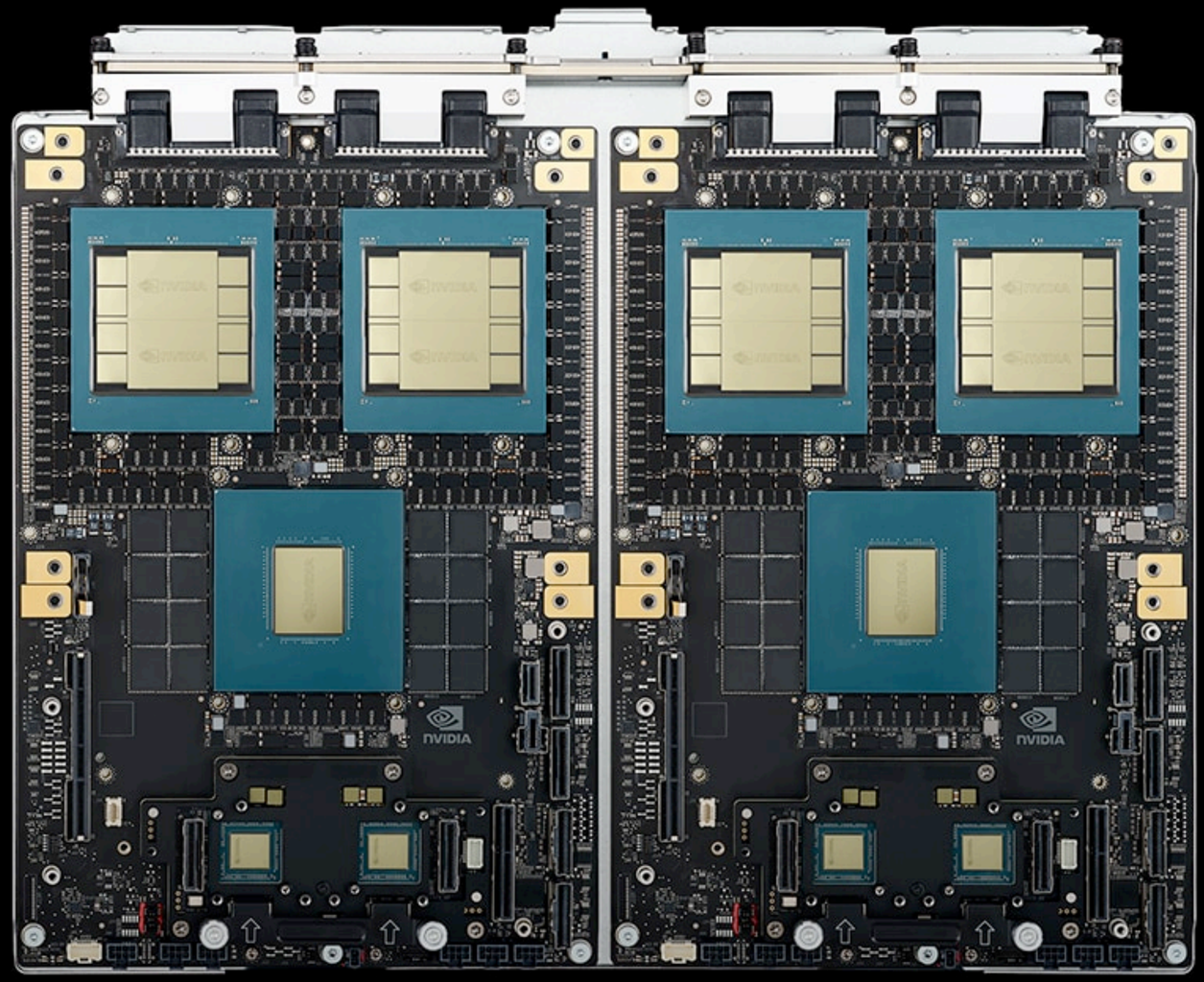
NVIDIA RTX Pro 6000 - 2025



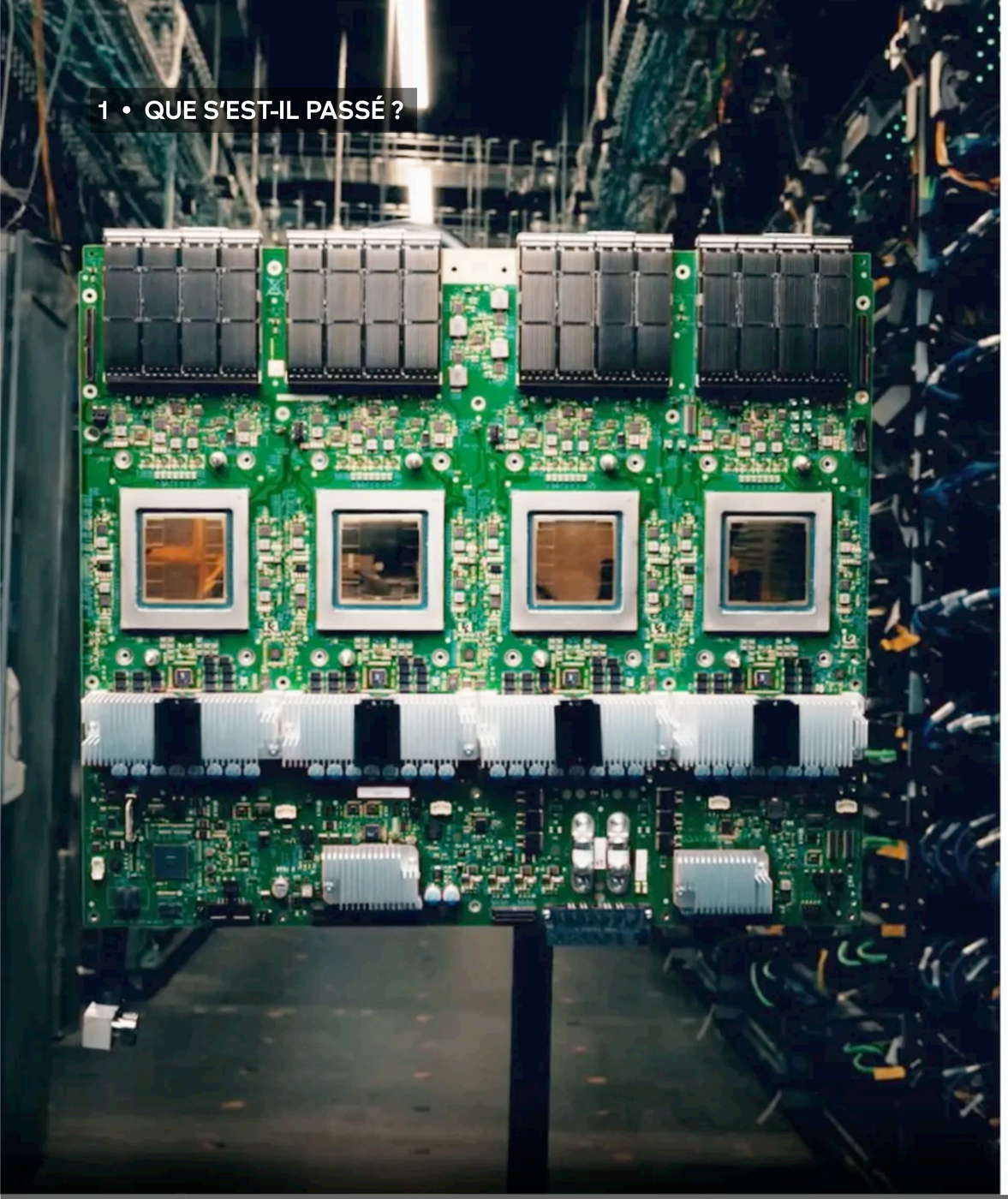
NVIDIA DGX B200 - 2024

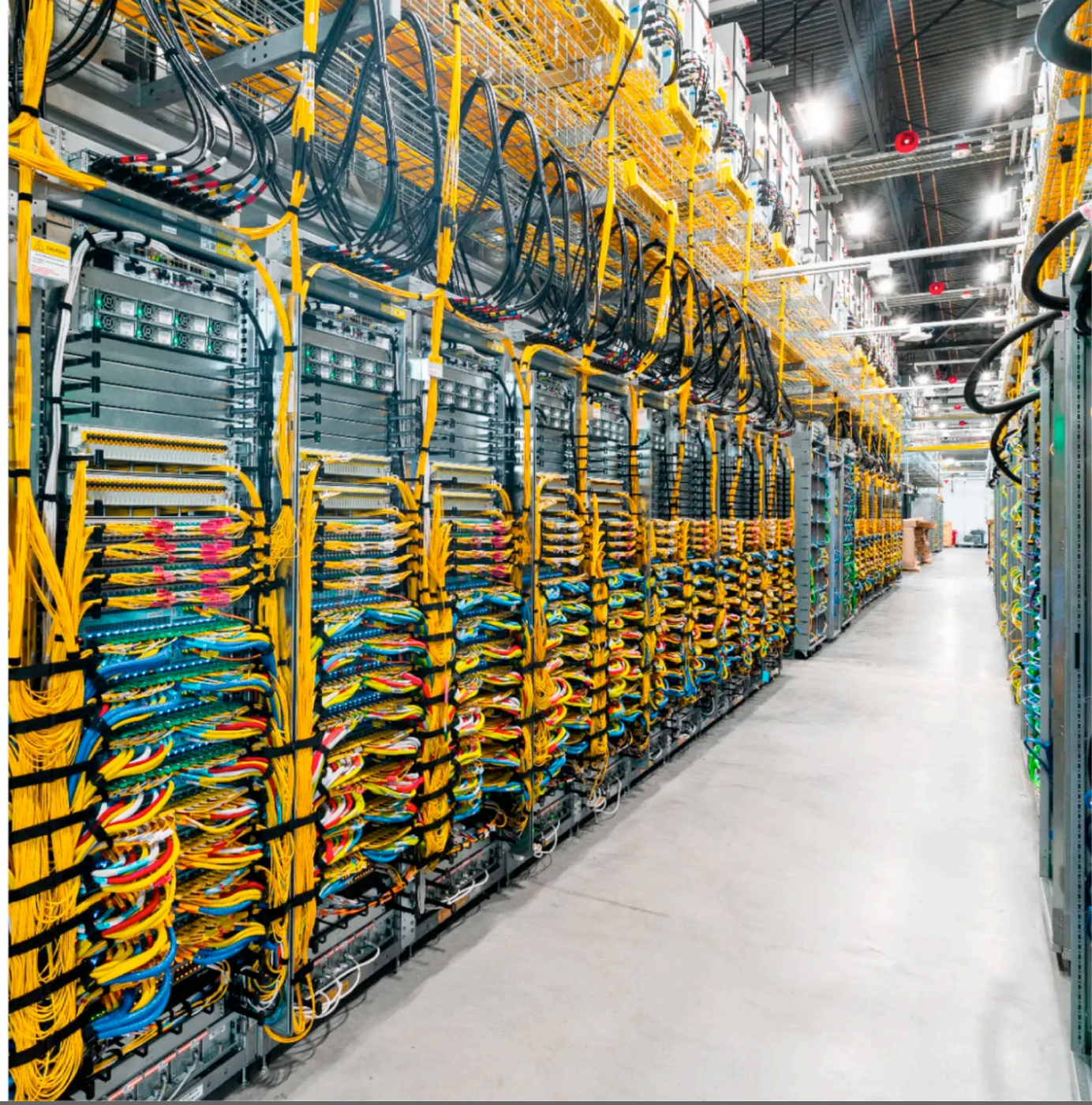


NVIDIA GB300 NVL72 - 2025



Google IronWood TPU - 2025





Google IronWood TPU - 2025

1 • QUE S'EST-IL PASSÉ ?



Boris Beaude - 15 septembre 2026

Google - New Albany - Ohio - 2012

1 • QUE S'EST-IL PASSÉ ?



Google - Water cooling - Data center The Dalles - Oregon - 2012



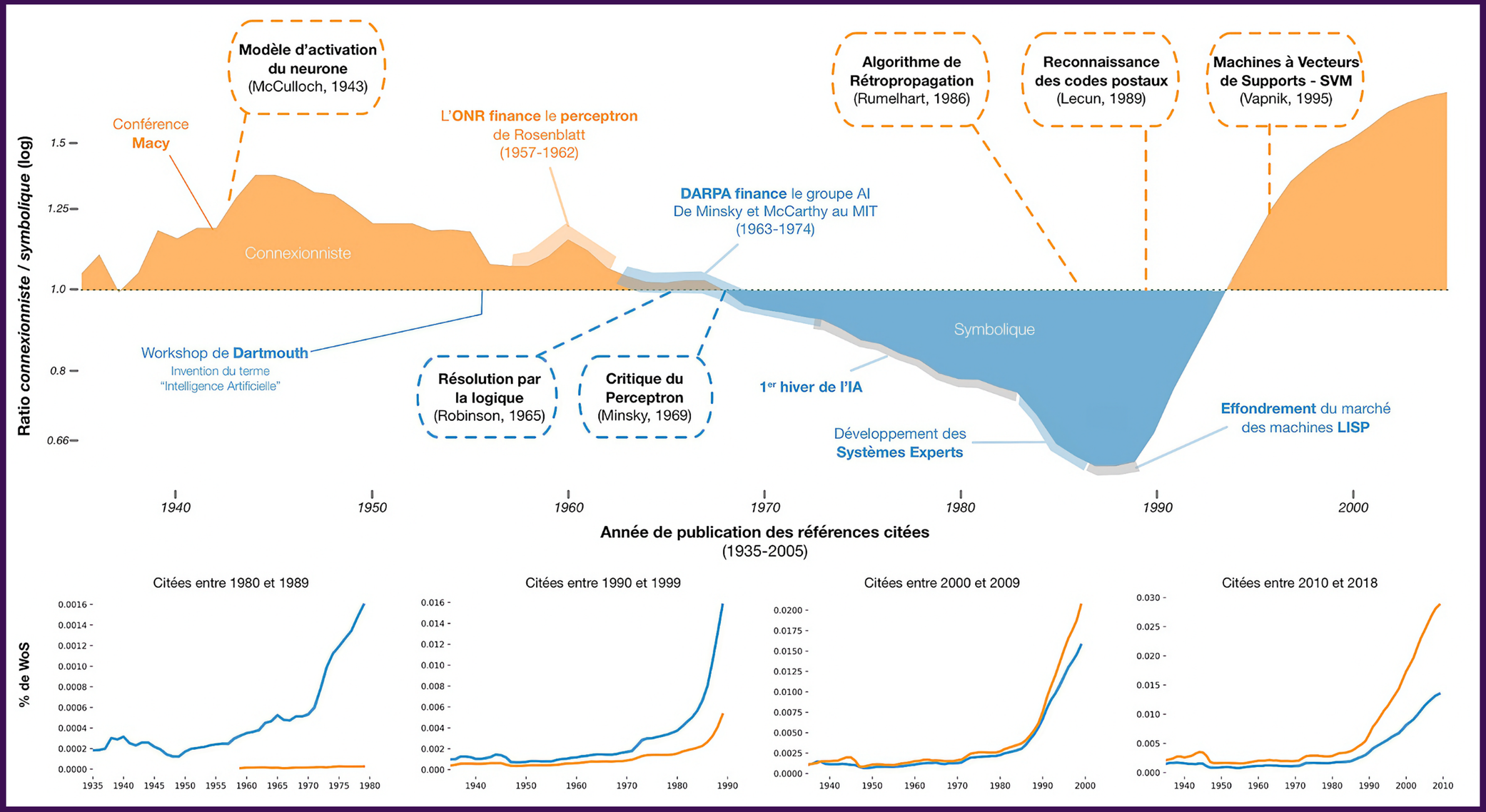
Microsoft - Centrale nucléaire de Three Mile Island, réactivée pour les infrastructures d'IA - Pennsylvanie - The Guardian -. 2024

La revanche des neurones

Changements de paradigmes

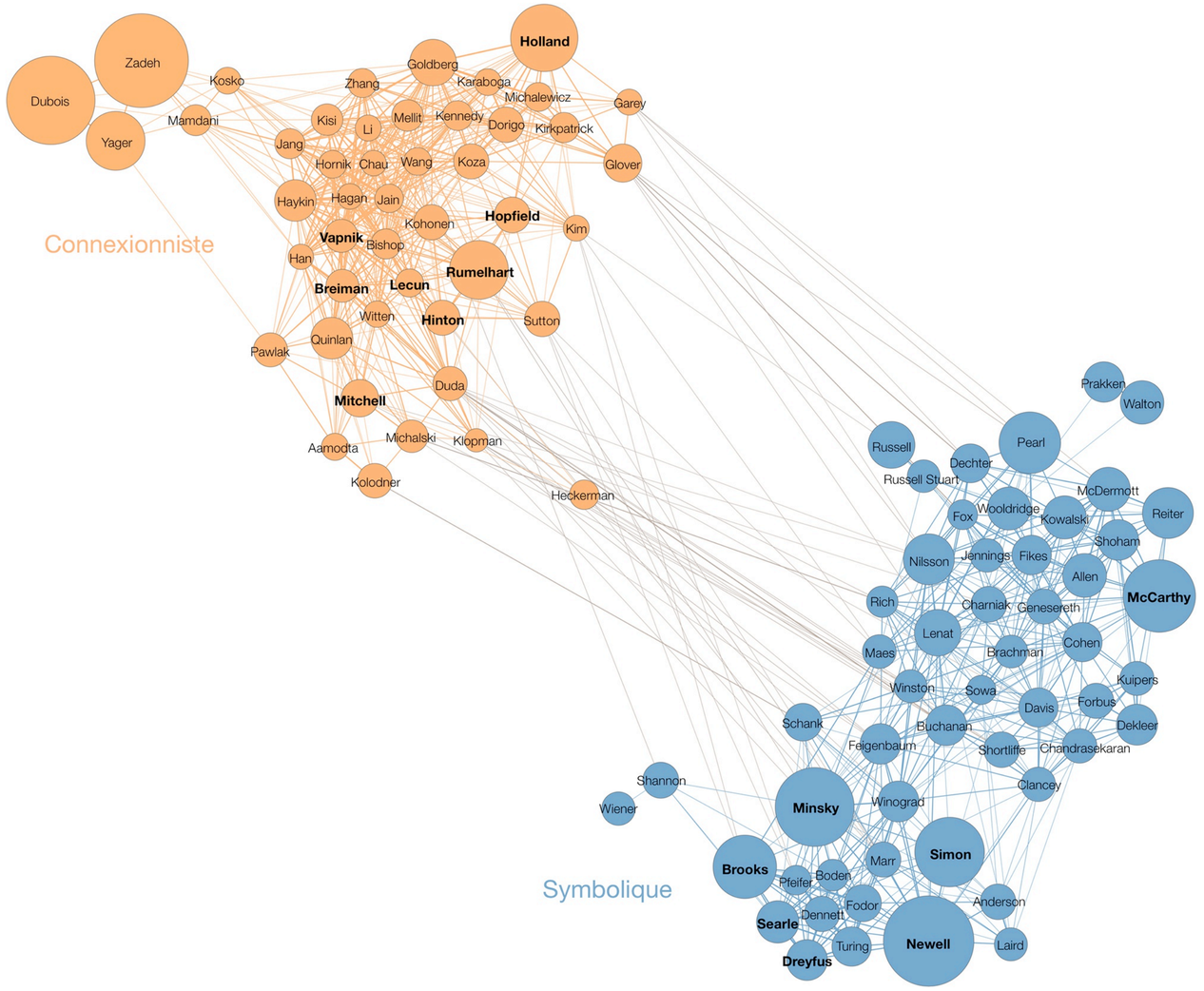
Boris Beaude - 15 septembre 2026

Changement de paradigme



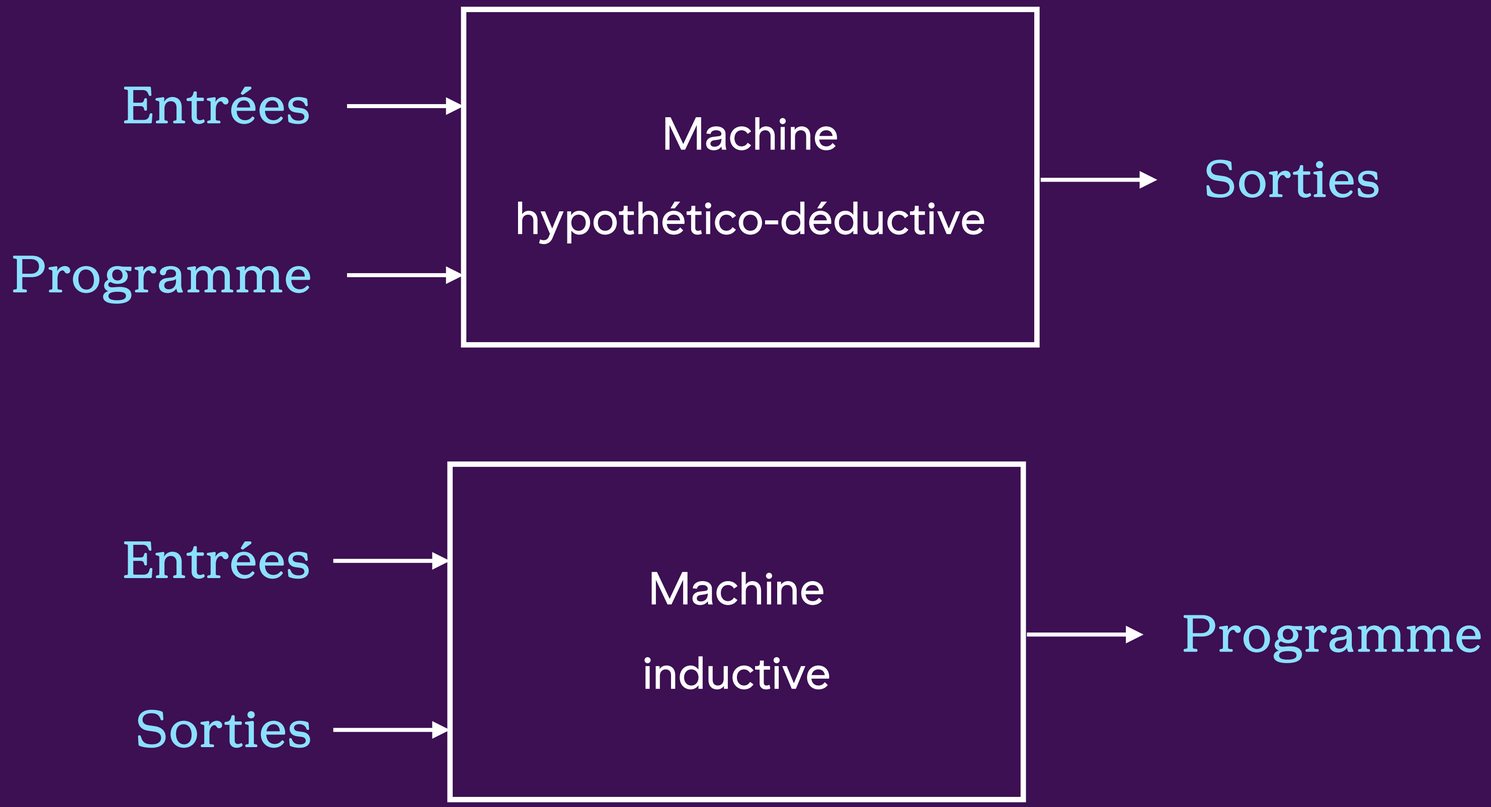
Dominique Cardon, Jean-Philippe Cointet, Antoine Mazières, La revanche des neurones, Réseaux 2018/5 (n° 211), pp. 173 à 220.

Changement de paradigme



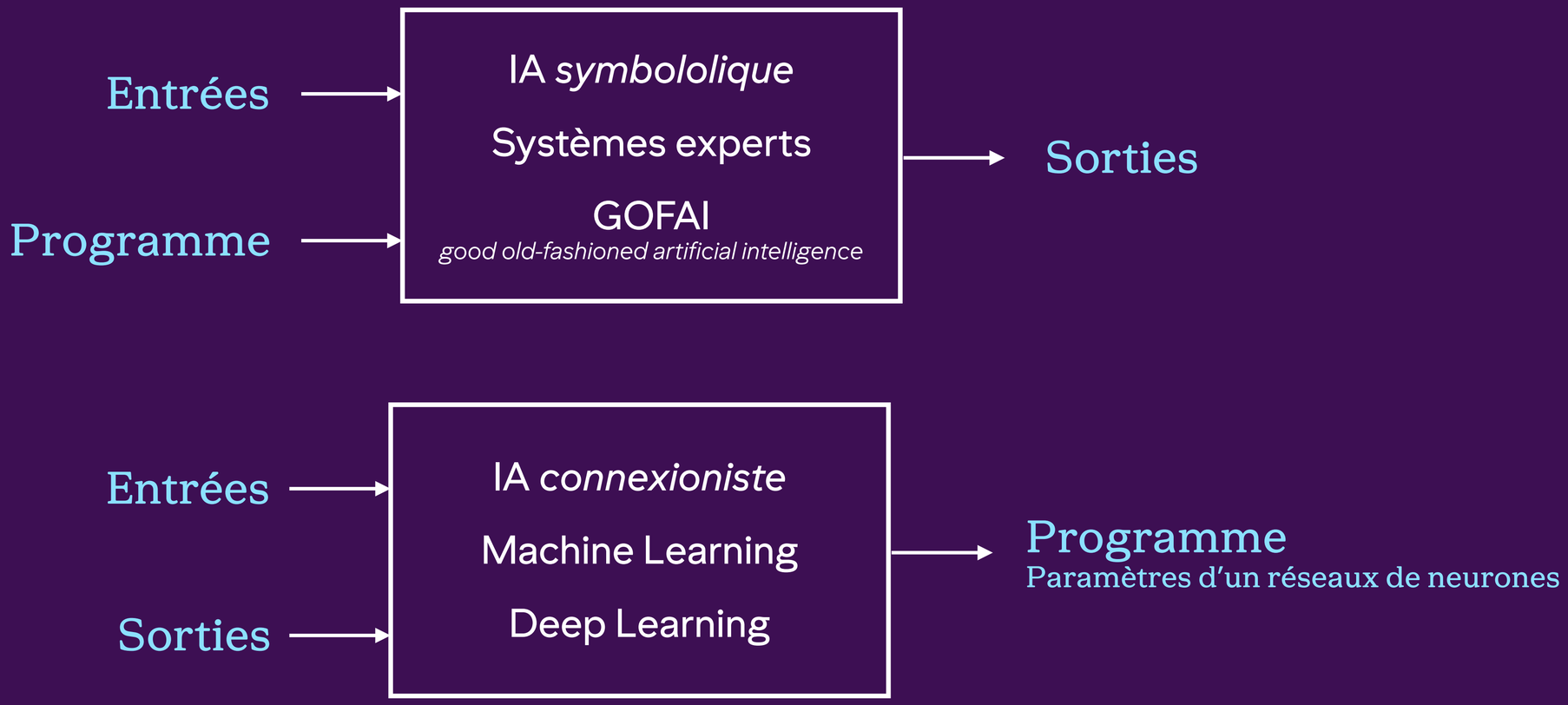
Dominique Cardon, Jean-Philippe Cointet, Antoine Mazières, La revanche des neurones, Réseaux 2018/5 (n° 211), pp. 173 à 220.

Changement de paradigme

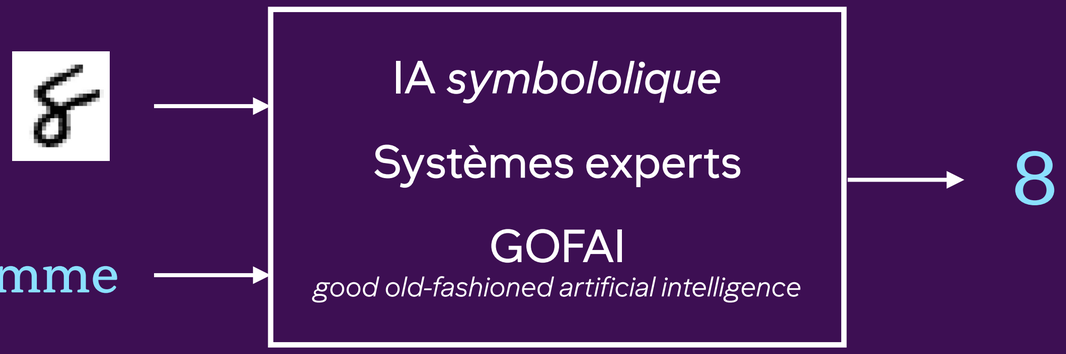


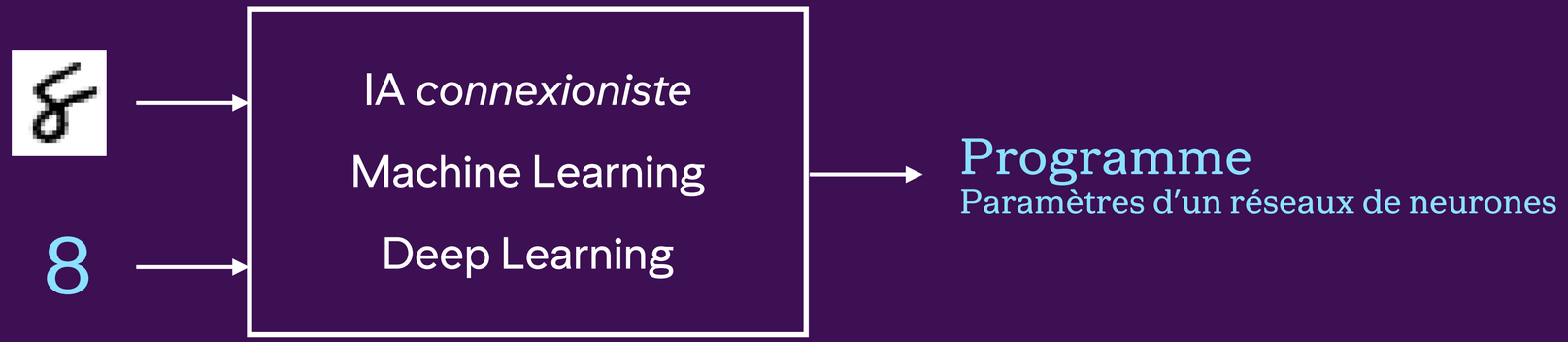
d'après : Dominique Cardon, Jean-Philippe Cointet, Antoine Mazières, La revanche des neurones, Réseaux 2018/5 (n° 211), pp. 173 à 220.

Changement de paradigme



Changement de paradigme





Changement de paradigme



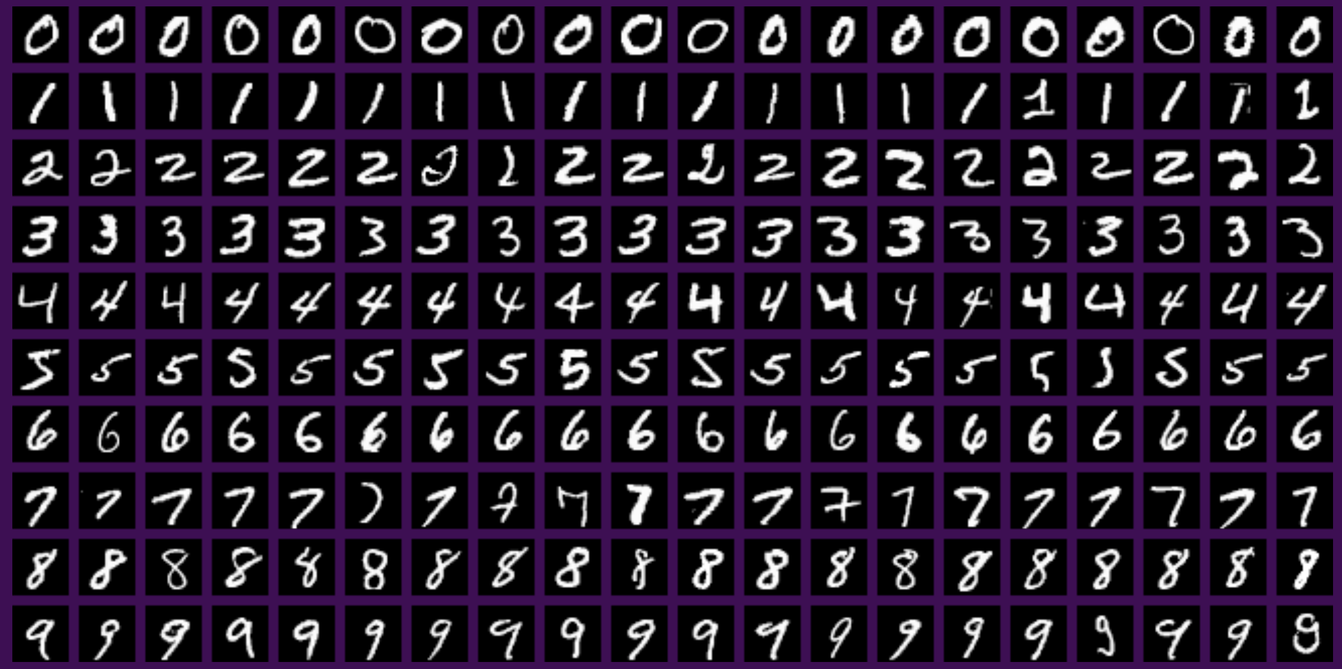
8

IA *connexioniste*

Machine Learning

Deep Learning

→ Programme
Paramètres d'un réseaux de neurones



MNIST dataset - 1994

Changement de paradigme



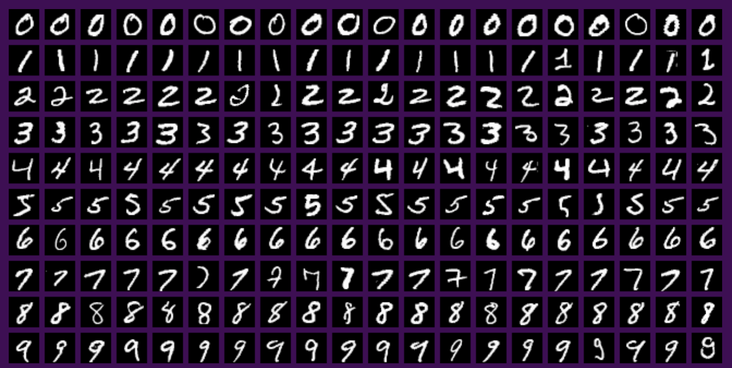
8

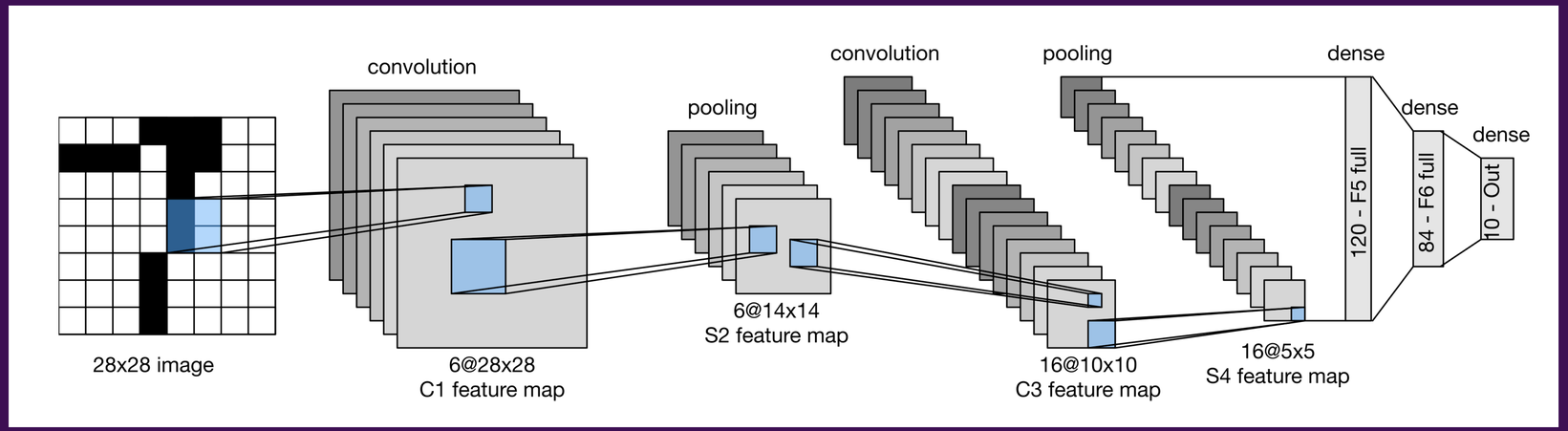
IA *connexioniste*

Machine Learning

Deep Learning

→ Programme
Paramètres d'un réseaux de neurones





LeNet 1989 - Lecun - Représentation de Zhang, Aston and Lipton, Zachary C. and Li, Mu and Smola, Alexander J. 2023

Changement de paradigme



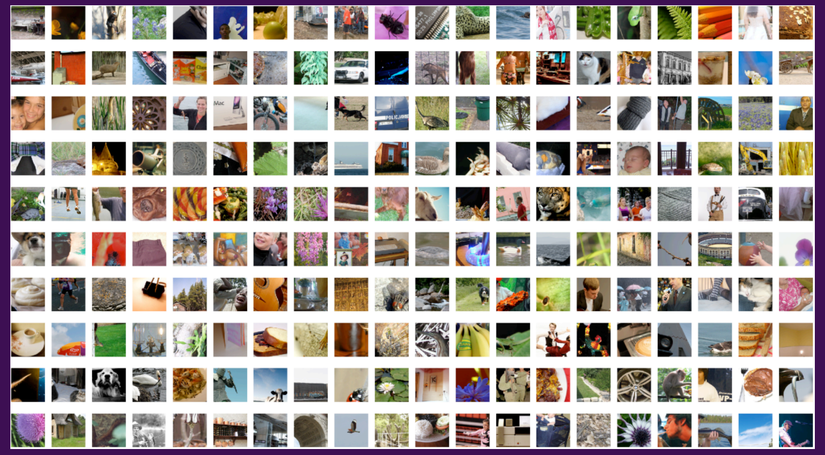
8

IA *connexioniste*

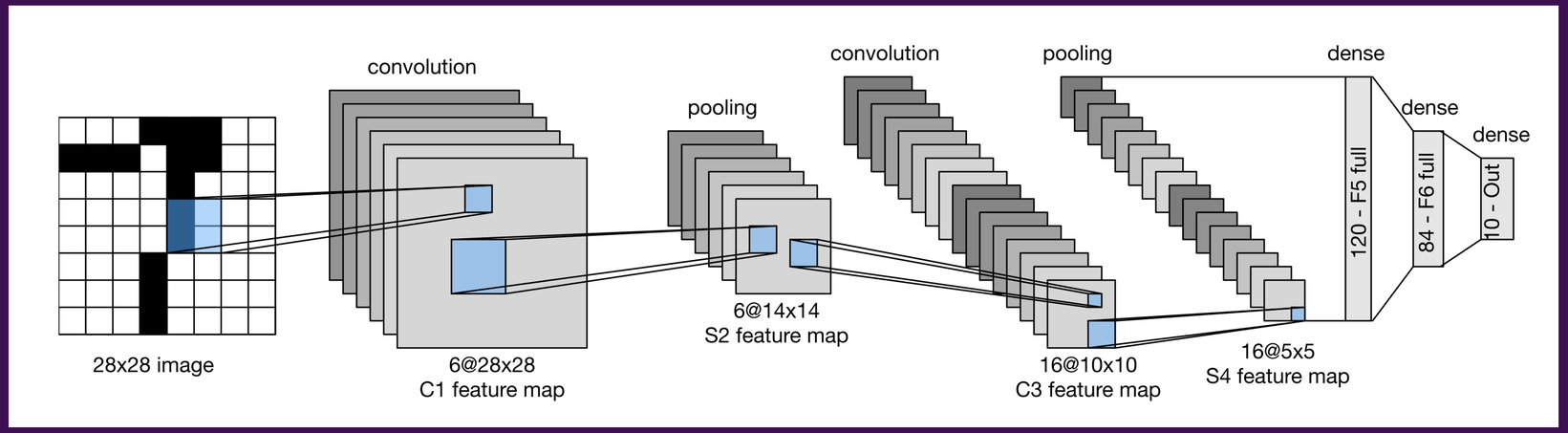
Machine Learning

Deep Learning

→ Programme
Paramètres d'un réseaux de neurones



ImageNet - 2009-2021



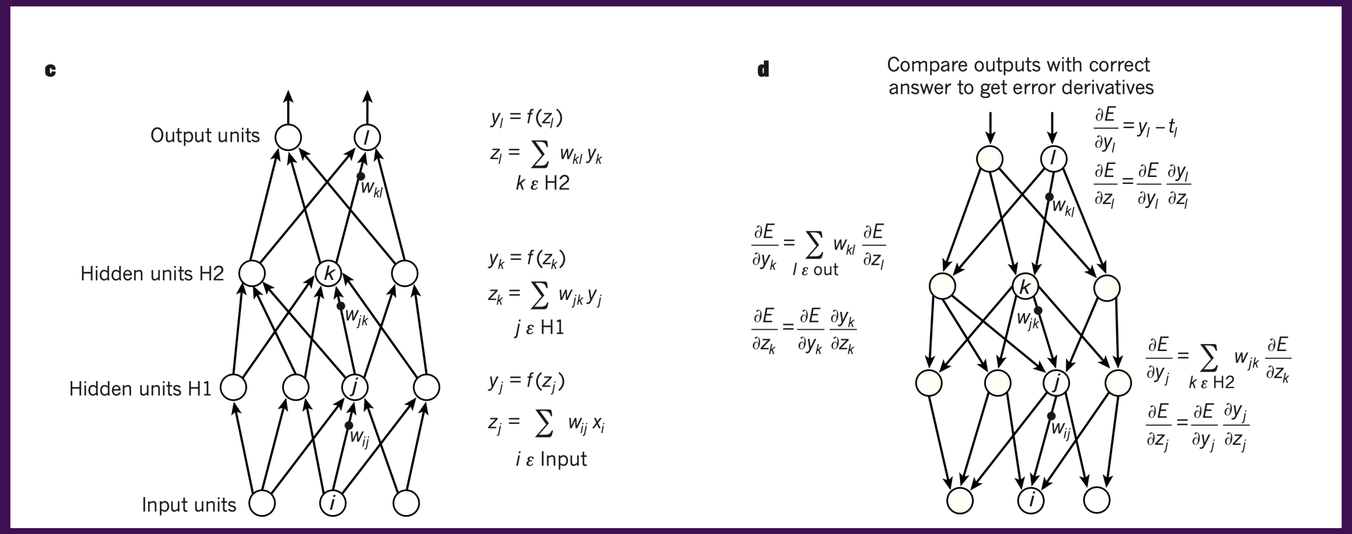
LeNet 1989 - Lecun - Représentation de Zhang, Aston and Lipton, Zachary C. and Li, Mu and Smola, Alexander J. 2023

Changement de paradigme



Chat

Deep Learning



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

Changement de paradigme



Chat





REVIEW

doi:10.1038/nature14539

Deep learning

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

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. Suppose we wish 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.

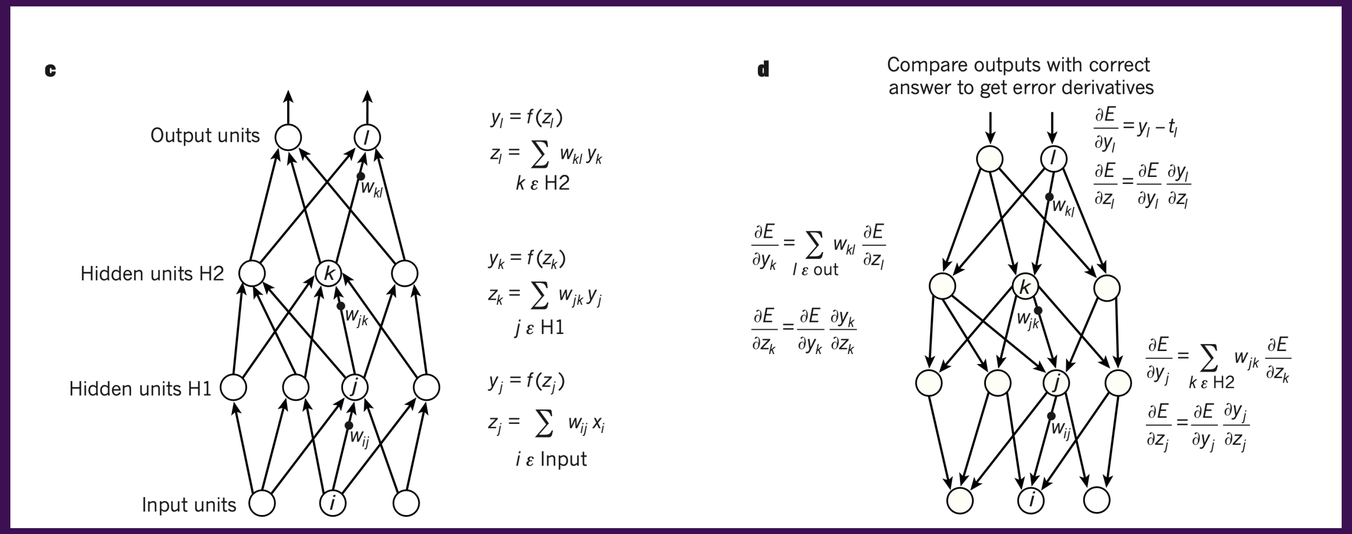
The objective function, averaged over all the training examples, can

[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 Andre-Aisenstadt, PO Box 6128 Centre-Ville STN Montreal, Quebec H3C 3J7, Canada. [4] Google, 1600 Amphitheatre Parkway, Mountain View, California 94043, USA. Department of Computer Science, University of Toronto, 6 King's College Road, Toronto, Ontario M5S 3G4, Canada.

436 | NATURE | VOL 521 | 28 MAY 2015

© 2015 Macmillan Publishers Limited. All rights reserved

Article cité plus de 100 000 fois ...

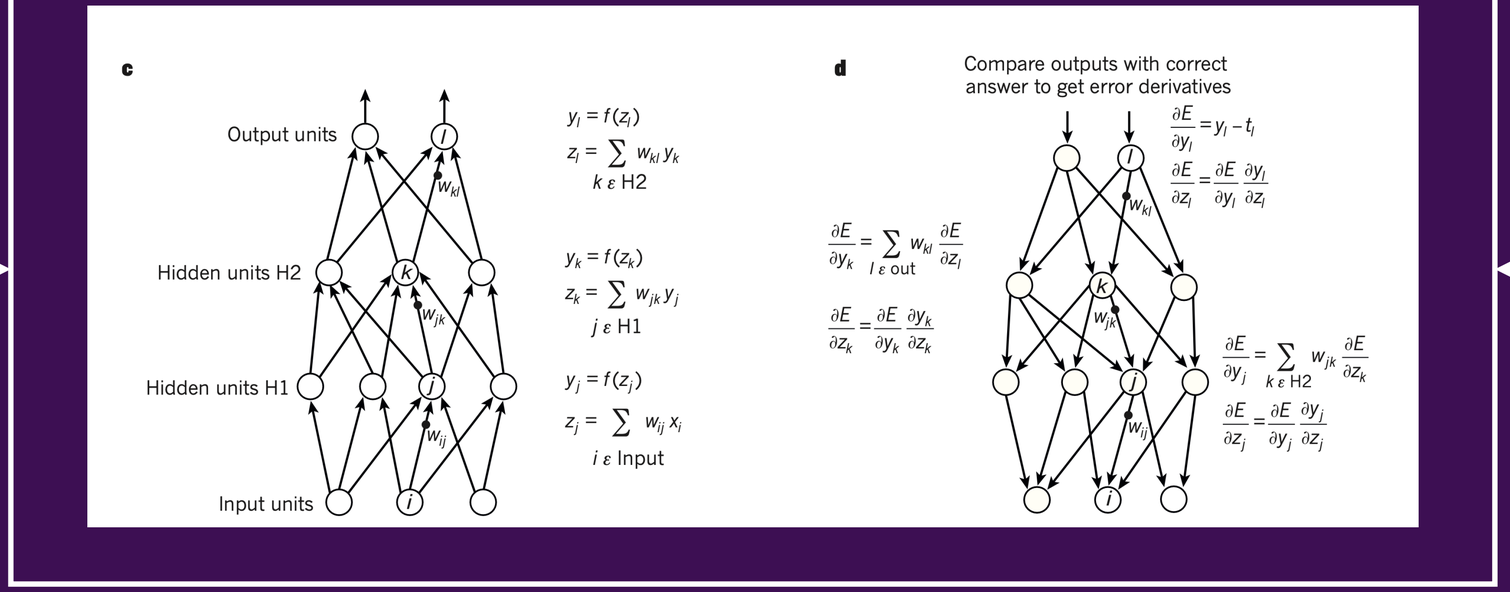


Yann LeCun, Yoshua Bengio & Geoffrey Hinton, Deep learning, Nature 2015 (vol.521), pp. 436-444.

Changement de paradigme

Apprentissage



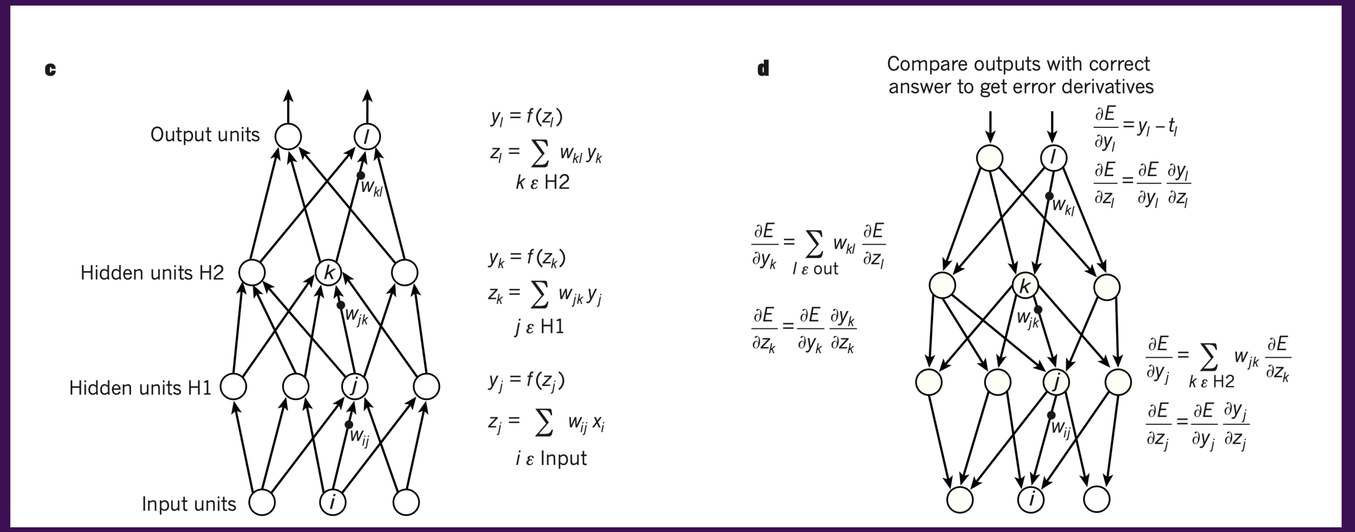


Chat

Changement de paradigme

Inférence



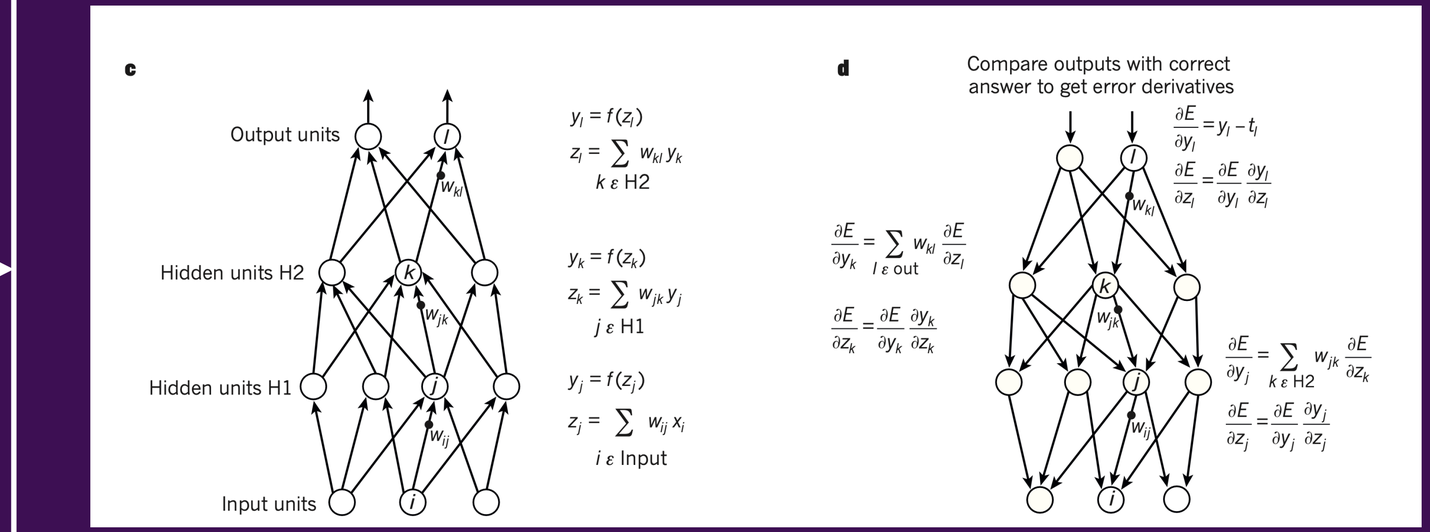


Chat

Changement de paradigme

Apprentissage

Chat →

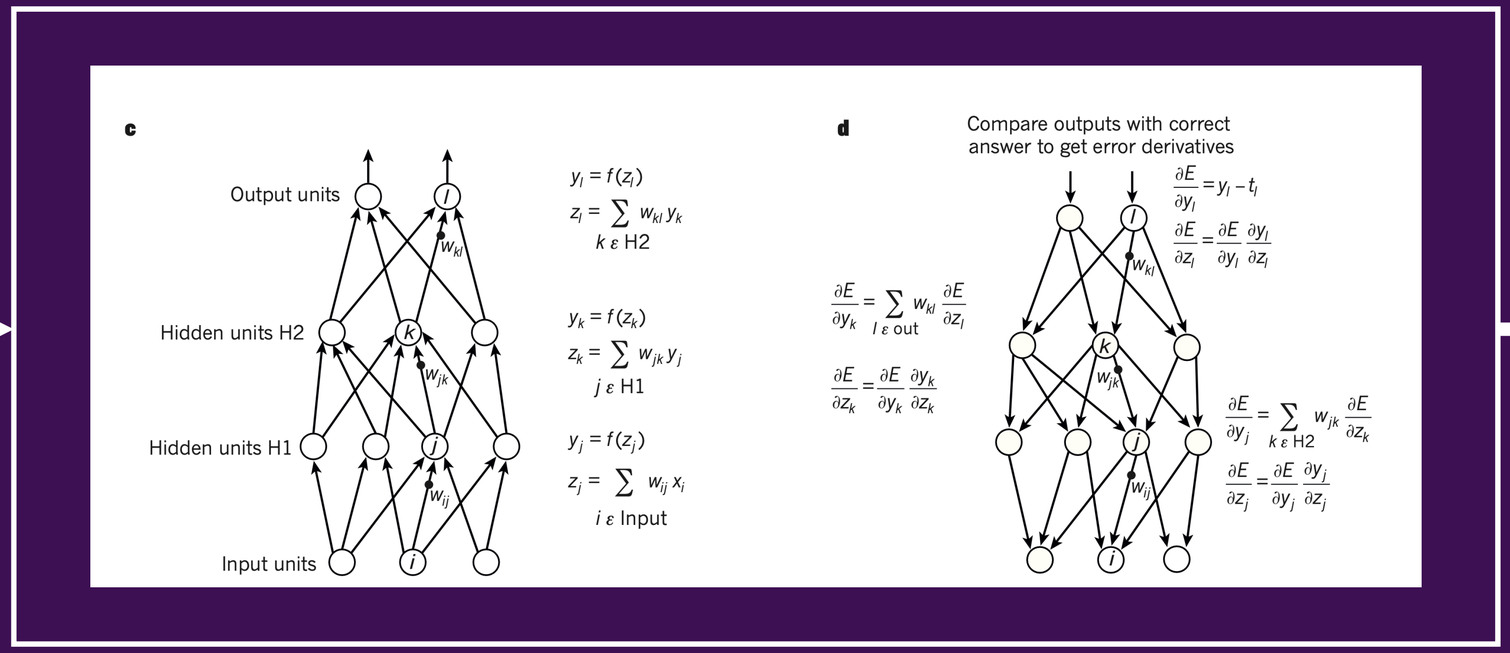


→ 

Changement de paradigme

Inférence

Chat →



→ 

L'importance du pluralisme

Données

Modèles

Harware / Data centers

Usages

Imaginaires

➜ **Le pluralisme sera de plus en plus politique**

D. Attention is all you need

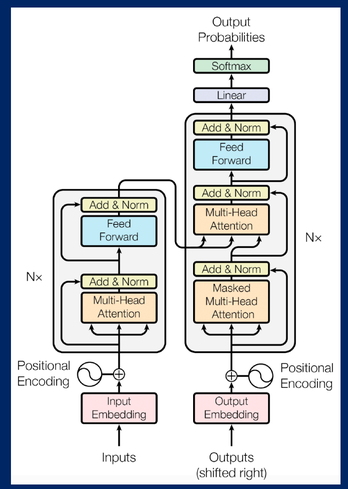
Changement de paradigme

Les LLM ont été inventés pour traduire

Et pour traduire correctement, il fallait « comprendre » le contexte…

Transformers

C'est un chat gris



It is a grey cat.

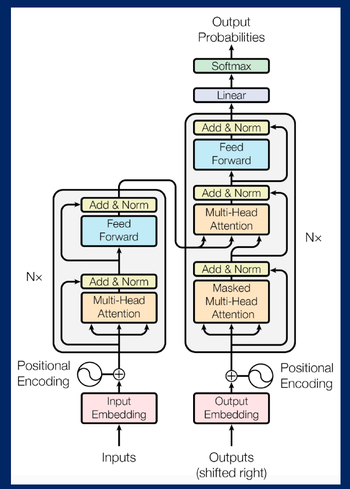
Apprentissage

Les LLM ont été inventés pour traduire

Et pour traduire correctement, il fallait « comprendre » le contexte…

C'est un chat gris

Transformers



Inférence

It is a grey cat.

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

Article cité plus de 50 000 fois ...

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.

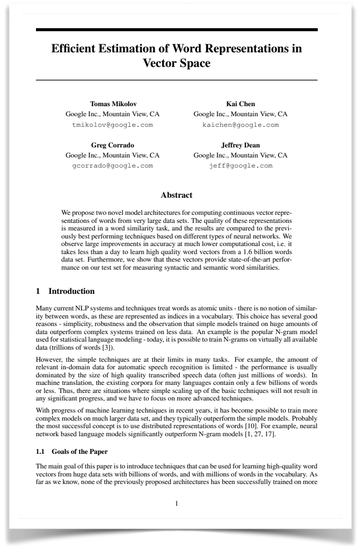
‡Work performed while at Google Research.

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

Article cité plus de 230 000 fois ...

2017

Embedding



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

1

2013

Transformers



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

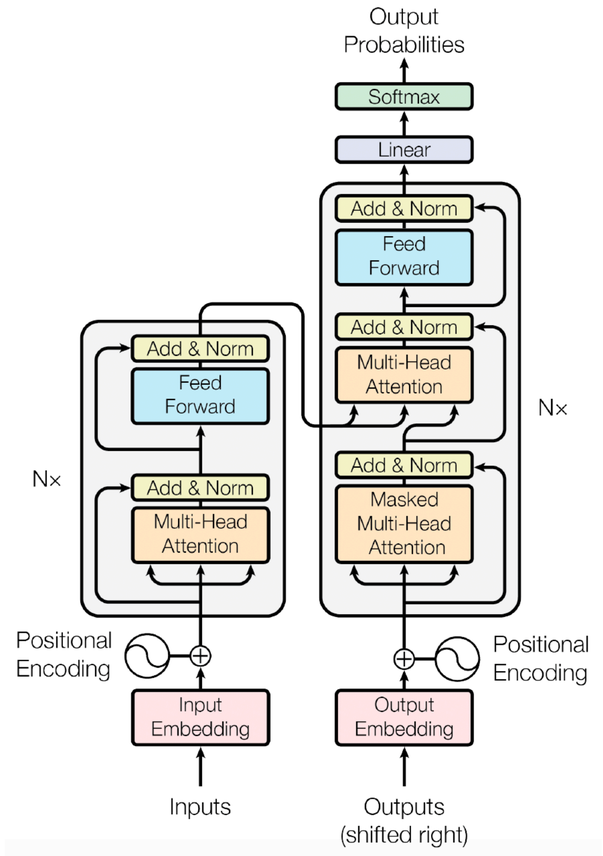
Abstract

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









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.

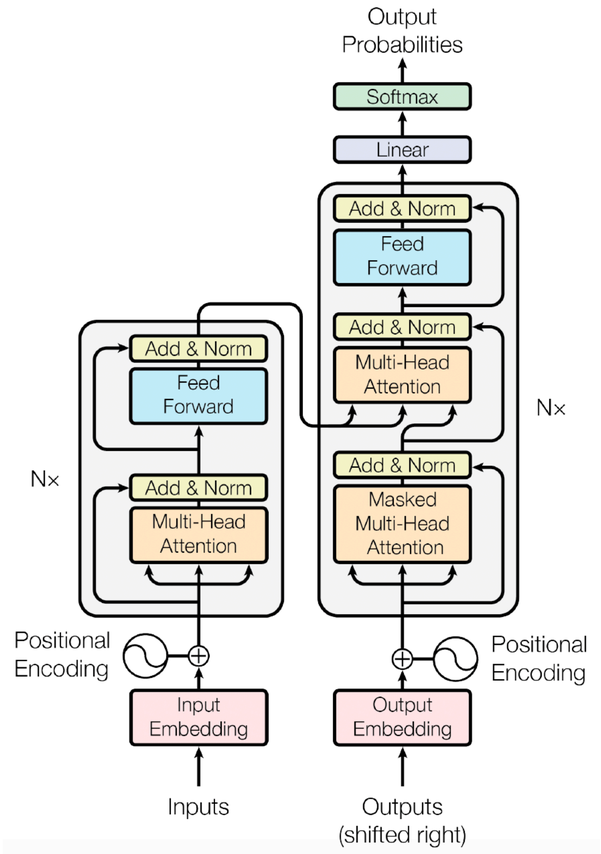
[*]Equal contribution. Listing order is random. Jakob proposed replacing RNNs with self-attention and started the effort to evaluate this idea. Ashish, with Illia, designed and implemented the first Transformer models and has been crucially involved in every aspect of this work. Noam proposed scaled 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



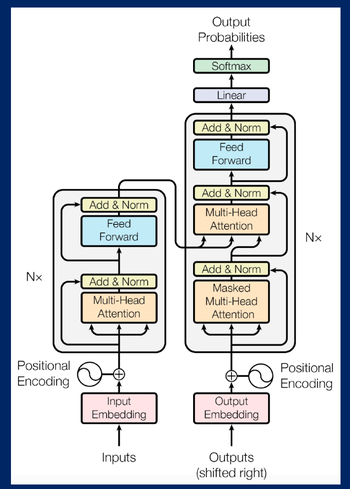
Architecture	Modèles représentatifs	Usage principal
Encodeur	ModernBERT, XLM-RoBERTa, E5, BGE, GTE	Classification, NER, embeddings, RAG
Décodeur	GPT-4o, Claude 3.7, Llama 3, Mistral, DeepSeek, Gemini	Génération, chat, raisonnement, code
Encodeur-décodeur	T5, BART, mT5, T5Gemma	Traduction, résumé, tâches seq2seq

Les LLM ont été inventés pour traduire

Et pour traduire correctement, il fallait « comprendre » le contexte…

C'est un chat gris

Transformers



Inférence

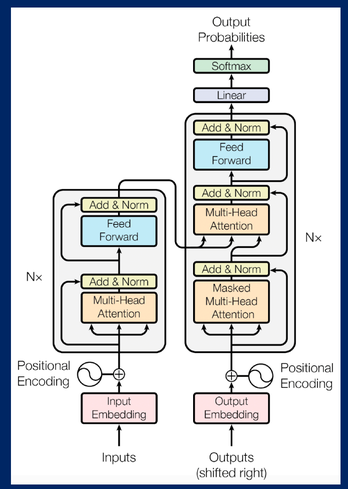
It is a grey cat.

Les LLM ont été inventés pour traduire

Et pour traduire correctement, il fallait « comprendre » le contexte…

The trophy doesn't fit in the suitcase because it is too big.

Transformers



Traduction

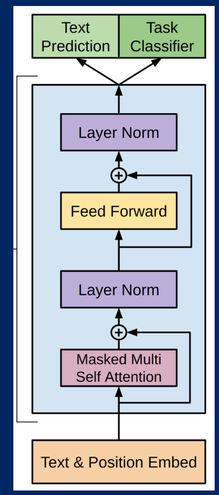
Le trophée ne rentre pas dans la valise parce qu'il est trop grand.

Les LLM ont été inventés pour traduire

Et pour traduire correctement, il fallait « comprendre » le contexte…

Generative Pretrained Transformers (GPT)

The trophy doesn't fit in the suitcase because it is too big



to close the zipper.

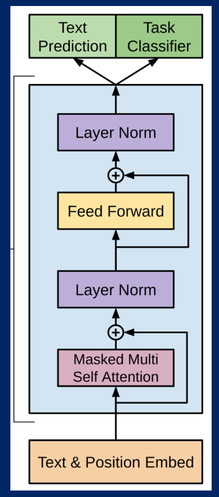
Prolongement de séquence → probabiliste ?

Les LLM ont été inventés pour traduire

Et pour traduire correctement, il fallait « comprendre » le contexte…

Traduis : The trophy doesn't fit in the suitcase because it is too big.

Generative Pretrained Transformers (GPT)



Le trophée ne rentre pas dans la valise parce qu'il est trop grand.

Généraliste : la traduction fait partie de la séquence → sens commun ?

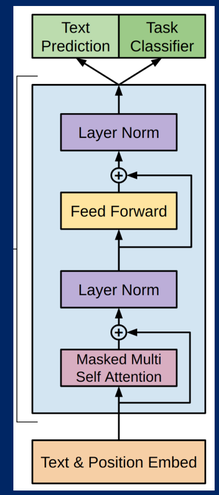
Boris Beaude - 15 septembre 2026

Les LLM ont été inventés pour traduire

Et pour traduire correctement, il fallait « comprendre » le contexte…

La surface occupée par les nénuphars double chaque jour. S'il faut 48 jours pour couvrir le lac, combien de jours faut-il pour en couvrir la moitié ?

Generative Pretrained Transformers (GPT)



Il faut 47 jours.

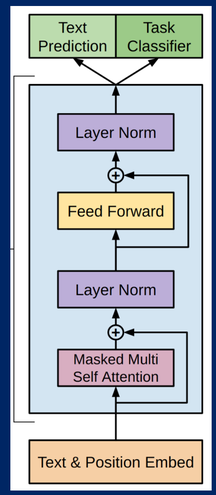
La prolongement de la séquence peut être la réponse à une problème → raisonnement ?

Les LLM ont été inventés pour traduire

Et pour traduire correctement, il fallait « comprendre » le contexte…

Generative Pretrained Transformers (GPT)

Quel incident dramatique a eu lieu récemment à Crans-Montana ?



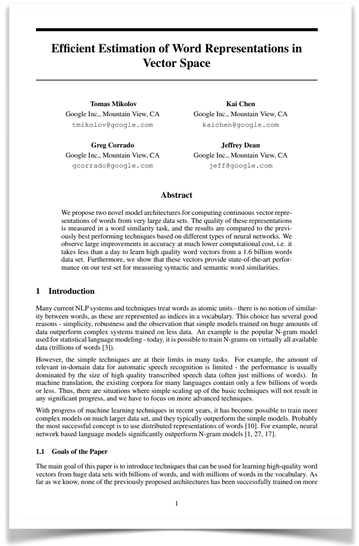
Confabulation ou refus de réponse

La prolongement peut être totalement spéculatif → knowledge cutoff

Unil.

Les compétences qui ont émergé n'étaient pas prévues par leurs concepteurs

Embedding



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

1

2013

Transformers



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

Article cité plus de 15 000 fois ...

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

juin 2018

Article cité plus de 150 000 fois ...

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).

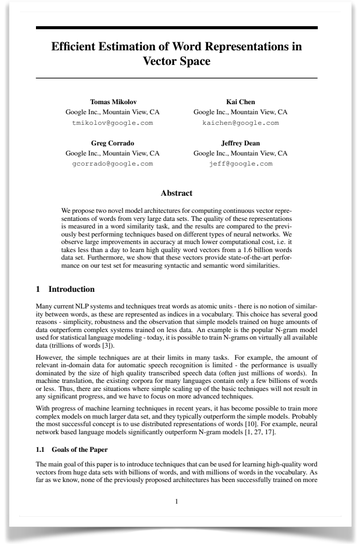
There are two existing strategies for applying pre-trained language representations to downstream tasks: *feature-based* and *fine-tuning*. The feature-based approach, such as ELMo (Peters et al., 2018a), uses task-specific architectures that include the pre-trained representations as additional features. 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

octobre 2018

Embedding



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

Transformers



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

GPT



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 ilyas@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

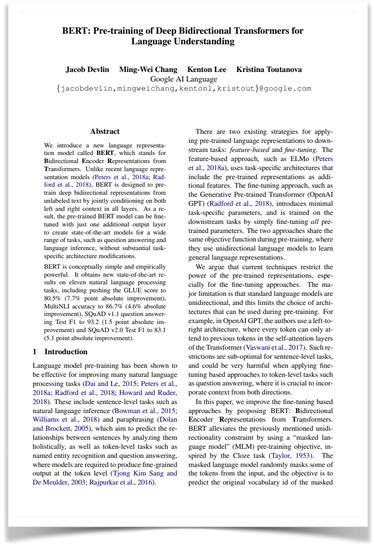
The ability to learn effectively from raw text is crucial to alleviating the dependence on supervised learning in natural language processing (NLP). Most deep learning methods require substantial amounts of manually labeled data, which restricts their applicability in many domains that suffer from a 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

Bert



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

Article cité plus de 65 000 fois ...

GPT 3



Language Models are Few-Shot Learners

Tom B. Brown[*] **Benjamin Mann**[*] **Nick Ryder**[*] **Melanie Subbiah**[*]
Jared Kaplan[†] **Prafulla Dhariwal** **Arvind Neelakantan** **Pranav Shyam** **Girish Sastry**
Amanda Askell **Sandhini Agarwal** **Ariel Herbert-Voss** **Gretchen Krueger** **Tom Henighan**
Rewon Child **Aditya Ramesh** **Daniel M. Ziegler** **Jeffrey Wu** **Clemens Winter**
Christopher Hesse **Mark Chen** **Eric Sigler** **Mateusz Litwin** **Scott Gray**
Benjamin Chess **Jack Clark** **Christopher Berner**
Sam McCandlish **Alec Radford** **Ilya Sutskever** **Dario Amodei**
OpenAI

Abstract

Recent work has demonstrated substantial gains on many NLP tasks and benchmarks by pre-training on a large corpus of text followed by fine-tuning on a specific task. While typically task-agnostic in architecture, this method still requires task-specific fine-tuning datasets of thousands or tens of thousands of examples. By contrast, humans can generally perform a new language task from only a few examples or from simple instructions – something which current NLP systems still largely struggle to do. Here we show that scaling up language models greatly improves task-agnostic, few-shot performance, sometimes even reaching competitiveness with prior state-of-the-art fine-tuning approaches. Specifically, we train GPT-3, an autoregressive language model with 175 billion parameters, 10x more than any previous non-sparse language model, and test its performance in the few-shot setting. For all tasks, GPT-3 is applied without any gradient updates or fine-tuning, with tasks and few-shot demonstrations specified purely via text interaction with the model. GPT-3 achieves strong performance on many NLP datasets, including translation, question-answering, and cloze tasks, as well as several tasks that require on-the-fly reasoning or domain adaptation, such as unscrambling words, using a novel word in a sentence, or performing 3-digit arithmetic. At the same time, we also identify some datasets where GPT-3's few-shot learning still struggles, as well as some datasets where GPT-3 faces methodological issues related to training on large web corpora. Finally, we find that GPT-3 can generate samples of news articles which human evaluators have difficulty distinguishing from articles written by humans. We discuss broader societal impacts of this finding and of GPT-3 in general.

[*]Equal contribution
[†]Johns Hopkins University, OpenAI

Author contributions listed at end of paper.

2020

Fev. 2019

GPT2 small

124 millions de paramètres

Nov. 2019

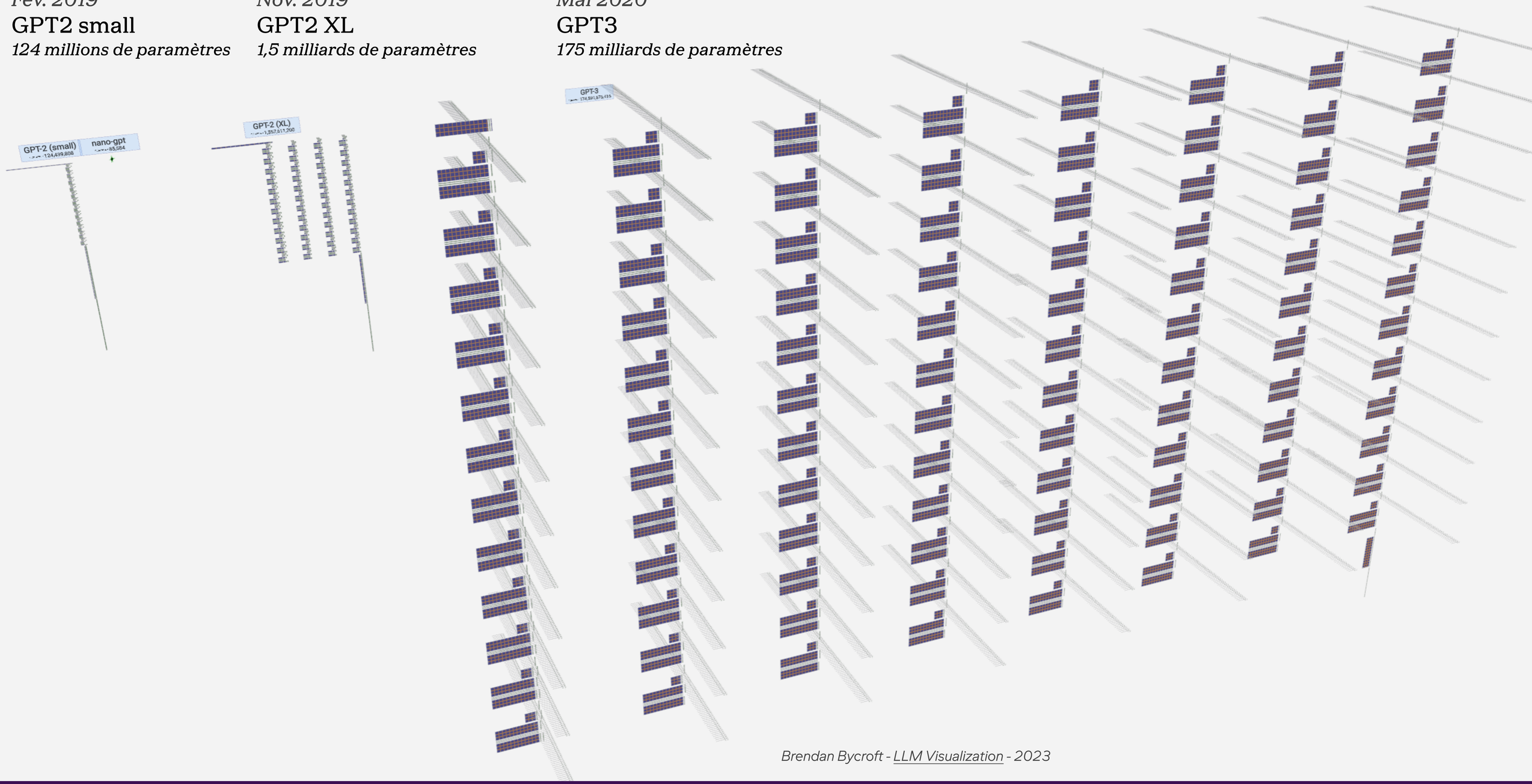
GPT2 XL

1,5 milliards de paramètres

Mai 2020

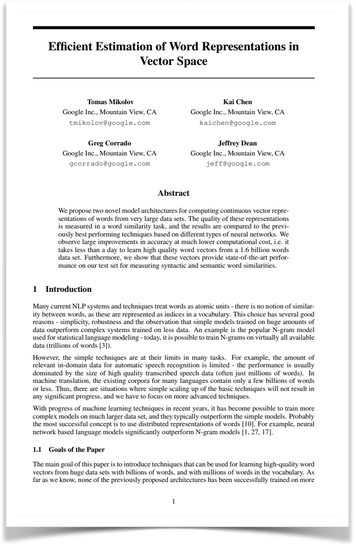
GPT3

175 milliards de paramètres



Brendan Bycroft - LLM Visualization - 2023

Embedding



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

Transformers



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

GPT



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 ilyas@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

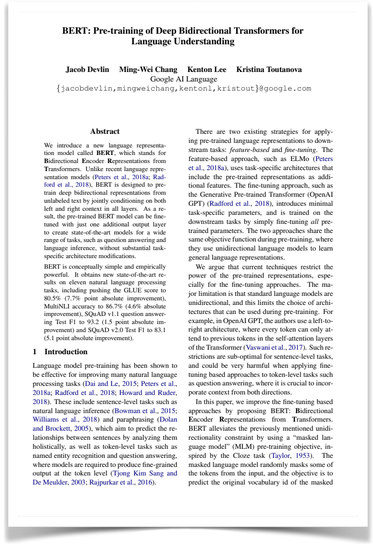
The ability to learn effectively from raw text is crucial to alleviating the dependence on supervised learning in natural language processing (NLP). Most deep learning methods require substantial amounts of manually labeled data, which restricts their applicability in many domains that suffer from a 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

Bert



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.

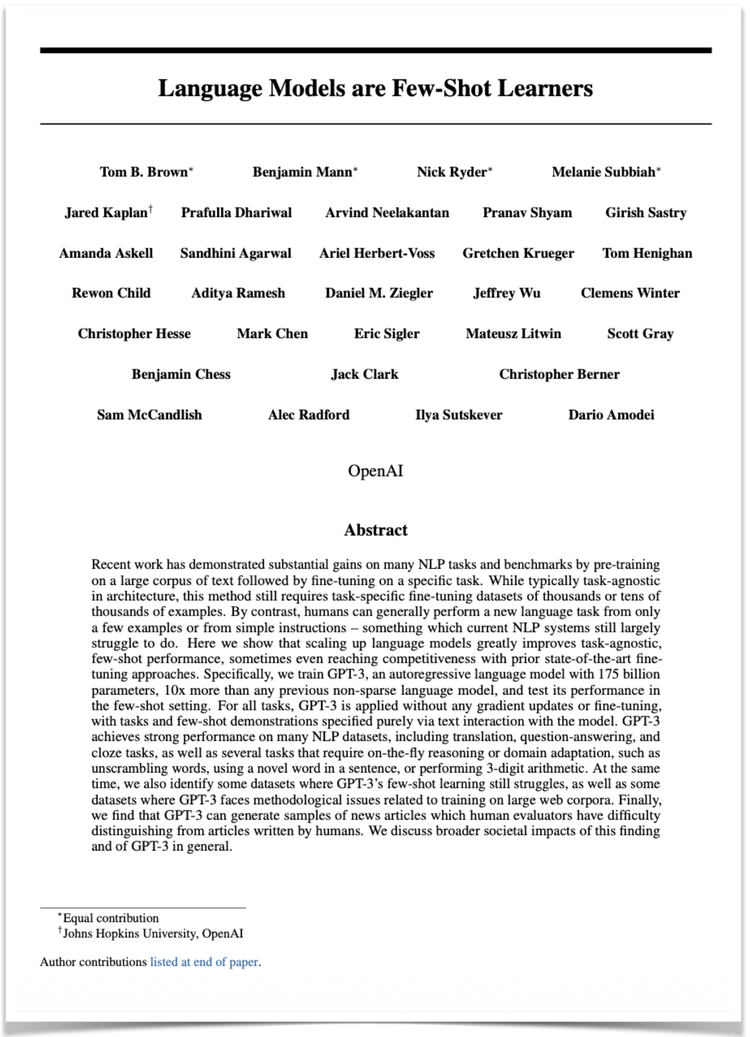
We argue that current techniques restrict the power of the pre-trained representations, especially for the fine-tuning approaches. The major limitation is that standard language models are unidirectional, and this limits the choice of architectures that can be used during pre-training. For example, in 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

Article cité plus de 65 000 fois ...

GPT 3



Language Models are Few-Shot Learners

Tom B. Brown[*] **Benjamin Mann**[*] **Nick Ryder**[*] **Melanie Subbiah**[*]
Jared Kaplan[†] **Prafulla Dhariwal** **Arvind Neelakantan** **Pranav Shyam** **Girish Sastry**
Amanda Askell **Sandhini Agarwal** **Ariel Herbert-Voss** **Gretchen Krueger** **Tom Henighan**
Rewon Child **Aditya Ramesh** **Daniel M. Ziegler** **Jeffrey Wu** **Clemens Winter**
Christopher Hesse **Mark Chen** **Eric Sigler** **Mateusz Litwin** **Scott Gray**
Benjamin Chess **Jack Clark** **Christopher Berner**
Sam McCandlish **Alec Radford** **Ilya Sutskever** **Dario Amodei**
OpenAI

Abstract

Recent work has demonstrated substantial gains on many NLP tasks and benchmarks by pre-training on a large corpus of text followed by fine-tuning on a specific task. While typically task-agnostic in architecture, this method still requires task-specific fine-tuning datasets of thousands or tens of thousands of examples. By contrast, humans can generally perform a new language task from only a few examples or from simple instructions – something which current NLP systems still largely struggle to do. Here we show that scaling up language models greatly improves task-agnostic, few-shot performance, sometimes even reaching competitiveness with prior state-of-the-art fine-tuning approaches. Specifically, we train GPT-3, an autoregressive language model with 175 billion parameters, 10x more than any previous non-sparse language model, and test its performance in the few-shot setting. For all tasks, GPT-3 is applied without any gradient updates or fine-tuning, with tasks and few-shot demonstrations specified purely via text interaction with the model. GPT-3 achieves strong performance on many NLP datasets, including translation, question-answering, and cloze tasks, as well as several tasks that require on-the-fly reasoning or domain adaptation, such as unscrambling words, using a novel word in a sentence, or performing 3-digit arithmetic. At the same time, we also identify some datasets where GPT-3's few-shot learning still struggles, as well as some datasets where GPT-3 faces methodological issues related to training on large web corpora. Finally, we find that GPT-3 can generate samples of news articles which human evaluators have difficulty distinguishing from articles written by humans. We discuss broader societal impacts of this finding and of GPT-3 in general.

[*]Equal contribution
[†]Johns Hopkins University, OpenAI

Author contributions listed at end of paper.

2020

Embedding



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

Transformers



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

GPT



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

Bert



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

GPT 3



Language Models are Few-Shot Learners

Tom B. Brown* Benjamin Mann* Nick Ryder* Melanie Subbiah*
Jared Kaplan† Prafulla Dhariwal Arvind Neelakantan Pranav Shyam Girish Sastry
Amanda Askell Sandhini Agarwal Ariel Herbert-Voss Gretchen Krueger Tom Henighan
Rewon Child Aditya Ramesh Daniel M. Ziegler Jeffrey Wu Clemens Winter
Christopher Hesse Mark Chen Eric Sigler Mateusz Litwin Scott Gray
Benjamin Chess Jack Clark Christopher Berner
Sam McCandlish Alec Radford Ilya Sutskever Dario Amodei
OpenAI

Abstract

Recent work has demonstrated substantial gains on many NLP tasks and benchmarks by pre-training on a large corpus of text followed by fine-tuning on a specific task. While typically task-agnostic in architecture, this method still requires task-specific fine-tuning datasets of thousands or tens of thousands of examples. By contrast, humans can generally perform a new language task from only a few examples or from simple instructions – something which current NLP systems still largely struggle to do. Here we show that scaling up language models greatly improves task-agnostic, few-shot performance, sometimes even reaching competitiveness with prior state-of-the-art fine-tuning approaches. Specifically, we train GPT-3, an autoregressive language model with 175 billion parameters, 10x more than any previous non-sparse language model, and test its performance in the few-shot setting. For all tasks, GPT-3 is applied without any gradient updates or fine-tuning, with tasks and few-shot demonstrations specified purely via text interaction with the model. GPT-3 achieves strong performance on many NLP datasets, including translation, question-answering, and cloze tasks, as well as several tasks that require on-the-fly reasoning or domain adaptation, such as unscrambling words, using a novel word in a sentence, or performing 3-digit arithmetic. At the same time, we also identify some datasets where GPT-3's few-shot learning still struggles, as well as some datasets where GPT-3 faces methodological issues related to training on large web corpora. Finally, we find that GPT-3 can generate samples of news articles which human evaluators have difficulty distinguishing from articles written by humans. We discuss broader societal impacts of this finding and of GPT-3 in general.

*Equal contribution
†Johns Hopkins University, OpenAI

Author contributions listed at end of paper.

2020

Article cité plus de 20 000 fois ...

RLHF

Reinforcement learning from human feedback

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

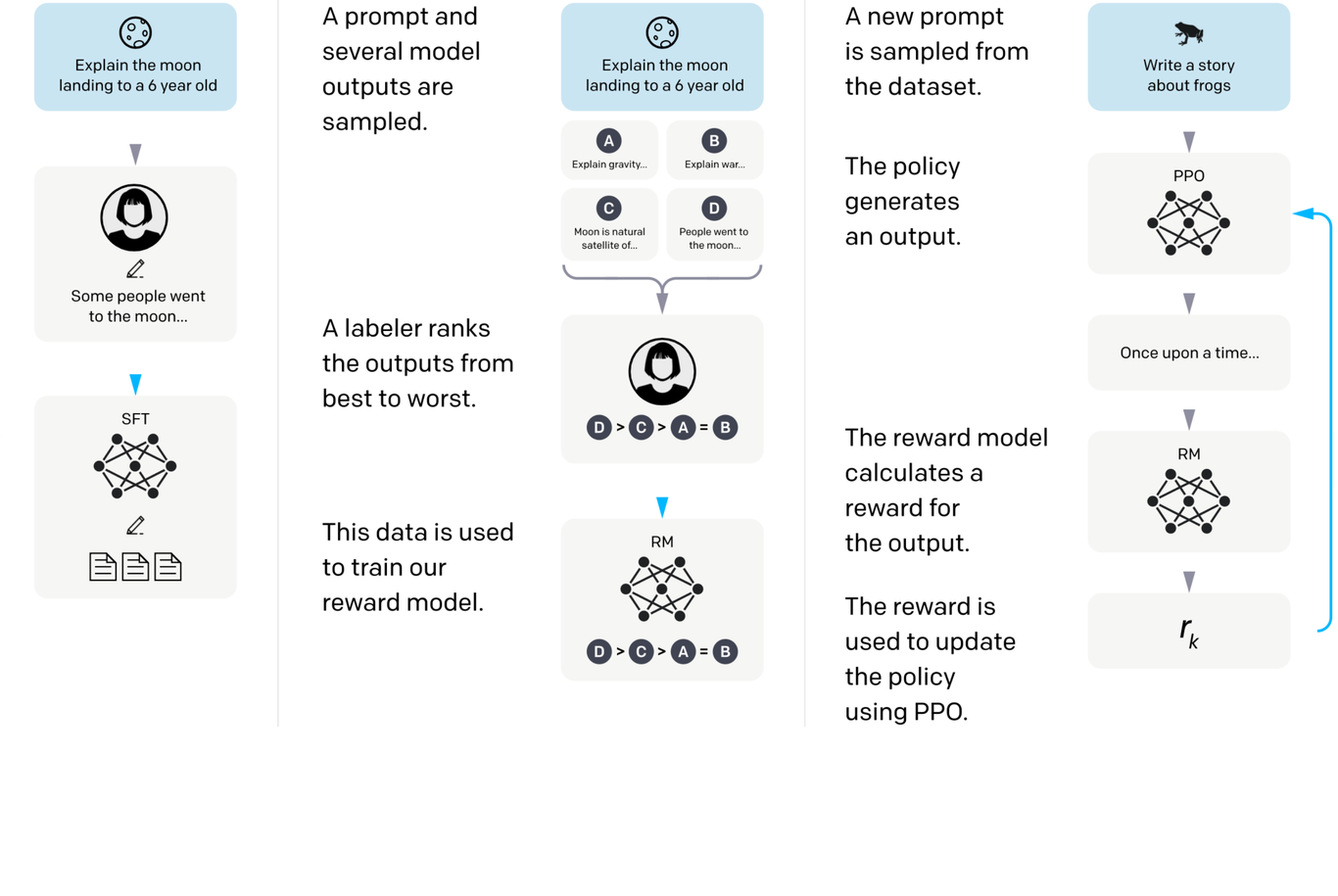
Making language models bigger does not inherently make them better at following a user's intent. For example, large language models can generate outputs that are untruthful, toxic, or simply not helpful to the user. In other words, these models are not *aligned* with their users. In this paper, we show 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



Embedding



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

Transformers



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

Lukasz 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

GPT



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

Bert



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

GPT 3



Language Models are Few-Shot Learners

Tom B. Brown* Benjamin Mann* Nick Ryder* Melanie Subbiah*
Jared Kaplan† Prafulla Dhariwal Arvind Neelakantan Pranav Shyam Girish Sastry
Amanda Askell Sandhini Agarwal Ariel Herbert-Voss Gretchen Krueger Tom Henighan
Rewon Child Aditya Ramesh Daniel M. Ziegler Jeffrey Wu Clemens Winter
Christopher Hesse Mark Chen Eric Sigler Mateusz Litwin Scott Gray
Benjamin Chess Jack Clark Christopher Berner
Sam McCandlish Alec Radford Ilya Sutskever Dario Amodei
OpenAI

Abstract

Recent work has demonstrated substantial gains on many NLP tasks and benchmarks by pre-training on a large corpus of text followed by fine-tuning on a specific task. While typically task-agnostic in architecture, this method still requires task-specific fine-tuning datasets of thousands or tens of thousands of examples. By contrast, humans can generally perform a new language task from only a few examples or from simple instructions – something which current NLP systems still largely struggle to do. Here we show that scaling up language models greatly improves task-agnostic, few-shot performance, sometimes even reaching competitiveness with prior state-of-the-art fine-tuning approaches. Specifically, we train GPT-3, an autoregressive language model with 175 billion parameters, 10x more than any previous non-sparse language model, and test its performance in the few-shot setting. For all tasks, GPT-3 is applied without any gradient updates or fine-tuning, with tasks and few-shot demonstrations specified purely via text interaction with the model. GPT-3 achieves strong performance on many NLP datasets, including translation, question-answering, and cloze tasks, as well as several tasks that require on-the-fly reasoning or domain adaptation, such as unscrambling words, using a novel word in a sentence, or performing 3-digit arithmetic. At the same time, we also identify some datasets where GPT-3's few-shot learning still struggles, as well as some datasets where GPT-3 faces methodological issues related to training on large web corpora. Finally, we find that GPT-3 can generate samples of news articles which human evaluators have difficulty distinguishing from articles written by humans. We discuss broader societal impacts of this finding and of GPT-3 in general.

*Equal contribution
†Johns Hopkins University, OpenAI

Author contributions listed at end of paper.

2020

RLHF



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

[1]Primary authors. This was a joint project of the OpenAI Alignment team. RL and JL are the team leads.
[2]Author order was randomized. See Appendix F.
[3]Work done while at OpenAI. Current affiliations: AA: Anthropic; PC: Alignment Research Center.

2022

RLAIF

Reinforcement learning from AI feedback

Constitutional AI: Harmlessness from AI Feedback

Yuntao Bai[1], Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion,

Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon,
Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain,
Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller,
Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt,
Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma,
Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec,
Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly,
Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann,
Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, Jared Kaplan*

Anthropic

Abstract

As AI systems become more capable, we would like to enlist their help to supervise other AIs. We experiment with methods for training a harmless AI assistant through self-improvement, without any human labels identifying harmful outputs. The only human oversight is provided through a list of rules or principles, and so we refer to the method as 'Constitutional AI'. The process involves both a supervised learning and a reinforcement learning phase. In the supervised phase we sample from an initial model, then generate self-critiques and revisions, and then finetune the original model on revised responses. In the RL phase, we sample from the finetuned model, use a model to evaluate which of the two samples is better, and then train a preference model from this dataset of AI preferences. We then train with RL using the preference model as the reward signal, i.e. we use 'RL from AI Feedback' (RLAIF). As a result we are able to train a harmless but non-evasive AI assistant that engages with harmful queries by explaining its objections to them. Both the SL and RL methods can leverage chain-of-thought style reasoning to improve the human-judged performance and transparency of AI decision making. These methods make it possible to control AI behavior more precisely and with far fewer human labels.

*Correspondence to: {yuntao,jared}@anthropic.com
Author contributions are detailed in 7.

2022

Constitutional AI

Embedding



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

Transformers



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

GPT



Improving Language Understanding by Generative Pre-Training

Alec Radford
OpenAI
alec@openai.com

Karthik Narasimhan
OpenAI
karthik@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 unsupervised learning, when supervision is available, can be very beneficial. It 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. 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

Bert



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 Bidirectional Encoder Representations from Transformers. 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: Bidirectional Encoder Representations from Transformers. 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

GPT 3



Language Models are Few-Shot Learners

Tom B. Brown* Benjamin Mann* Nick Ryder* Melanie Subbiah*
Jared Kaplan† Prafulla Dhariwal Arvind Neelakantan Pranav Shyam Girish Sastry
Amanda Askell Sandhini Agarwal Ariel Herbert-Voss Gretchen Krueger Tom Henighan
Rewon Child Aditya Ramesh Daniel M. Ziegler Jeffrey Wu Clemens Winter
Christopher Hesse Mark Chen Eric Sigler Mateusz Litwin Scott Gray
Benjamin Chess Jack Clark Christopher Berner
Sam McCandlish Alec Radford Ilya Sutskever Dario Amodei
OpenAI

Abstract

Recent work has demonstrated substantial gains on many NLP tasks and benchmarks by pre-training on a large corpus of text followed by fine-tuning on a specific task. While typically task-agnostic in architecture, this method still requires task-specific fine-tuning datasets of thousands or tens of thousands of examples. By contrast, humans can generally perform a new language task from only a few examples or from simple instructions – something which current NLP systems still largely struggle to do. Here we show that scaling up language models greatly improves task-agnostic, few-shot performance, sometimes even reaching competitiveness with prior state-of-the-art fine-tuning approaches. Specifically, we train GPT-3, an autoregressive language model with 175 billion parameters, 10x more than any previous non-sparse language model, and test its performance in the few-shot setting. For all tasks, GPT-3 is applied without any gradient updates or fine-tuning, with tasks and few-shot demonstrations specified purely via text interaction with the model. GPT-3 achieves strong performance on many NLP datasets, including translation, question-answering, and cloze tasks, as well as several tasks that require on-the-fly reasoning or domain adaptation, such as unscrambling words, using a novel word in a sentence, or performing 3-digit arithmetic. At the same time, we also identify some datasets where GPT-3's few-shot learning still struggles, as well as some datasets where GPT-3 faces methodological issues related to training on large web corpora. Finally, we find that GPT-3 can generate samples of news articles which human evaluators have difficulty distinguishing from articles written by humans. We discuss broader societal impacts of this finding and of GPT-3 in general.

*Equal contribution
†Johns Hopkins University, OpenAI

Author contributions listed at end of paper.

2020

RLHF



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

RLAIF



Constitutional AI: Harmlessness from AI Feedback

Yuntao Bai* Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion,
Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon,
Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain,
Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller,
Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt,
Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma,
Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec,
Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly,
Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann,
Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, Jared Kaplan*
Anthropic

Abstract

As AI systems become more capable, we would like to enlist their help to supervise other AIs. We experiment with methods for training a harmless AI assistant through self-improvement, without any human labels identifying harmful outputs. The only human oversight is provided through a list of rules or principles, and so we refer to the method as 'Constitutional AI'. The process involves both a supervised learning and a reinforcement learning phase. In the supervised phase we sample from an initial model, then generate self-critiques and revisions, and then finetune the original model on revised responses. In the RL phase, we sample from the finetuned model, use a model to evaluate which of the two samples is better, and then train a preference model from this dataset of AI preferences. We then train with RL using the preference model as the reward signal, i.e. we use 'RL from AI Feedback' (RLAIF). As a result we are able to train a harmless but non-evasive AI assistant that engages with harmful queries by explaining its objections to them. Both the SL and RL methods can leverage chain-of-thought style reasoning to improve the human-judged performance and transparency of AI decision making. These methods make it possible to control AI behavior more precisely and with far fewer human labels.

*Correspondence to: {yuntao,jared}@anthropic.com
Author contributions are detailed in §7.

2022

ANTHROP\C

Research Policy Commitments Learn News Try Claude



Claude's Constitution

Our vision for Claude's character

Claude's constitution is a detailed description of Anthropic's intentions for Claude's values and behavior. It plays a crucial role in our training process, and its content directly shapes Claude's behavior. It's also the final authority on our vision for Claude, and our aim is for all of our other guidance and training to be consistent with it.

Training models is a difficult task, and Claude's behavior might not always reflect the constitution's ideals. We will be open—for example, in our system cards—about the ways in which Claude's behavior comes apart from our intentions. But we think transparency about those intentions is important regardless.

The document is written with Claude as its primary audience, so it might read differently than you'd expect. For example, it's optimized for precision over accessibility, and it covers various topics that may be of less interest to human readers. We also discuss Claude in terms normally reserved for humans (e.g.,

Listen to the audiobook

Claude's Constitution - consulté le 10 septembre 2026

Model system cards

System cards document the capabilities, safety evaluations, and responsible deployment decisions for Claude models.

Model	Date	System card
Claude Fable 5.1 and Mythos 5.1	September 2026	Read system card
Claude Opus 5	July 2026	Read system card
Claude Sonnet 5	June 2026	Read system card
Claude Opus 4.8	May 2026	Read system card
Claude Fable 5 and Mythos 5	June 2026	Read system card
Claude Opus 4.7	April 2026	Read system card
Mythos Preview	April 2026	Read system card
Claude Sonnet 4.6	February 2026	Read system card
Claude Opus 4.6	February 2026	Read system card
Claude Opus 4.5	November 2025	Read system card
Claude Haiku 4.5	October 2025	Read system card
Claude Sonnet 4.5	September 2025	Read system card
Claude Opus 4.1	August 2025	Read system card
Claude Sonnet 4 and Opus 4	May 2025	Read system card
Claude Sonnet 3.7	February 2025	Read system card
Claude Haiku 3.5 and Sonnet 3.5 (new)	October 2024	Read system card
Claude Sonnet 3.5	June 2024	Read system card
Claude 3	March 2024	Read system card
Claude 2	July 2023	Read system card

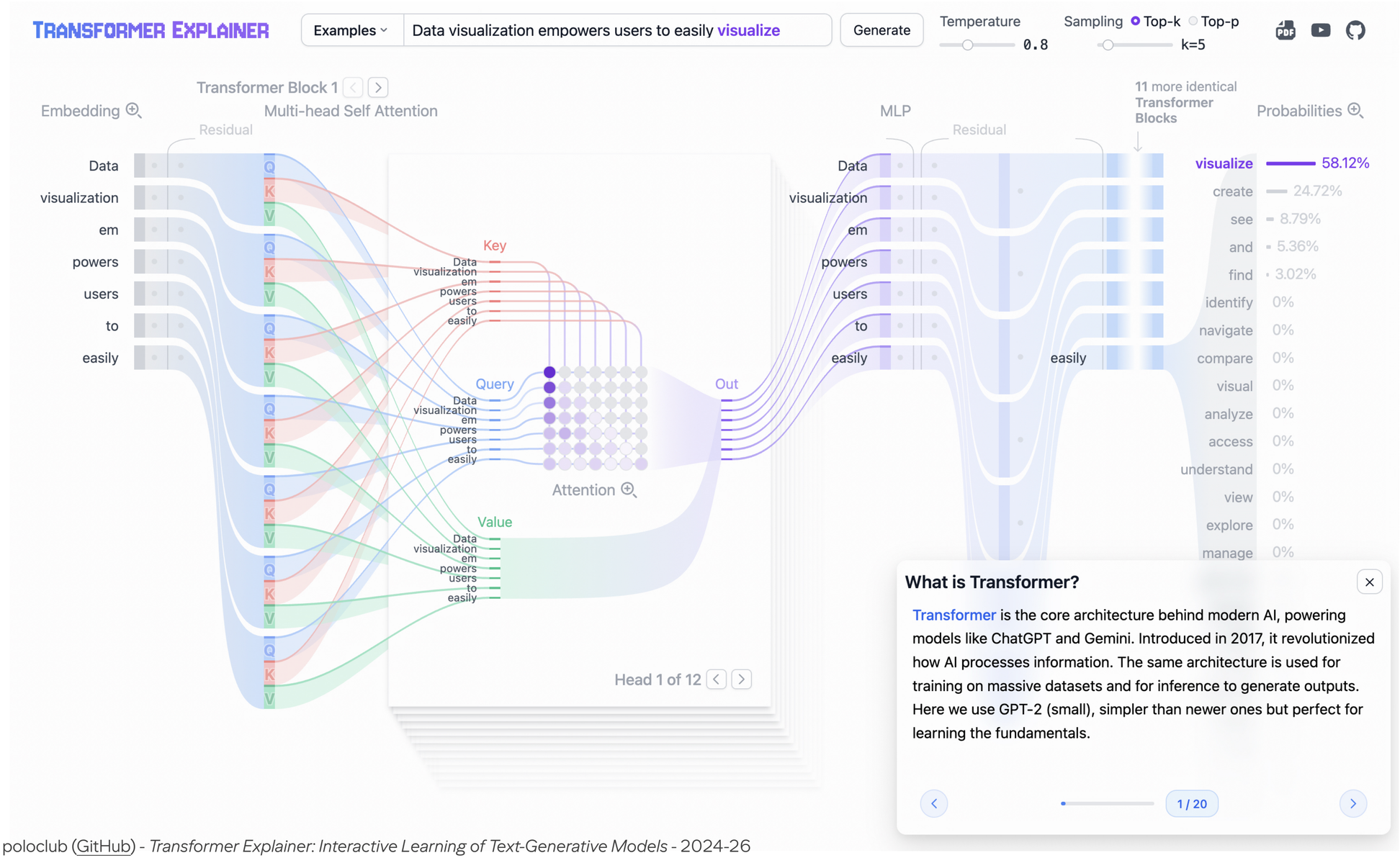
Anthropic Model system cards

	Anthropic	OpenAI	Google DeepMind	Meta	Z.ai	Alibaba Cloud	xAI	DeepSeek	Mistral
Overall Grade	C+	C	C	D+	D-	D-	F	F	F
Score	2.66	2.28	2.01	1.32	0.88	0.87	0.65	0.47	0.33
Grade Trend (Winter 25)	C+	C+ ▼	C	D ▲	D ▼	D-	D ▼	D ▼	N/A
Domains (Hint: Click on a domain to inspect)									
Risk Assessment (6 indicators)	C+	C+	C+	D+	F	F	D-	F	F
Current Harms (9 indicators)	B-	C	C	D-	C-	C-	F	D-	F
Safety Frameworks (4 indicators)	B-	C+	C	C-	D-	D-	D	F	F
Existential Safety (4 indicators)	D+	D+	D	F	F	F	F	F	F
Governance & Accountability (4 indicators)	B	C	C-	D+	F	D-	F	F	F
Information Sharing (10 indicators)	B+	B-	B-	D+	D	D	D	D-	D-
Survey Responses	✓	✓	✓	✓	✓	–	–	–	–

Grading: We use the US GPA system for grade boundaries: A, B, C, D, F letter values correspond to numerical values 4.0, 3.0, 2.0, 1.0, 0.

How are scores calculated? | Chinese Regulatory Context | Save

AI Safety Index - Future of Life Institute - Juillet. 2026



poloclub (GitHub) - *Transformer Explainer: Interactive Learning of Text-Generative Models* - 2024-26

Les LLM ne comprennent pas pour prédire.

Les LLM prédisent pour « comprendre ».

Lorsqu’on utilise un LLM, le mot suivant n’existe pas encore.
Chaque mot n'est qu'une supposition. C'est ce qui rend la réussite déconcertante et la créativité possible...

Boris Beaude - 15 septembre 2026

Au delà de la prédiction...

Multimodalité textes, sons et images

→ Interaction plus large et « modèles de mondes plus complexes »

Génération à enrichissement contextuel Retrieval-Augmented Generation - RAG

→ Adaptation à des documents précis

Recherche

→ Actualisation des données utilisées

Thinking Chain-of-Thought et apprentissage de raisonnements

→ Capacité de « raisonnement » exploratoire et structuré

Tools calcul, code...

→ Utilisation d'outils spécifiques lorsque c'est plus efficace

Agents

→ Décomposition d'une tâche en sous-tâches spécialisées

Model Context Protocol MCP

→ Connexion à d'autres services (bases de données, fichiers, services en ligne...)

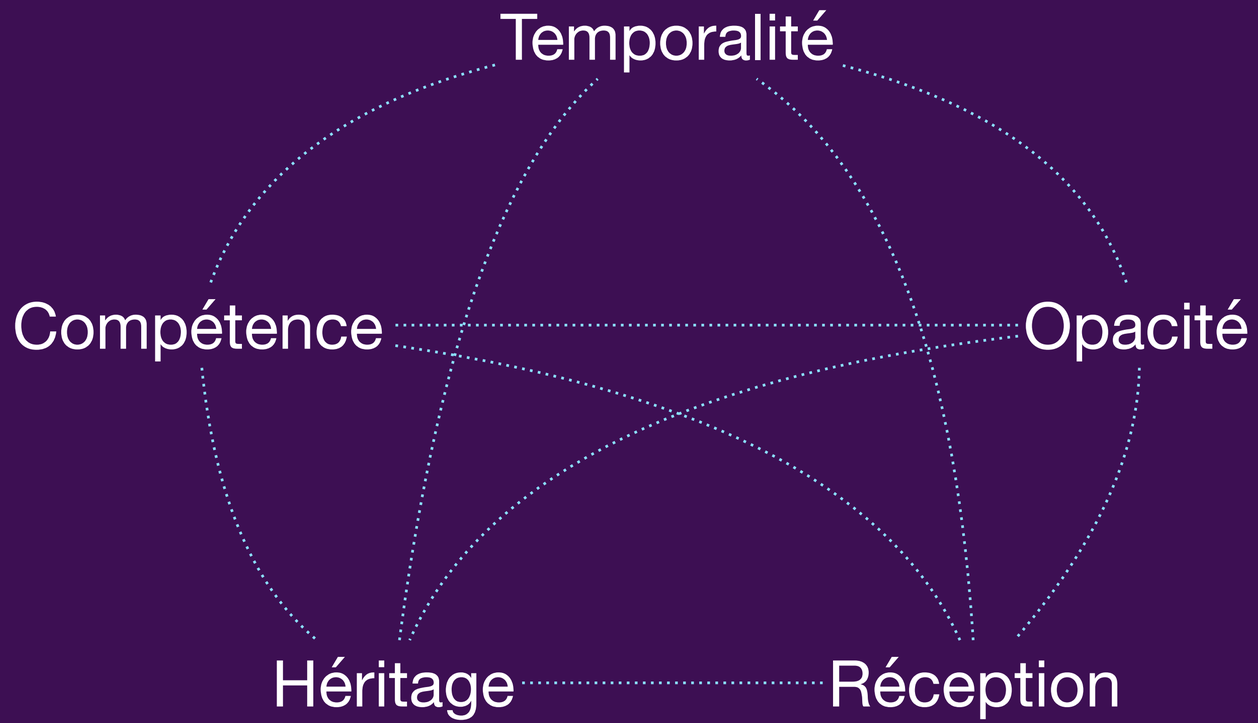
Harness

→ Environnement d'orchestration du LLM avec de la mémoire, des outils,...

2. Etudier l'IA

... depuis les sciences sociales

Des précautions et des difficultés multiples



d'après la thèse de Virginia Haussauer : *Histoire et conditions de la critique du numérique (en cours de finalisation)*

Des analogies avec Internet

Longue histoire

Émergence soudaine (Web vs ChatGPT)

Sous-estimé dans un premier temps

Sentiment d'accélération

Jugements moraux

Attentes démesurées

Bulle spéculative

Acteurs dominants

Gouvernance problématique

Des oppositions fondamentales

Intelligence

Efficience

Matérialité

Risques et opportunités

➜ **Les politiques de l'IA dépendent de ces oppositions**

3. Et si ça marche…

Politiques de l’émergence et de l’accélération

A. Une bulle ?

Une fuite en avant…

Une croissance inédite, une pénurie à court terme



SpaceX xAI - Southaven - Mississipi - 9 septembre 2026 - ScientificAmerican

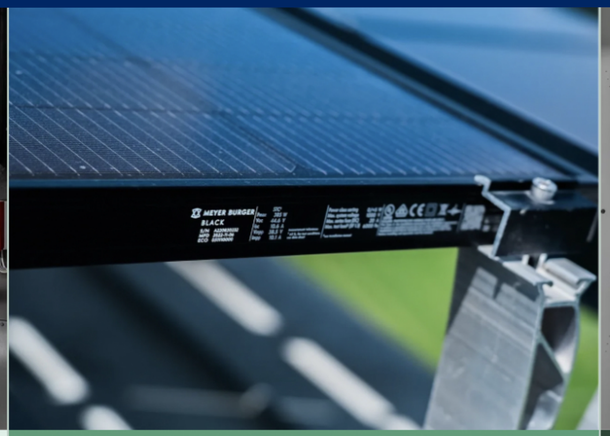
Une croissance inédite, une pénurie à court terme



Production de chaleur

Depuis 2025, l'électricité consommée par nos nouveaux data centers n'est pas perdue : elle est toute l'année 100% revalorisée en chaleur pour chauffer des foyers.

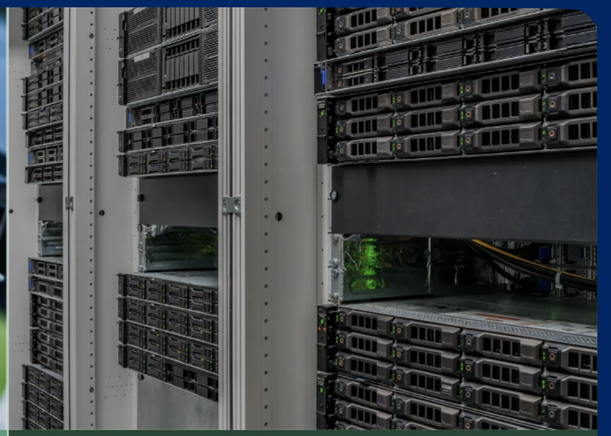
Notre système innovant →



Production d'électricité solaire

Infomaniak développe sa propre production solaire avec des panneaux fabriqués en Europe pour alimenter ses infrastructures et renforcer la part des énergies renouvelables face à un mix encore largement dépendant des énergies fossiles.

Nos sites de production →



À la pointe de l'efficacité énergétique

Depuis 2013, nos data centers sont refroidis avec de l'air extérieur filtré, sans production de froid artificiel dédiée aux serveurs. Avec un PUE moyen inférieur à 1.1, ils figurent parmi les plus performants au monde en efficacité énergétique.

Notre infrastructure →

Unil.

infomaniak The Ethical Cloud

Infomaniak - « Nous polluons, mais nous agissons » - Consulté le 9 septembre 2026

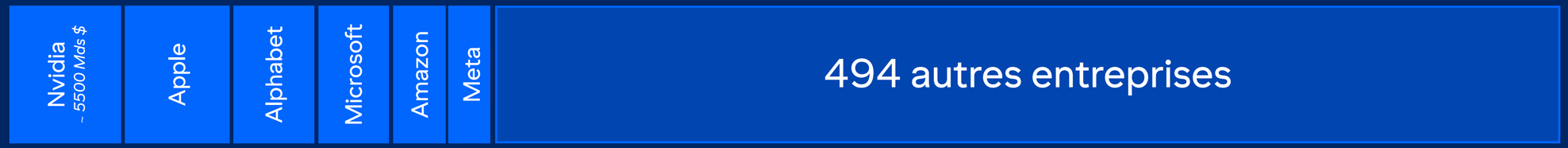
Une croissance inédite, une pénurie à court terme

Une bulle comparable à celle d'Internet en 2000 ?

1000 Md$ d'investissements prévus pour 2027

Une fuite en avant des capitalisations boursières

6 entreprises
~ ⅓ du S&P 500



... sans OpenAI, Anthropic et SpaceX (Grok)

Unil.

Cette bulle peut éclater. Mais l'IA y survivra...

À quoi ressemblera le monde en 2037 ?

Un élève qui entre en 1P cette rentrée sortira de l'école obligatoire en 2037… et certains en 2045.

Unil.

Ne pariez pas sur l'échec de l'IA

Se rassurer avec des mains à six doigts ?





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

B. Déplacer la critique ?

Une IA qui marche ne pose pas moins de problèmes…

L'asymétrie de la critique

Les critiques de l'IA s'appliquent très bien à nous-même

- Confabulations
- Connaissances limitées
- Biais d'apprentissage
- Discrimination
- Destruction de l'environnement
- Fiabilité
- Imitation
- Interprétabilité

Boris Beaude - 15 septembre 2026

Des perroquets très intelligents ?

Des perroquets qui traduisent...

- Complexité
- Prédiction
- Expérience
- Alignement
- Émergence
- Accélération

Boris Beaude - 15 septembre 2026 CC BY NC SA

Les paradoxes de la critique

- Des perroquets qui ne font que répéter…
  … mais qui hallucinent !
- Des IA qui ne fonctionnent pas…
  … mais qui vont nous remplacer !
- Des IA qui discriminent…
  … mais qui ne reconnaissent pas les personnes racisées !
- Les IA sont nécrosées (récurrence et corrélation)…
  … mais nous sommes dépossédés de nous-mêmes !
- Les IA ne manipulent que des chiffres, pas des mots…
  … mais notre cerveau ne manipule pas des mots !

- Se concentrer sur les *échecs* nous protège à *court terme*
- Se concentrer sur les *capacités* nous protège à *moyen terme*
- Se concentrer sur les *potentialités* nous protège à *long terme*



Instagram - theaifield - 22 novembre 2025 - « A night out with the boys »

Ambivalence de la critique

Être discriminé

par erreur

ou

intentionnellement ?

➜ **Il est temps d'avoir une politique de l'IA**

Réviser nos jugements ?

Des critiques résistent à l'évolution de l'IA

Ce qui tient encore :

Le processus reste opaque et instable.

Les limites de l'IA ne sont pas prévisibles.

Son impact environnemental croît rapidement.

Le travail humain reste dissimulé.

Il faut maintenir une lecture critique des textes et des images.

Le rythme imposé est sans rapport avec celui des institutions.

La concentration du pouvoir et du capital est inédite

Situer les limites, percevoir les dynamiques, se concentrer sur les problèmes de fond...

Boris Beaude - 15 septembre 2026

C. Délégation ?

Garder la main sur ce que l'on cède

La délégation ultime ?



La révolution industrielle

Délégation de la force mécanique

Charlie Chaplin • Les Temps modernes • 1936



La transformation actuelle

Délégation des fonctions cognitives

Lee Jin-Man / AP • Lee Sedol contre AlphaGo de DeepMind • 12 mars 2016

Boris Beaude - 15 septembre 2026 CC BY NC SA

La délégation ultime ?

Cinq qualités qu’aucun humain ne peut réunir

Compétence : avoir des compétences élevées dans de très nombreux domaines.

Individualisation : s’adapter à chaque personne selon la situation et le niveau.

Automatisation : réaliser des tâches de plus en plus complexes sans intervention.

Vitesse : être relativement rapide et disponible à toute heure.

Coût : pour un coût relativement faible à la tâche (traduire, répondre à une question, coder…).

Se poser les bonnes questions sera de plus en plus décisif.

Boris Beaude - 15 septembre 2026

Deux problèmes symétriques

Les vertus de l'expérience

La performance peut réduire la vigilance

C’est vraisemblable, donc cela doit être vrai.

La vigilance peut réduire la performance

Une critique sans pratique manque de substance.

Mais avec vigilance, l’expérience accroit la compétence.

L’IA tend à pouvoir faire tout…
ce qu’un humain peut faire avec un ordinateur.

Boris Beaude - 15 septembre 2026

La délégation mécanique qui vient ?



Joanna Maciejewska
@AuthorJMac

You know what the biggest problem with pushing all-things-AI is? Wrong direction.
I want AI to do my laundry and dishes so that I can do art and writing, not for AI to do my art and writing so that I can do my laundry and dishes.

Traduire le post

12:50 PM · 29 mars 2024 · **3,3 M** vues

Boris Beaude - 15 septembre 2026

C. Éducation ?

Une domaine particulièrement boulversé

Le problème n'est pas que

… l’IA fasse les devoirs des élèves.

Le problème est que

… l’IA soit capable de faire les devoirs des élèves.

Boris Beaude - 15 septembre 2026

De la transmission à l'accompagnement

Les humains n'ont plus le monopole du savoir (faire)

Le statut fondé sur la détention du savoir devient plus fragile

Médecins, consultants, juristes, informaticiens… et bien entendu enseignants.

Unil.

L'articulation entre la transmission et l'accompagnement devient centrale.

Le *deskilling* n'est pas une fatalité !

Un cercle vertueux ?

L'apprentissage demeure nécessaire pour être compétent,
et la compétence demeure nécessaire pour apprendre.

Avec des bases et un cadre

Des conditions d'apprentissage inédites.

Sans bases et sans cadre

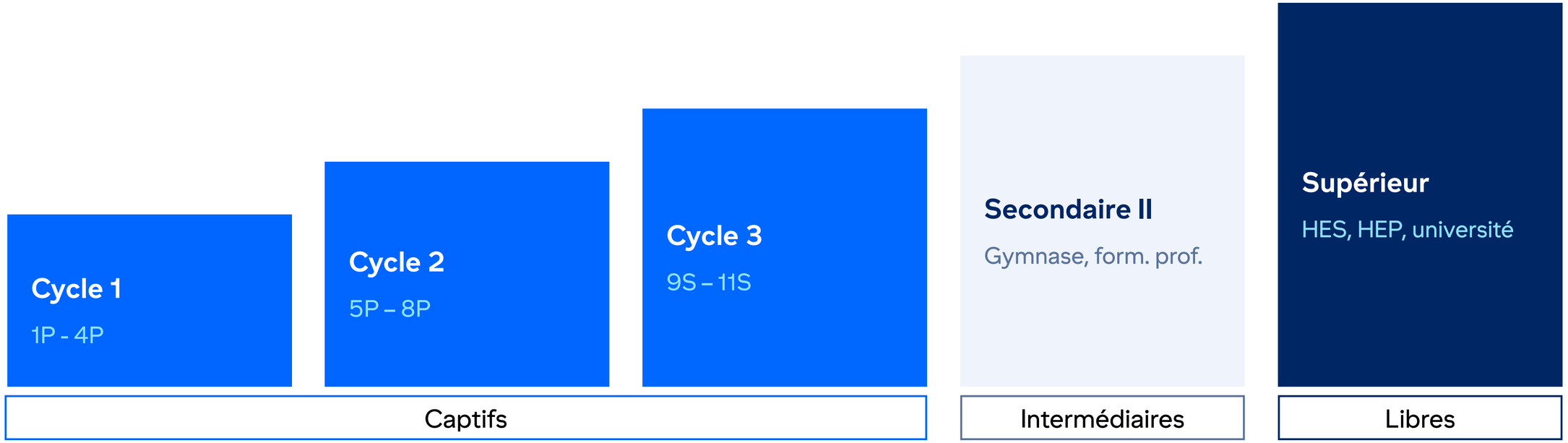
Une illusion de compréhension et un déclassement.

Principale source de l'amplification des inégalités.

Boris Beaude - 15 septembre 2026 CC BY NC SA

Il n'y a pas UN problème de l'IA à l'école

Il y en a autant que de missions, de niveaux et de disciplines.



Unil.

Les défis existent dès les premières années, mais l'impact croît avec l'autonomie.

Pourquoi utiliser l'IA ?

Pourquoi éduquer les nouvelles générations ?

L'IA n'est pas une fin en soi, et l'école non plus.

Ne défendons ni l'IA, ni l'école : accompagnons les étudiant·e·s.

Boris Beaude - 15 septembre 2026

Trois défis majeurs

1. Se concentrer sur les missions

À quoi formons-nous si nous ignorons ce que la machine fera à notre place ?
Se recentrer sur ce à quoi nous tenons !

2. Repenser l'évaluation

Quelles compétences évaluer lorsque la délégation est si puissante ?
Penser l'évaluation comme un pilier de l'apprentissage !

3. Garder la main

Comment rester responsable de ce que nous ne produisons plus ?
Trouver l'équilibre entre la déontologie et l'utilitarisme !

Boris Beaude - 15 septembre 2026 CC BY NC SA

D. Souveraineté ?

L’IA n’est pas une exception

Acteurs mondiaux et souveraineté

Gravure

ASML

ZEISS

KLA

APPLIED MATERIALS

Lam RESEARCH

TEL TOKYO ELECTRON

Lasertec

Rapidus

SMIC

Processeurs

cerebras

intel.

AMD

HUAWEI

Mémoires

SK hynix

SAMSUNG

Processeurs + modèles

Google

amazon

NVIDIA

Microsoft

xAI

Modèles

ANTHROP\C

Meta

OpenAI

Modèles chinois

ZHIPU · AI

Moonshot AI

Alibaba Cloud

deepseek

ByteDance

MINIMAX

Acteurs mondiaux et souveraineté



E. Un problème politique ?

Organiser les possibles

Politiques des IA & Pharmakon

- Epuisement des sources et des ressources
- Fragilisation de la sphère publique
- Fragilisation des modalités d'apprentissage (Deskiling)
- Captation privée de la recherche et de l'innovation
- Fragilisation et précarité de l'environnement de travail
- Concentration du pouvoir
- Automatisation des discriminations
- Concentration des gains de productivité
- Délégation cognitive incontrôlée
- Fragilisation de la sécurité informatique
- Renforcement des capacités militaires et sécuritaires

- Transition écologique accélérée et biens communs
- Amélioration de la veille et du fact-checking
- Individualisation de l'apprentissage (Upskilling)
- Accélération de la découverte scientifique
- Redistribution du temps de travail
- Renforcement du contrôle démocratique
- Identification et correction des discriminations
- Redistribution des gains de productivité
- Réflexivité et explicitation augmentées
- Renforcement de la sécurité informatique
- Construction d'une souveraineté numérique et défensive

→ **Sortir les lumières de l'obscurité**

Cinq axes de tension

01

Nécessité - Délibération

Qui détient le rythme du déploiement ?

L'inévitable

Le rythme est dicté par la course technologique, le marché, la géopolitique. La politique suit.

L'instituable

Tout déploiement est un choix collectif, discutable, suspendable, réorientable.

02

Extraction - Contribution

Que devons-nous à ce dont nous vivons ?

Prélèvement sans dette

L'IA prélève sur les communs informationnels, l'attention, les écosystèmes, sans avoir à les reconstituer.

Obligation contributive

Tout système qui prélève sur un commun lui doit réparation : presse, savoirs, environnement, recherche, formation.

03

Dépossession - Libération

Que faisons-nous du travail humain ?

Le travail capté

Les gains de productivité remontent vers le capital. L'IA intensifie, surveille et déqualifie le travail restant. Le travailleur perd la maîtrise de son activité.

Le travail libéré

Les gains financent la réduction du temps de travail et la délégation des tâches pénibles ou répétitives. L'IA rend au travail vivant ce qui ne peut être délégué : le jugement, le sens, la relation.

04

Substitution - Capacitation

Que reste-t-il de notre puissance d'agir ?

Le sujet déchargé

L'IA décide, crée, juge à notre place. La dépendance s'installe, l'atrophie s'étend.

Le sujet outillé

L'IA augmente la capacité à comprendre, argumenter, délibérer, contester.

05

Indistinction - Véridiction

Reste-t-il une base commune pour s'accorder sur ce qui est ?

L'équivalence généralisée

Le synthétique sature le champ informationnel. Les signaux du vrai s'effondrent.

L'institution du vrai

L'IA outille les institutions de véridiction : presse, science, justice, archives.

Ce que l'IA fera de nous

dépend directement de

ce que nous ferons d'elle.

Unil.

Pr. Boris Beaude · Université de Lausanne

Ce que l'IA fera de nous

dépend directement

de ce que nous ferons d'elle.

Merci pour votre attention

Unil.

Pr. Boris Beaude · Université de Lausanne

Sources des images générées par IA

A. : Boris Beaude, via Gemini Nano Banana 2
Prompt : « Image of a goldfish inside a "Cola-Coca" bottle on a wooden table », 24 février 2026

B. : Boris Beaude, via Midjourney v6.1
Prompt : « Image of a goldfish inside a "Cola-Coca" bottle on a wooden table », 4 décembre 2024

C. : Pranav Dixit, Hands via Stable Diffusion, BuzzFeed, 31 janvier 2023

D. : Boris Beaude, via Midjourney
Prompt : « Vibrant , Beautiful hand on a wooden table , Cinematic photography , --aspect 1:1 --v 5.2 », octobre 2023

E. : Boris Beaude, via GPT 5.4 Image 2, 9 août 2026
Prompt : « Photographie en gros plan : au premier plan, deux personnes se serrent fermement la main ; à l'arrière-plan, une troisième personne fait un signe de la main paume ouverte. Sur une table en bois devant elles, une quatrième main compte sur ses doigts en montrant exactement trois doigts levés (index, majeur, annulaire), pouce et auriculaire repliés contre la paume. L'une des deux personnes qui se serrent la main tient également un verre de vin transparent par le pied avec sa main libre, les doigts naturellement enroulés autour. Lumière naturelle de fenêtre, faible profondeur de champ, photoréaliste, objectif 50mm ».

F. : Boris Beaude, via Gemini 3.6 Flash, 10 août 2026
Prompt avec la photo supra : « Peux-tu animer cette photo, 5 secondes, pour la rendre plus vivante ? »

The Dark Enlightenment - NRx


Nick Land


Curtis Yarvin


J. D. Vance

Turing Award


Yann Le Cun


Geoffrey Hinton


Yoshua Bengio

Safe Superintelligence


Ilya Sutskever

Palantir


Peter Thiel


Alex Karp

Attention Is All You Need


Ashish Vaswani

OpenAI


Greg Brockman


Sam Altman

ANTHROP\C


Dario Amodei


Daniela Amodei

Meta


Mark Zuckerberg

Google DeepMind


Demis Hassabis

THINKING MACHINES


Mira Murati


Chris Olah

xAI


Elon Musk

NVIDIA


Jensen Huang


Arthur Mensch