



Intelligences artificielles *en sociétés*

1. De quoi l'IA est-elle le nom ?



De quoi l'IA est-elle le nom ?

HPC cluster, abstract, in an original artistic style, inspired by art deco, with a yellow-colored background

Midjourney - Style personnel - 7 novembre 2025

De quoi l'IA est-elle le nom ?



De quoi l'IA est-elle le nom ?



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

De quoi l'IA est-elle le nom ?

Préalable

Nous n'avons pas attendu l'IA pour ...

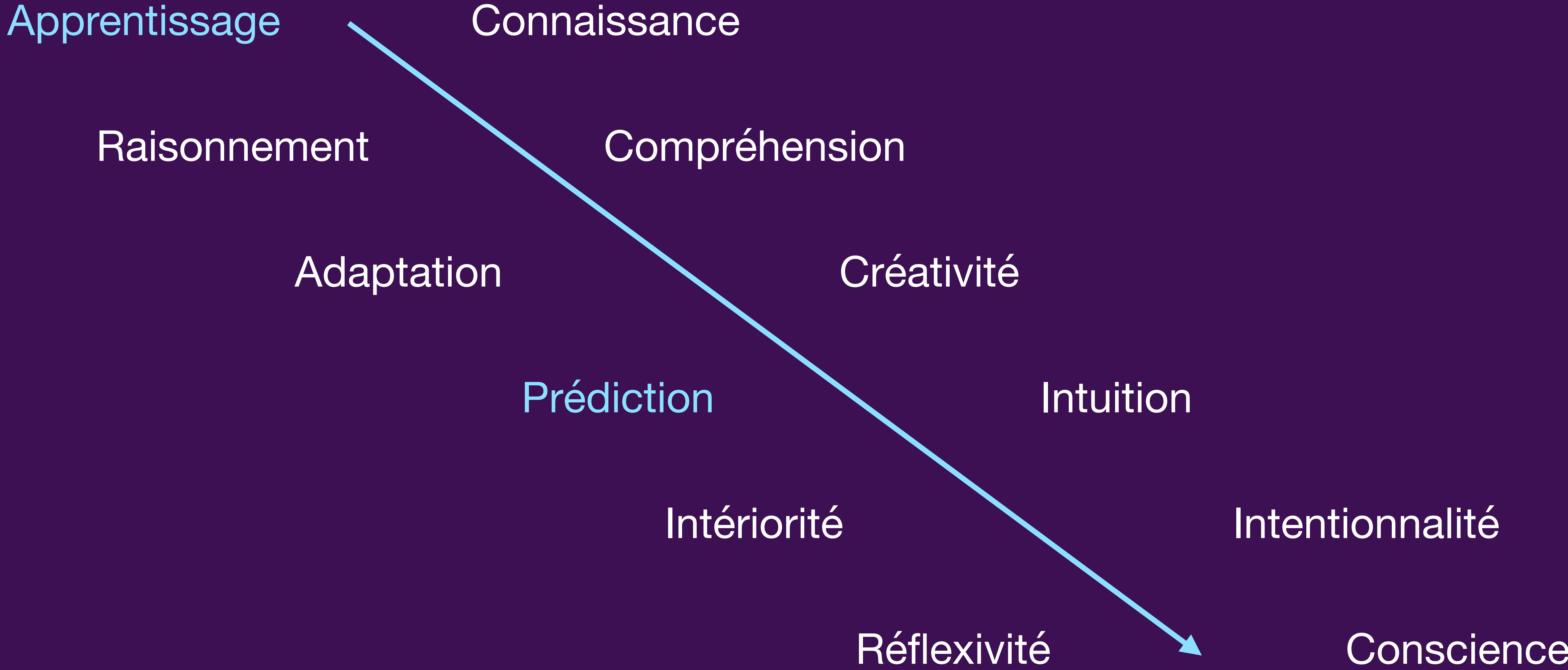
Nous entretenons massivement
Déshumaniser et exploiter nos semblables
Détruire notre environnement
Manipuler les imaginaires

...

Intelligences

Un concept fuyant aux acceptions multiples

Définitions multiples



L'AI effect



Reconnaissance de caractères (Lecun - Tri postal - 1989)

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)

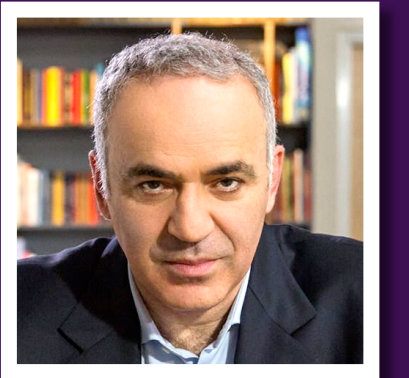
Art visuel ? (Dall-E - 2021, Midjourney - 2022)

Conduite automobile ? (Waymo - 2022)

Dialogue généraliste (ChatGPT - 2022)

Désambiguïsation lexicale (GPT 3.5 - 2022)

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



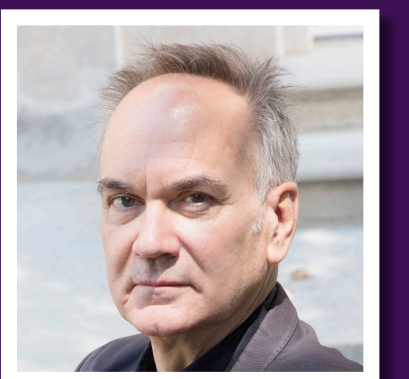
[Garry Kasparov](#)



[Ken Jennings](#)

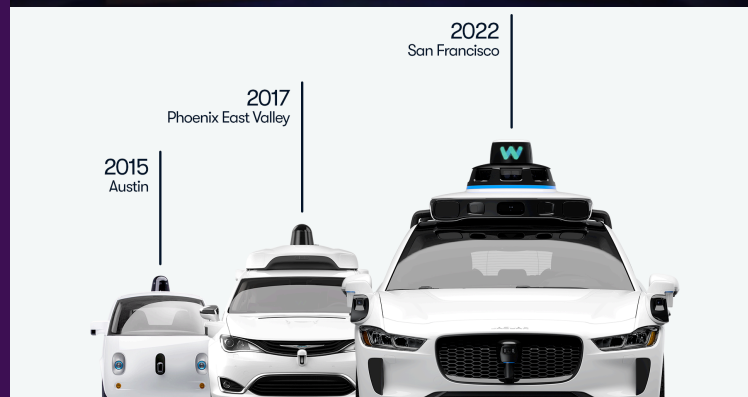


[Lee Sedol](#)



[Hervé Le Tellier](#)

L'AI effect



1990

Reconnaissance de caractères (Lecun - Tri postal - 1989)

Échecs (IBM Deep Blue - Kasparov - 1997)

2000

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

Go (DeepMind AlphaGo - Lee Sedol — 2016)

Traduction ? (DeepL - 2017)

Compréhension d'un texte ? (Alibaba - 2018)

2020

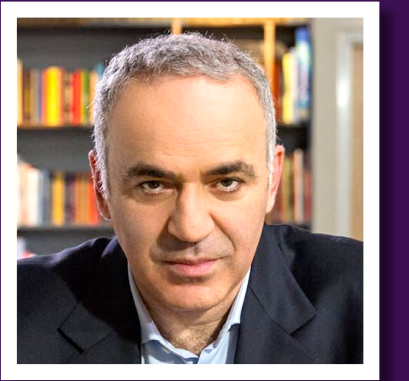
Art visuel ? (Dall-E - 2021, Midjourney - 2022)

Conduite automobile ? (Waymo - 2022)

Dialogue généraliste (ChatGPT - 2022)

Désambiguïsation lexicale (GPT 3.5 - 2022)

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



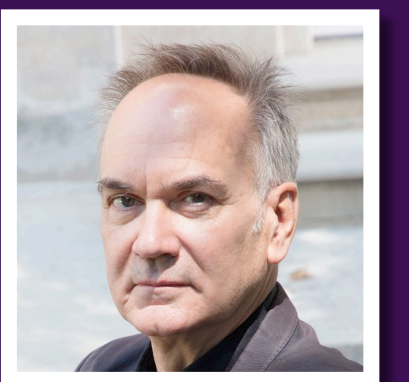
Garry Kasparov



Ken Jennings



Lee Sedol



Hervé Le Tellier

Select AI Index technical performance benchmarks vs. human performance

Source: AI Index, 2025 | Chart: 2025 AI Index report

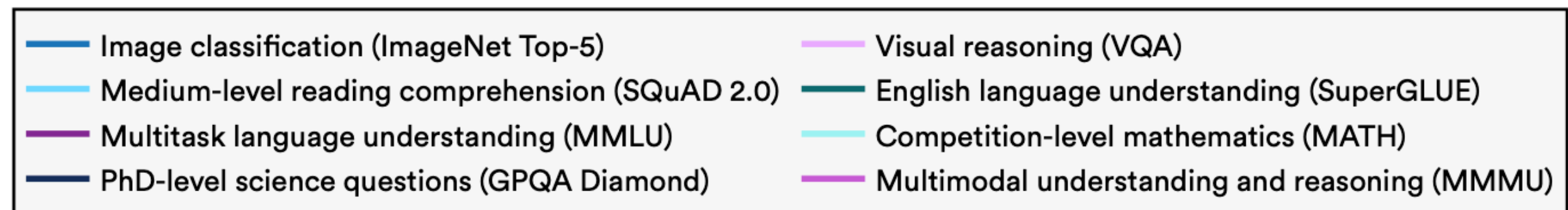
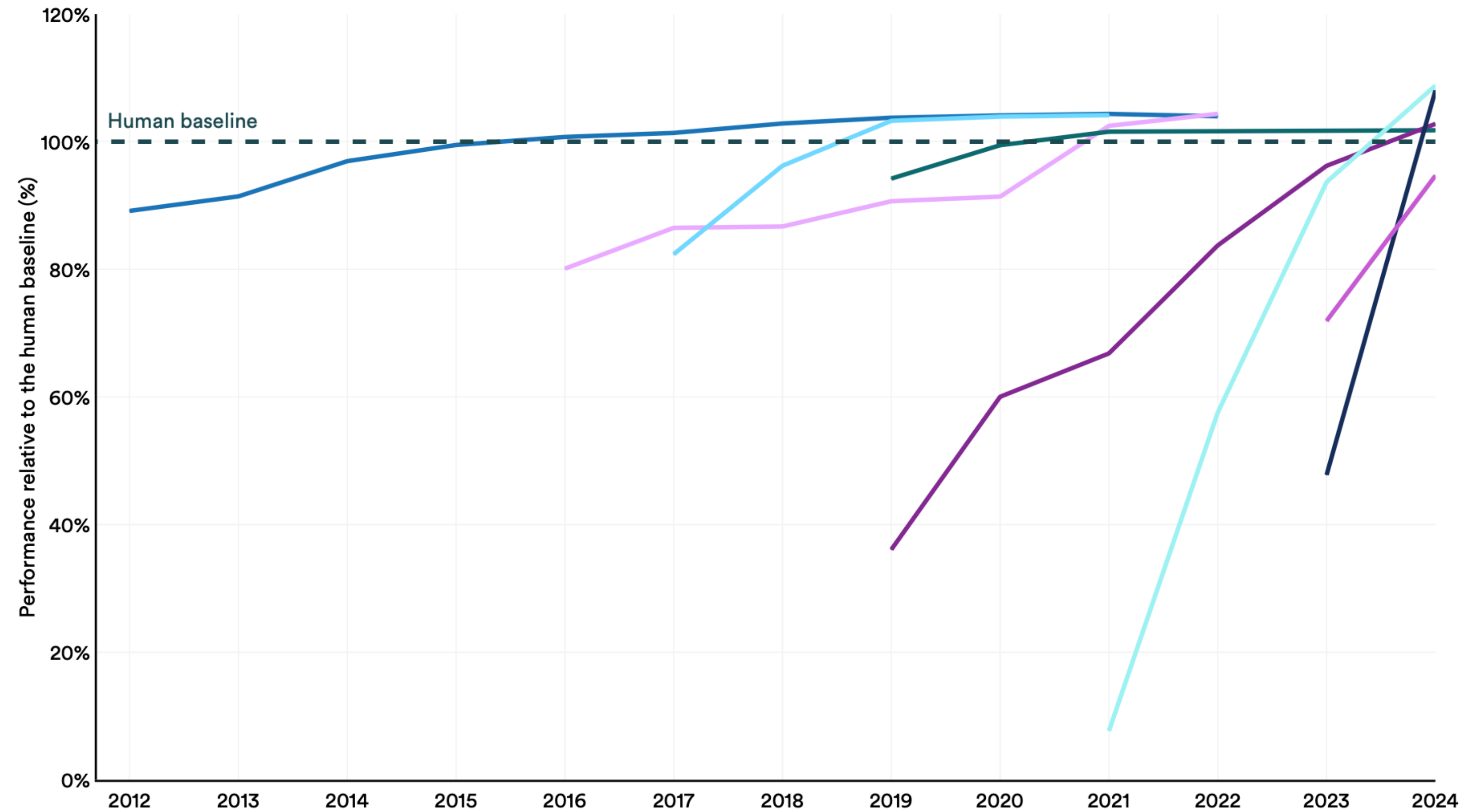


Figure 2.1.33²

Epoch Capabilities Index (ECI)

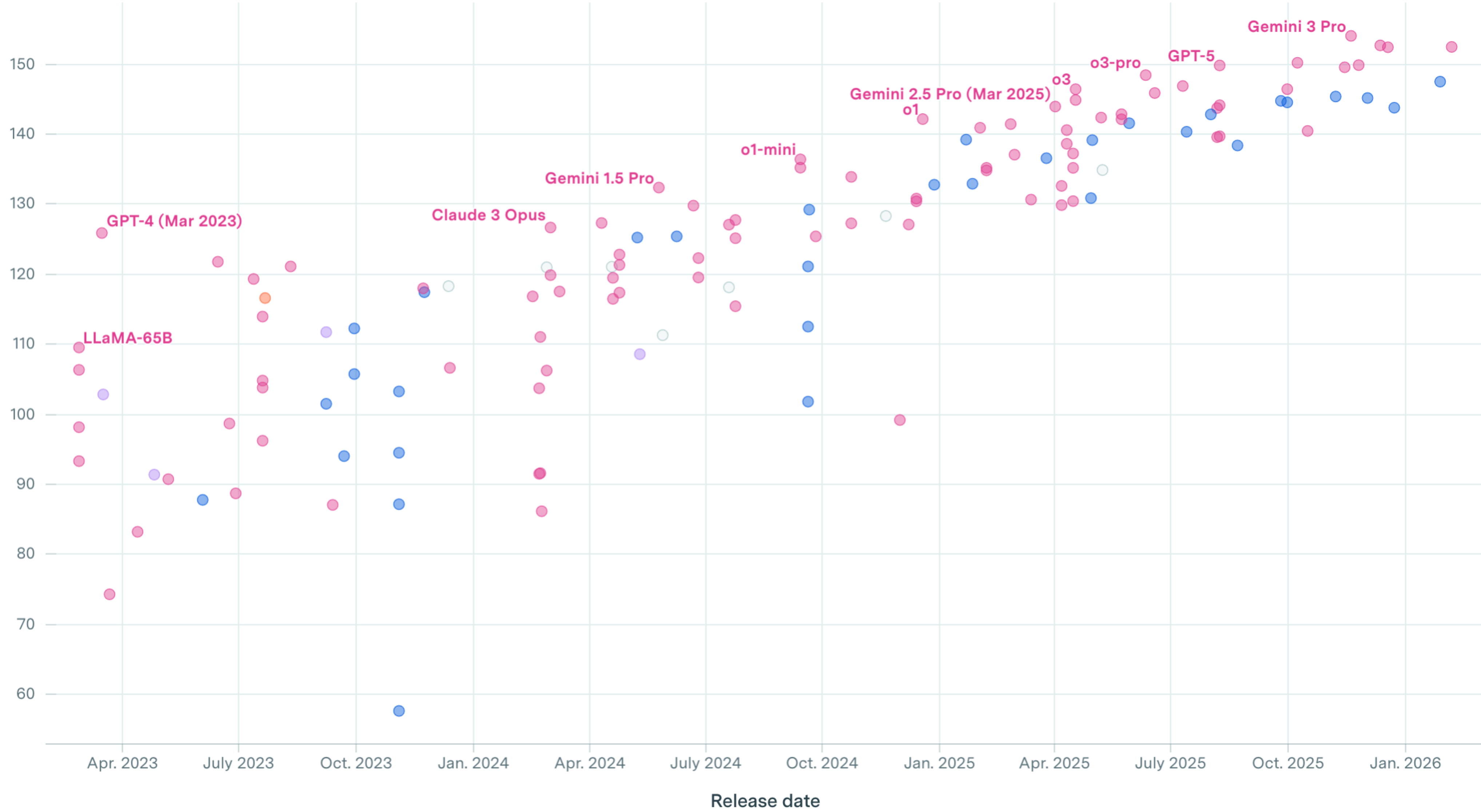
Graph Settings

Score

134 Results

Country

- United States of America
- China
- EU
- United Arab Emirates
- United Kingdom



Epoch Capabilities Index (ECI)

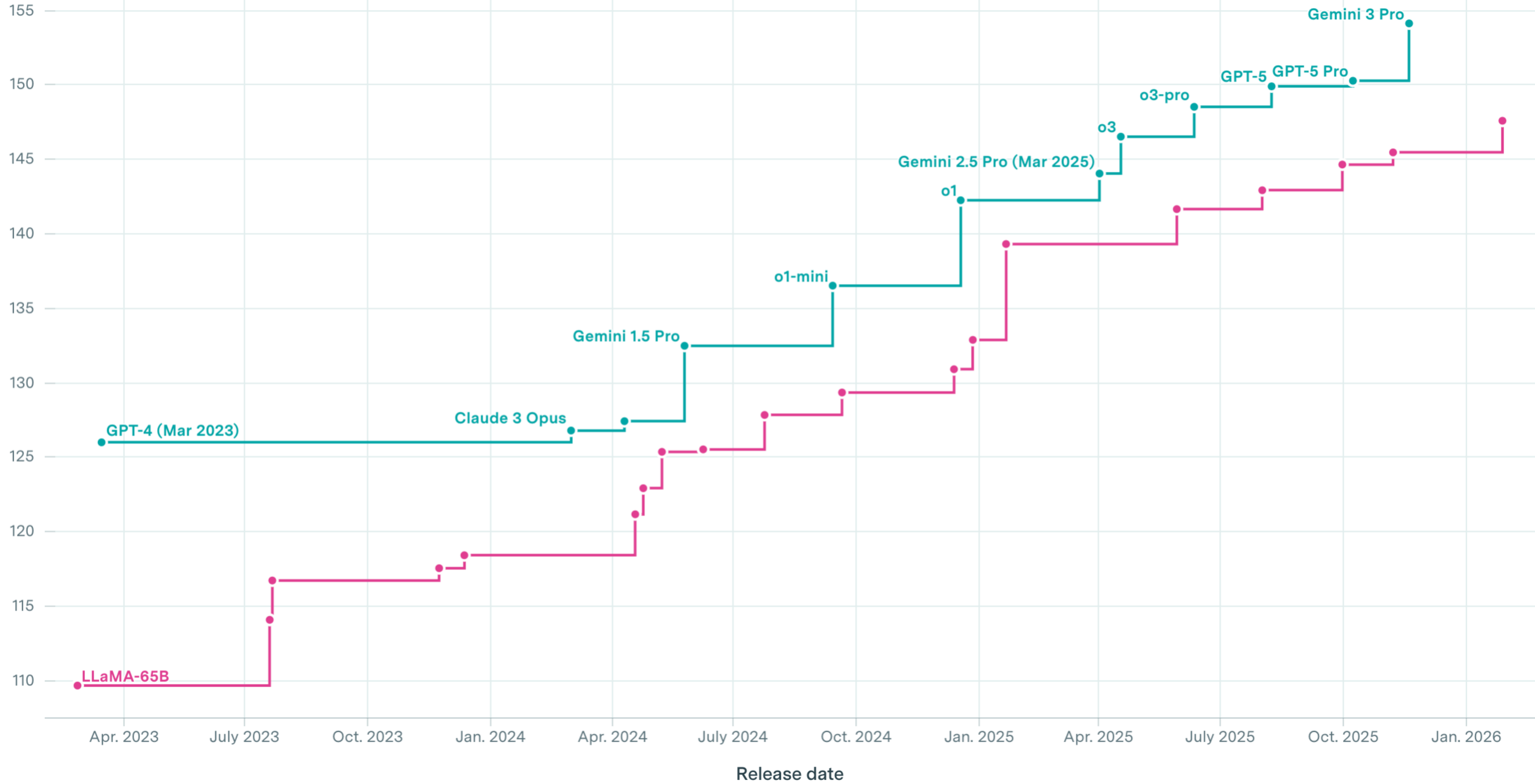
Graph Settings

Score

31 Results

Accessibility

- Open weights
- Closed weights



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

System ▼

Presets ▼

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

Effacer Envoyer

OpenAI: GPT-4.1 Nano ✕

Options ▼

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

ASSISTANT
La phrase corrigée est : **Les enfants sont partis jouer dans la cour.**
1.5s 40 tokens <\$0.0001

Continuer la conversation... ↑

Anthropic: Claude 3 Haiku ✕

Options ▼

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

ASSISTANT
Voici la phrase corrigée :
Les enfants sont partis jouer dans la cour.
Explications des corrections :
• "Enfen" -> "enfants" (orthographe correcte)
• "son parti" -> "sont partis" (accord avec le sujet "les enfants")
• "joué" -> "jouer" (infinitif du verbe)
• "den" -> "dans" (préposition correcte)
• "court" -> "cour" (orthographe correcte)
1.7s 150 tokens \$0.0002

Continuer la conversation... ↑

Google: Gemini 2.5 Flash Lite ✕

Options ▼

ASSISTANT
Raisonnement ▼
Voici la version corrigée :
Les enfants sont partis jouer dans la cour.
Explication des corrections :
• **enfen** -> **enfants** (faute d'orthographe)
• **son** -> **sont** (verbe être à la 3ème personne du pluriel, car le sujet "les enfants" est pluriel)
• **parti** -> **partis** (accord du participe passé "parti" avec le sujet masculin pluriel "les enfants")
• **joué** -> **jouer** (après un verbe de mouvement comme "partir", on utilise l'infinitif pour exprimer le but : ils sont partis *pour* jouer)
• **den** -> **dans** (préposition correcte)
• **la court** -> **la cour** (le nom "cour" désignant une aire de jeux, un espace extérieur dans une habitation ou une école, est féminin. "Court" est masculin s'il s'agit par exemple d'un court de tennis).
8.9s 2050 tokens \$0.0008

Continuer la conversation... ↑

Mistral: Ministral 3 3B 2512 ✕

Options ▼

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

ASSISTANT
Voici la phrase corrigée :
"Les enfants sont partis jouer dans la cour."
Explications des corrections :
• **"Enfen"** → **"Enfants"** (accord du pluriel)
• **"son"** → **"partis"** (verbe correct, accord avec "enfants")
• **"joué den"** → **"jouer dans la cour"** (conjugaison correcte et orthographe)
713ms 107 tokens <\$0.0001

Continuer la conversation... ↑

Les enfen son parti joué den la court

Langue

- Erreurs (1)
 - enfen → enfin
- Ruptures (1)
 - Les enfen son
- Alertes (1)
 - den → *Anglicisme*

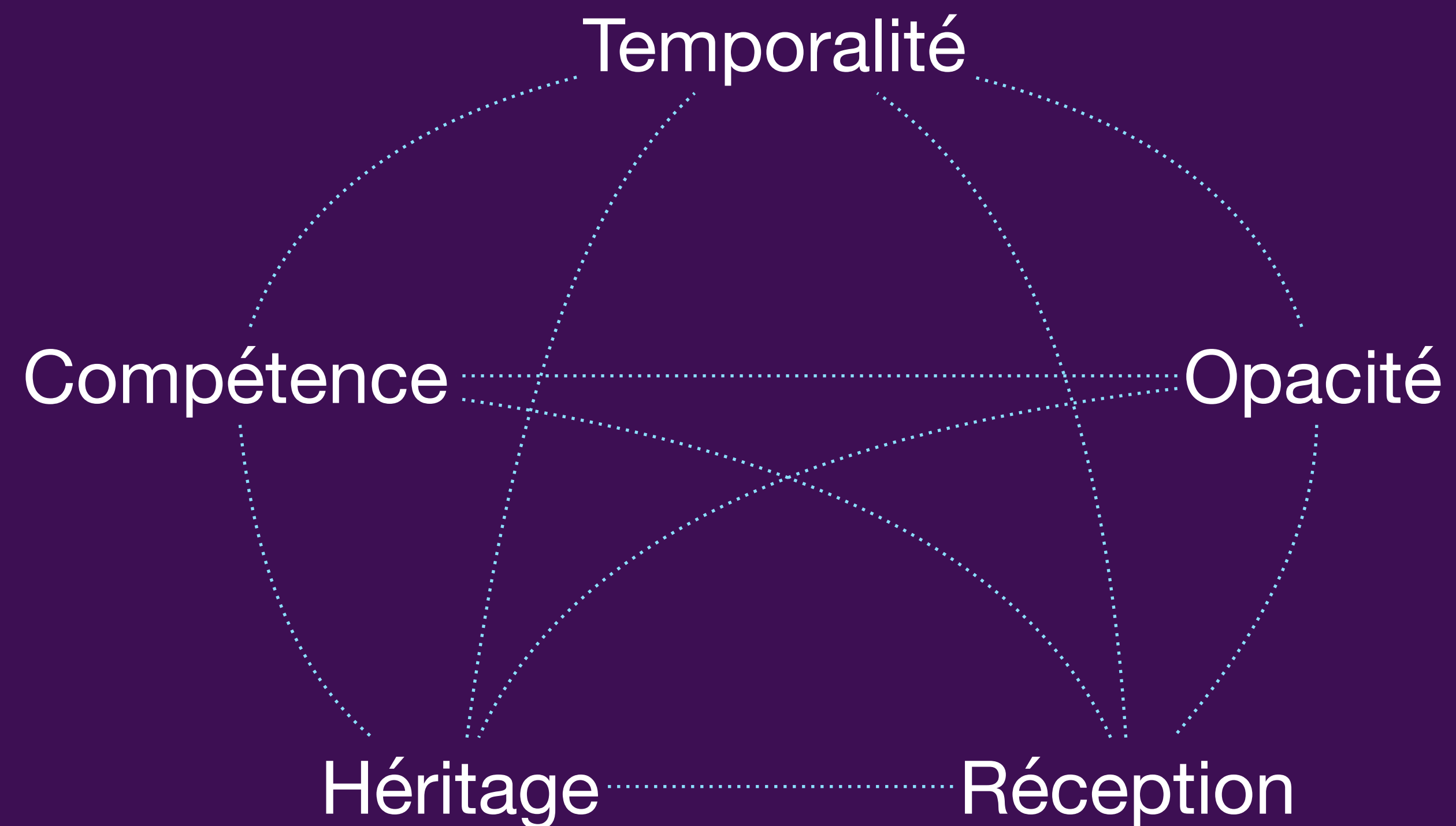
Correcteur

Dictionnaires

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 conditions particulières

Abondance des données
Capacité des infrastructures
Innovations algorithmiques

→ **Accélération des innovations**

L'importance du pluralisme

Données

Modèles

Acteurs

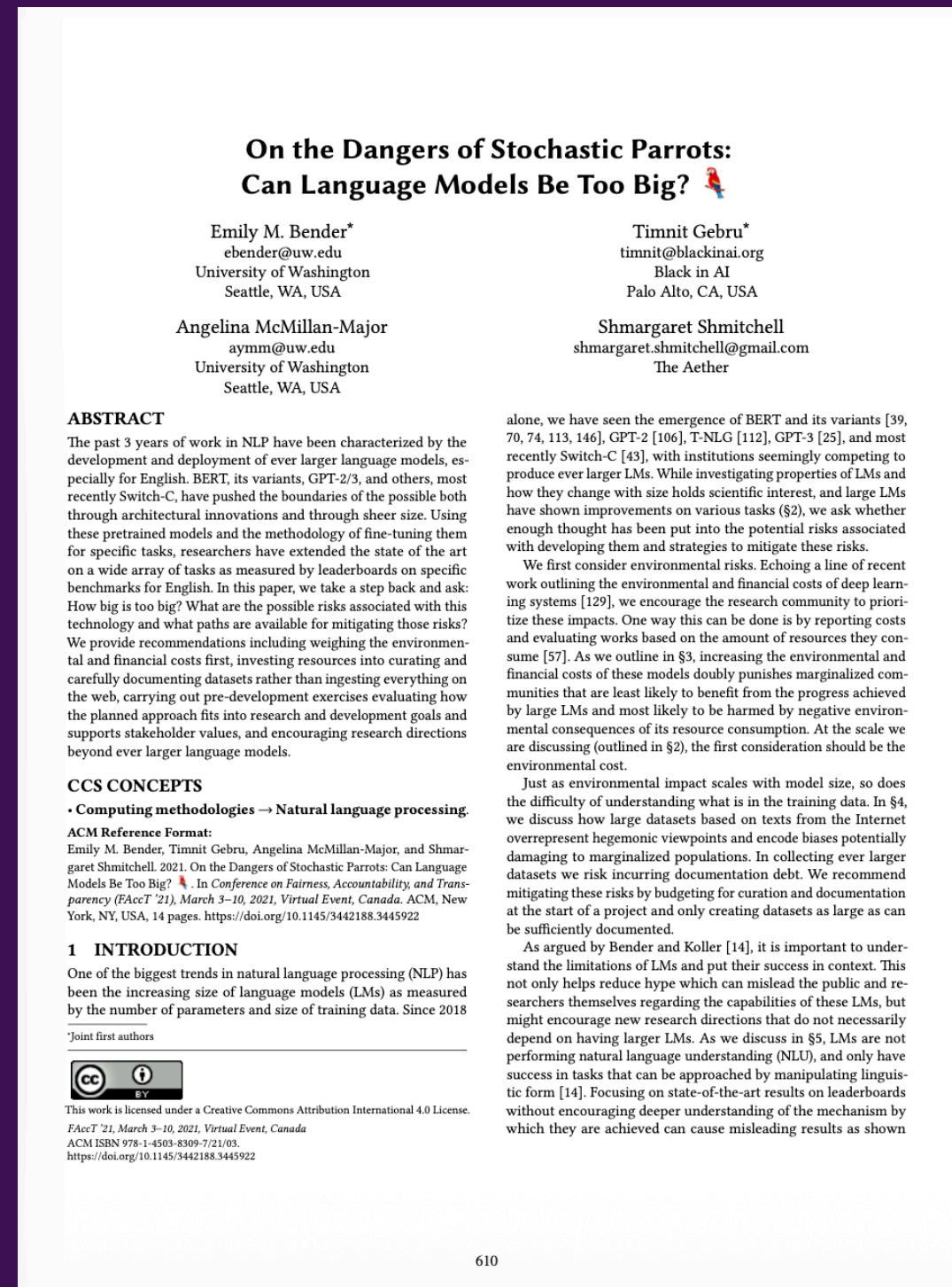
Usages

Imaginaires

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

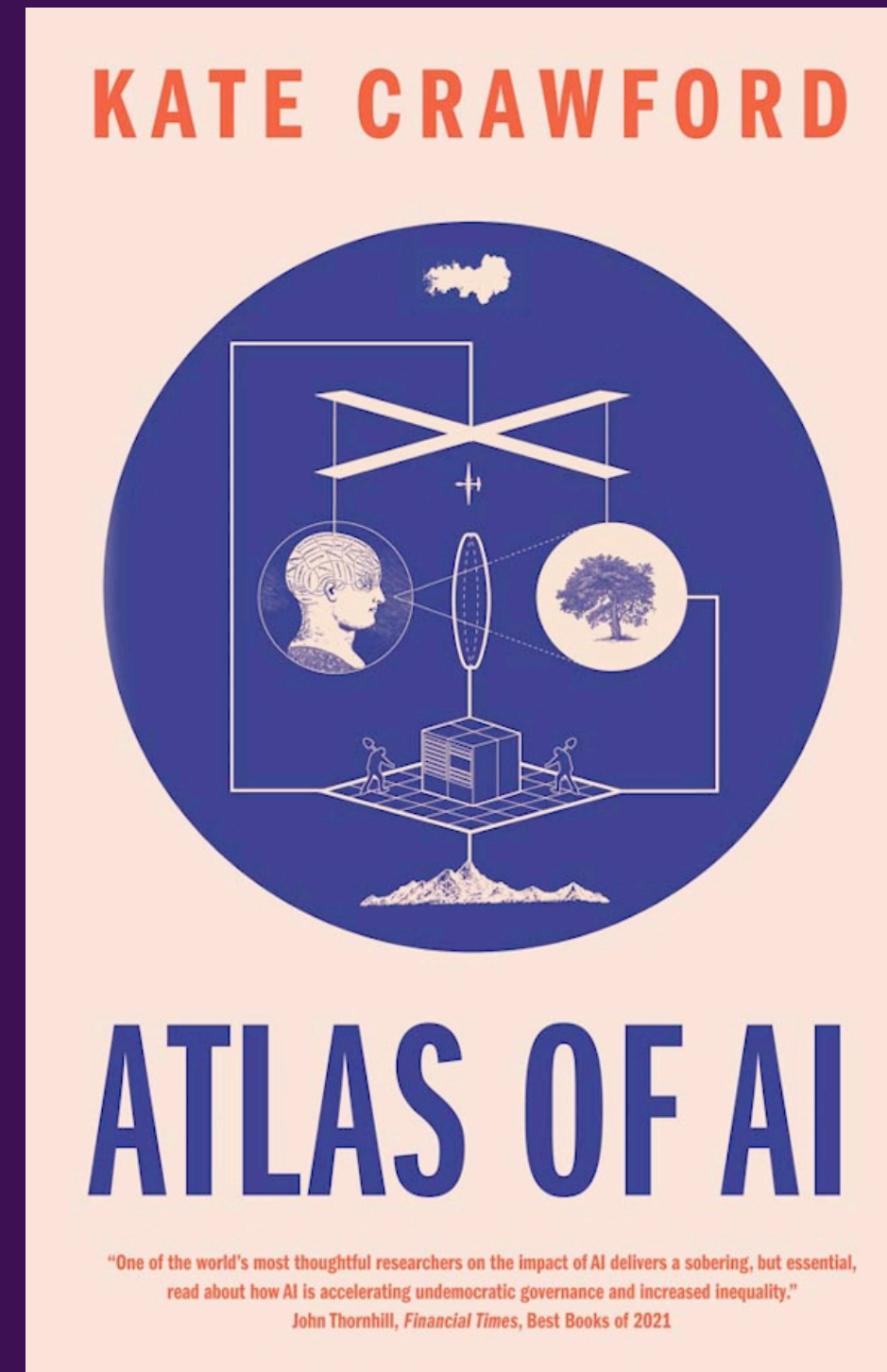
Des oppositions fondamentales

2021



Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21). Association for Computing Machinery, New York, NY, USA, 610–623. <https://doi.org/10.1145/3442188.3445922>

2021



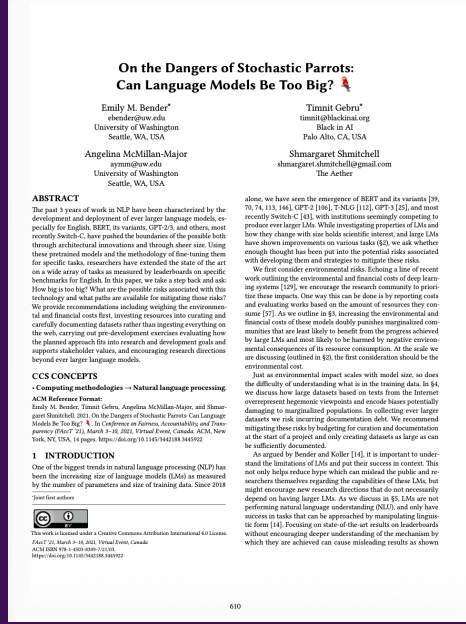
Kate Crawford, *Atlas of AI*, Yale University Press, 2021

2025

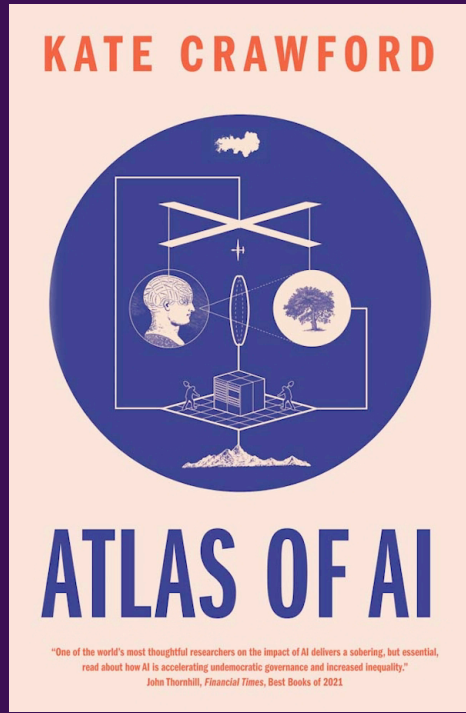


Emily M. Bender, Alex Hanna, *The AI Con: How to Fight Big Tech's Hype and Create the Future We Want*, HarperCollins Publishers, 2025

Des oppositions fondamentales



- Le sens est inaccessible aux LLM
- Les LLM ne font *que* prédire le mot suivant
- Les hallucinations sont structurelles, sans accès au sens
- Les langues et cultures moins représentées sont mal traitées
- Les données utilisées sont biaisées et discriminantes
- Les grands modèles augmentent l'illusion de performance
- Risque élevé d'anthropomorphisation et d'interprétation génèreuse
- L'impact environnemental est très important
- Atlas of AI : « L'IA n'est ni intelligente, ni artificielle »
- The AI Con : « hype » au service des levées de capital et de l'extractivisme



Des oppositions fondamentales

Intelligence

Efficiency

Matérialité

Risques et opportunités

→ Les politiques de l'IA dépendent de ces oppositions

Hypothèses principales

L'évolution de l'IA depuis 5 ans est difficile à concevoir
La notion d'émergence mérite d'être considérée

→ **Se concentrer sur les innovations récentes**

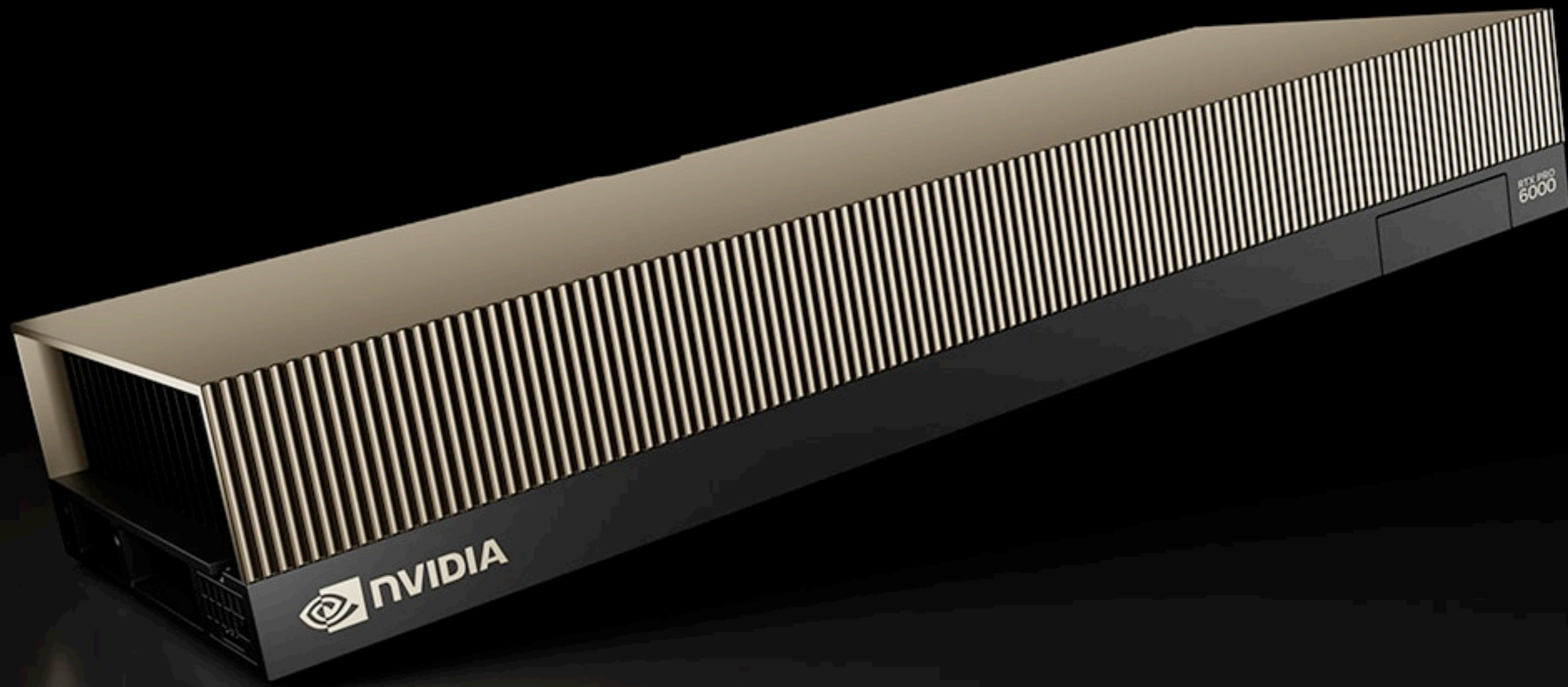
Sélection

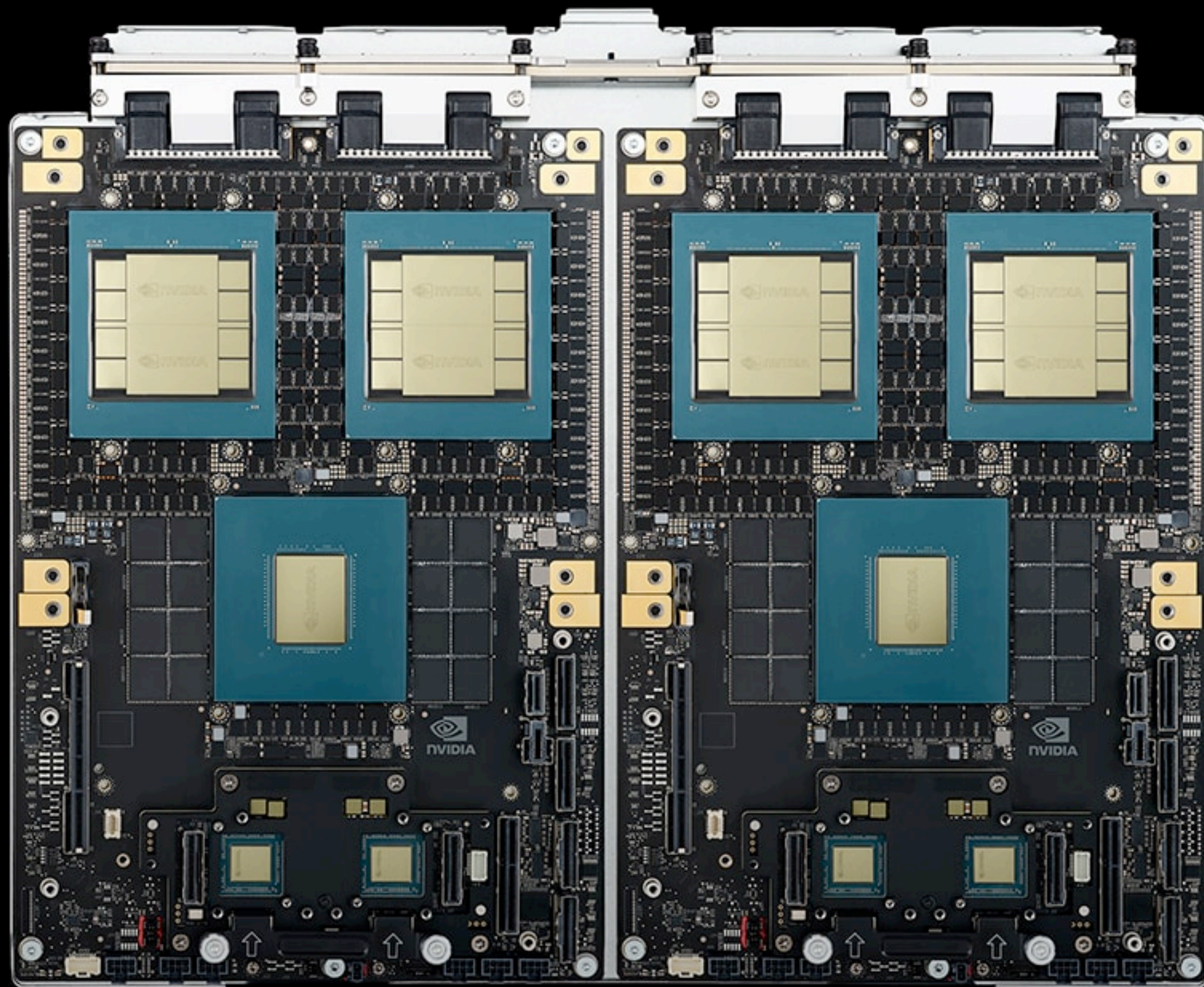
Concentration sur les 10 dernières années
Insistance sur les modèles de langage

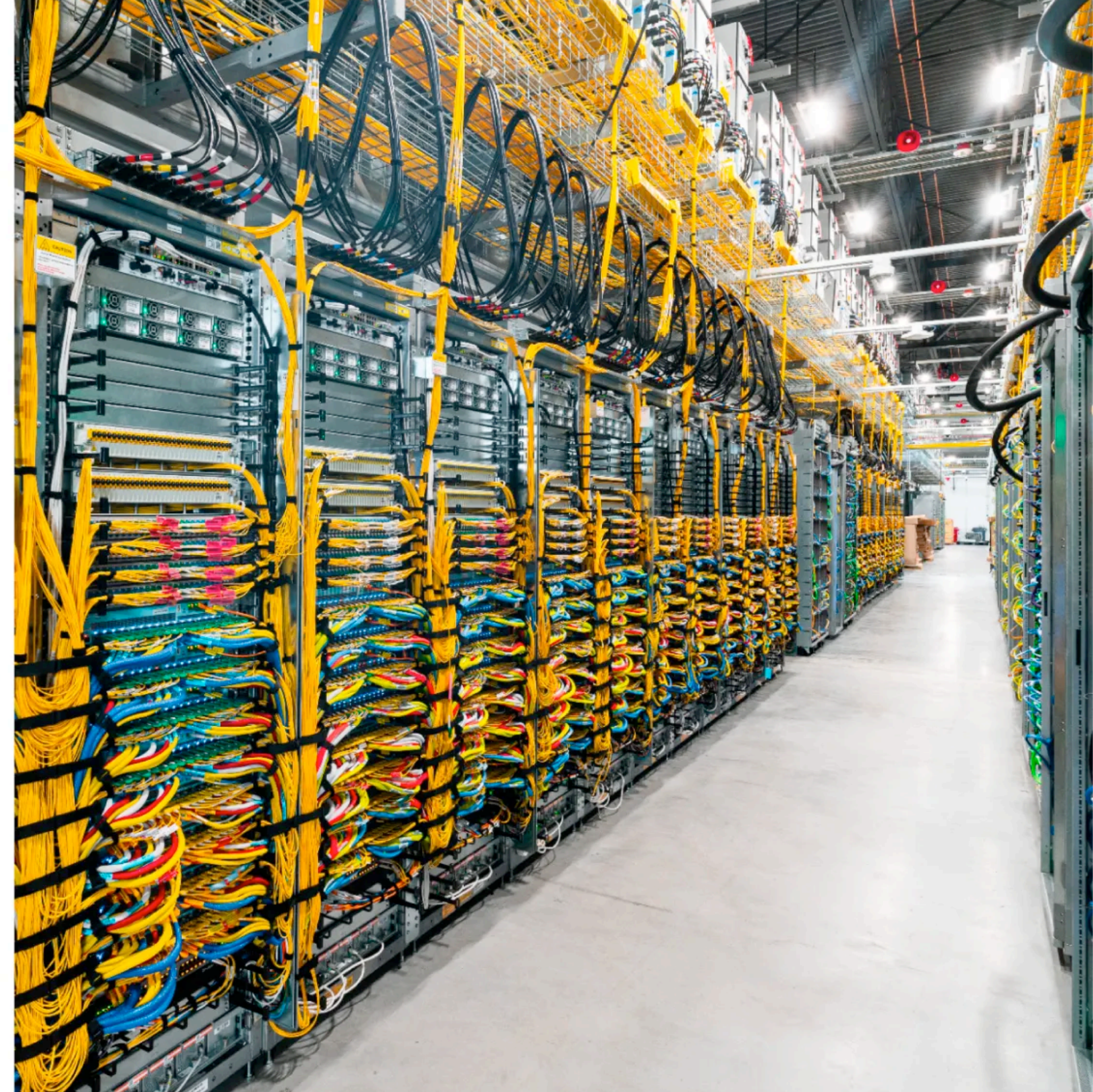
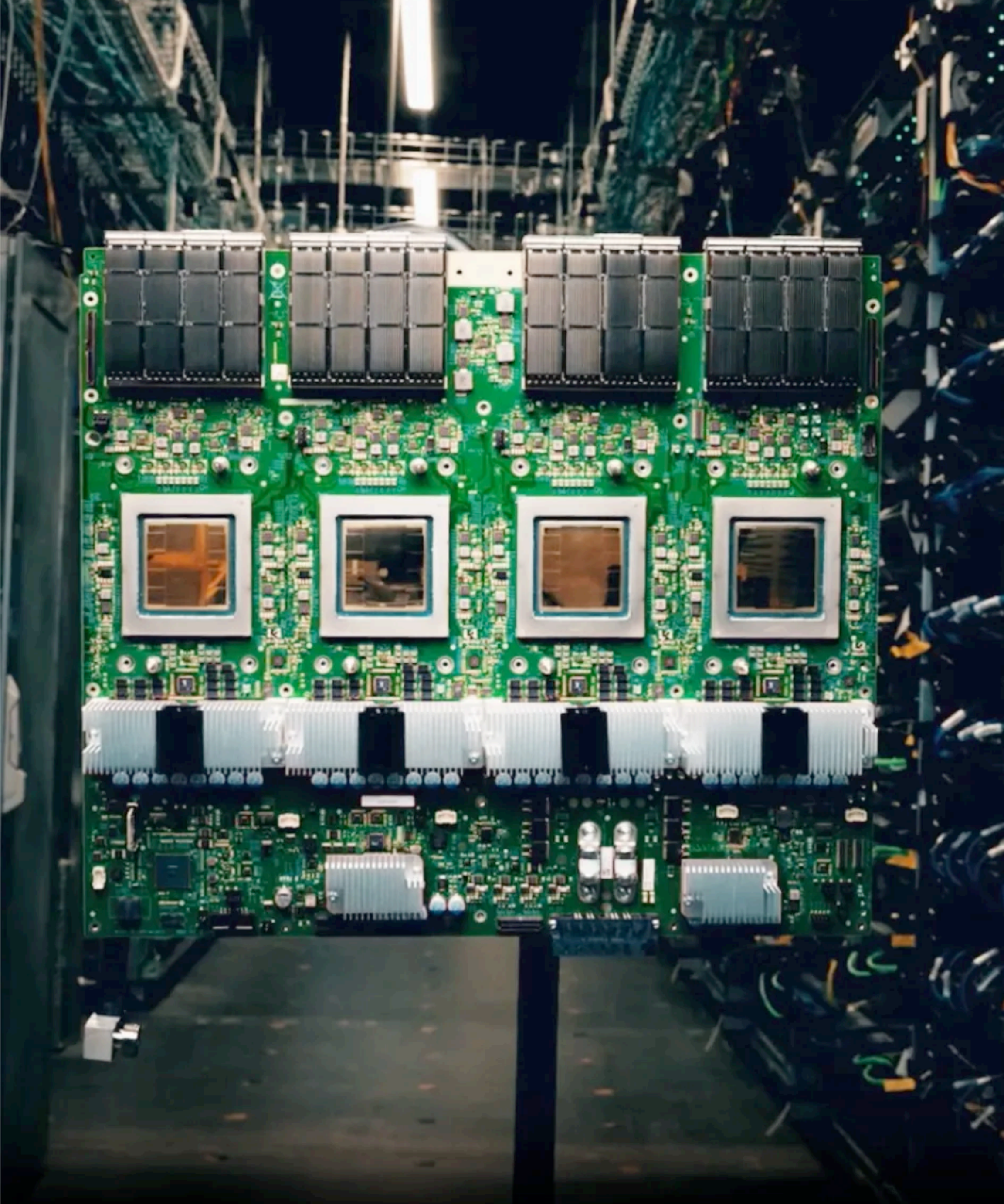
→ **Qualifier concrètement les évolutions récentes de l'IA**

Des infrastructures dédiées

Croissance exponentielle des capacités de traitement













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)

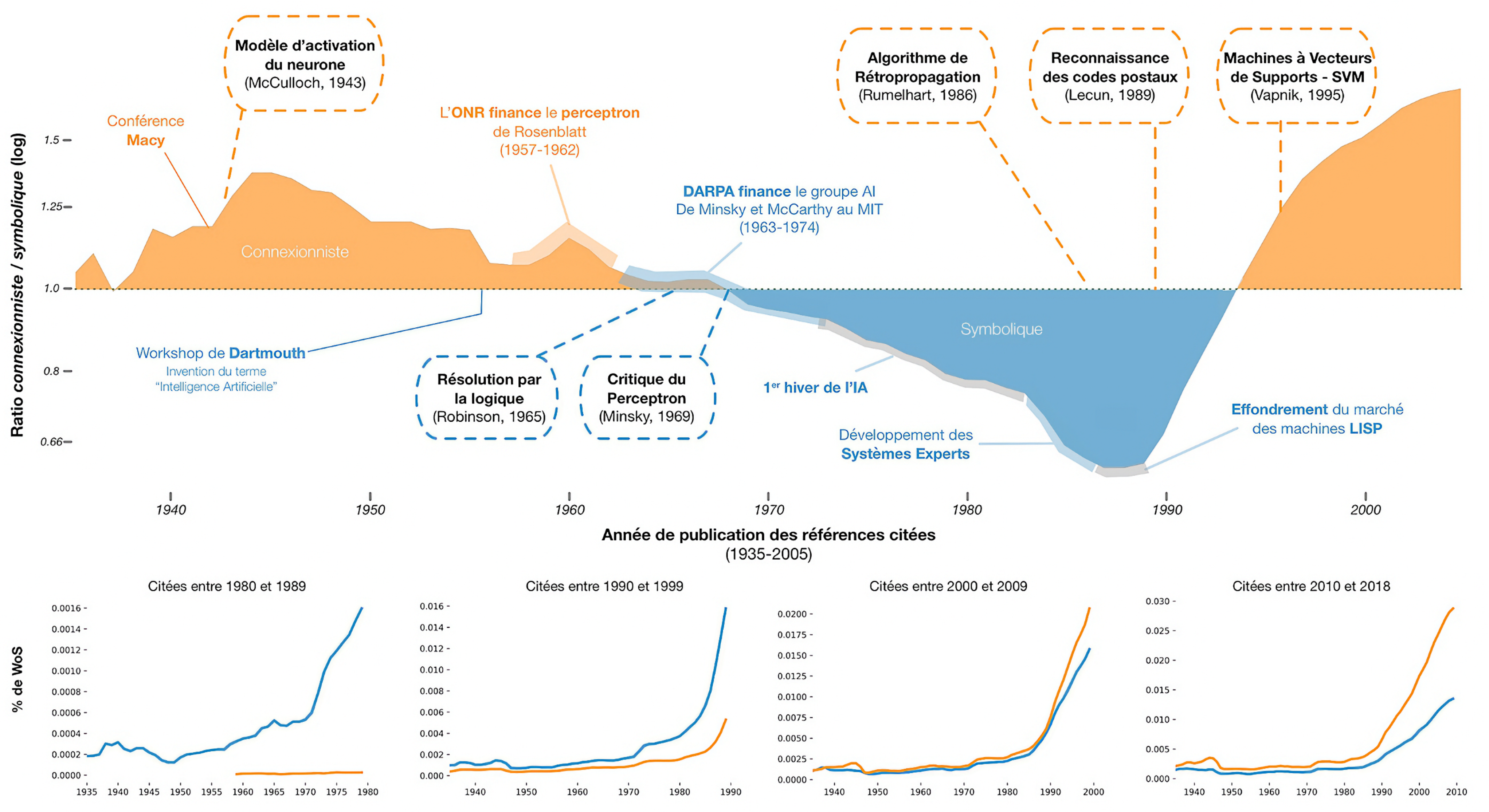
Dataset	Année	Taille	Thématique / Contenu	Source	Ouvert	Utilisé par
Anthropic HH-RLHF	2022	~161K paires	Helpfulness & harmlessness : paires (réponse choisie vs rejetée)	Annotation humaine	Oui	Claude (early), recherche
InstructGPT data	2022	~33K comparaisons	Comparaisons de réponses classées (prompts 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

La revanche des neurones

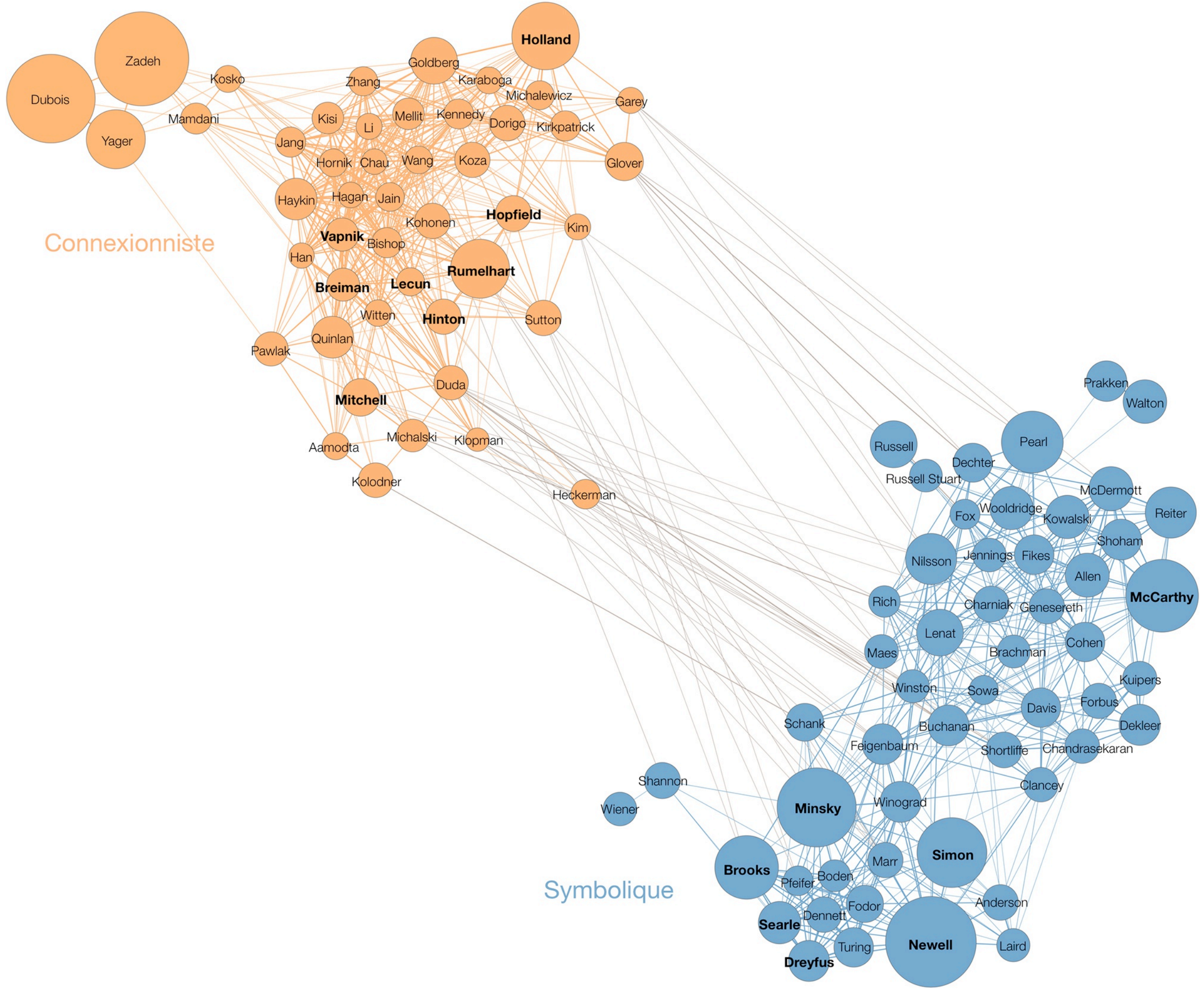
Changements de paradigmes

Changement de paradigme



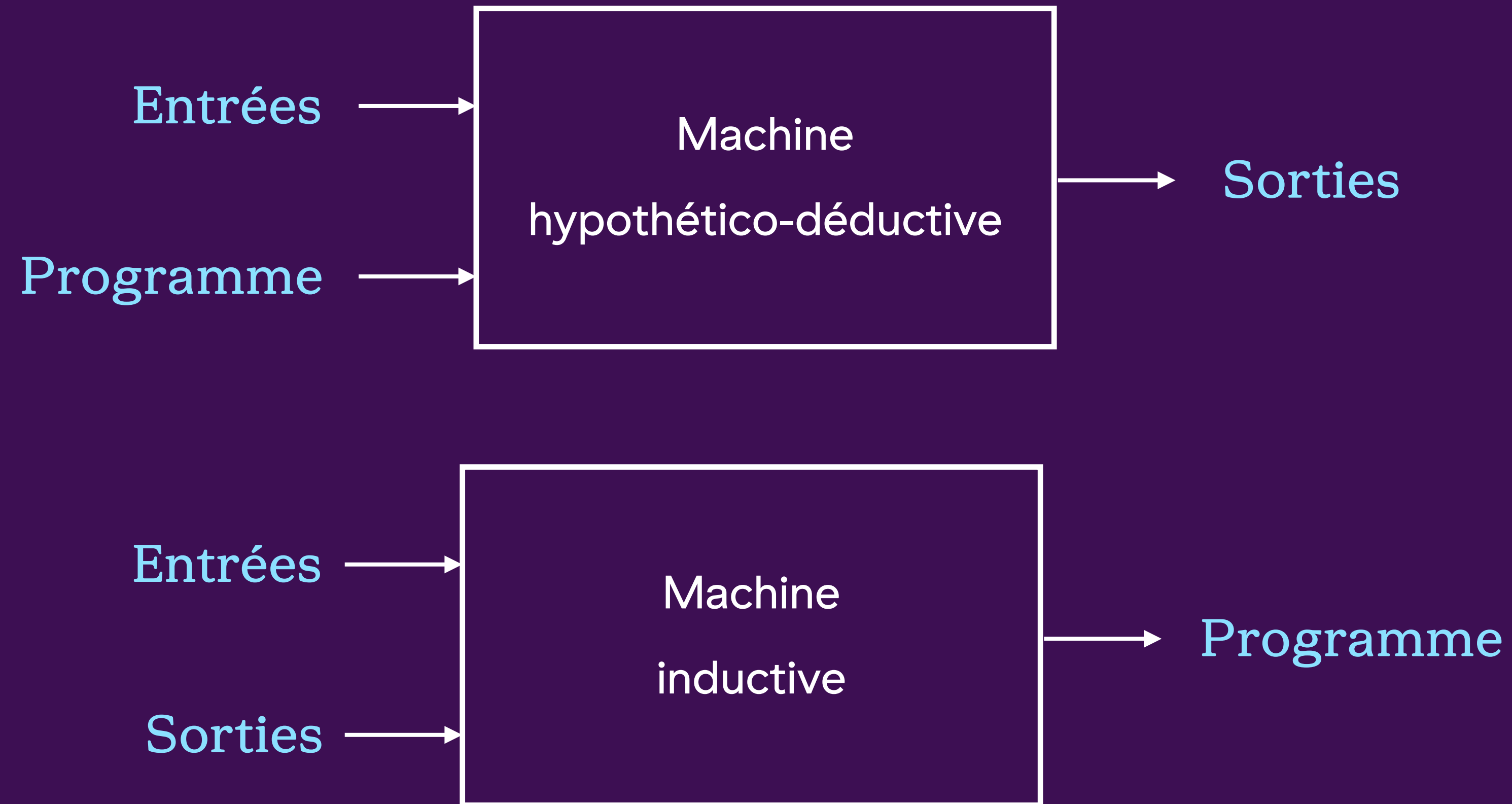
Dominique Cardon, Jean-Philippe Cointet, Antoine Mazières, *La revanche des neurones*, Réseaux 2018/5 (n° 211), pp. 173 à 220.

Changement de paradigme



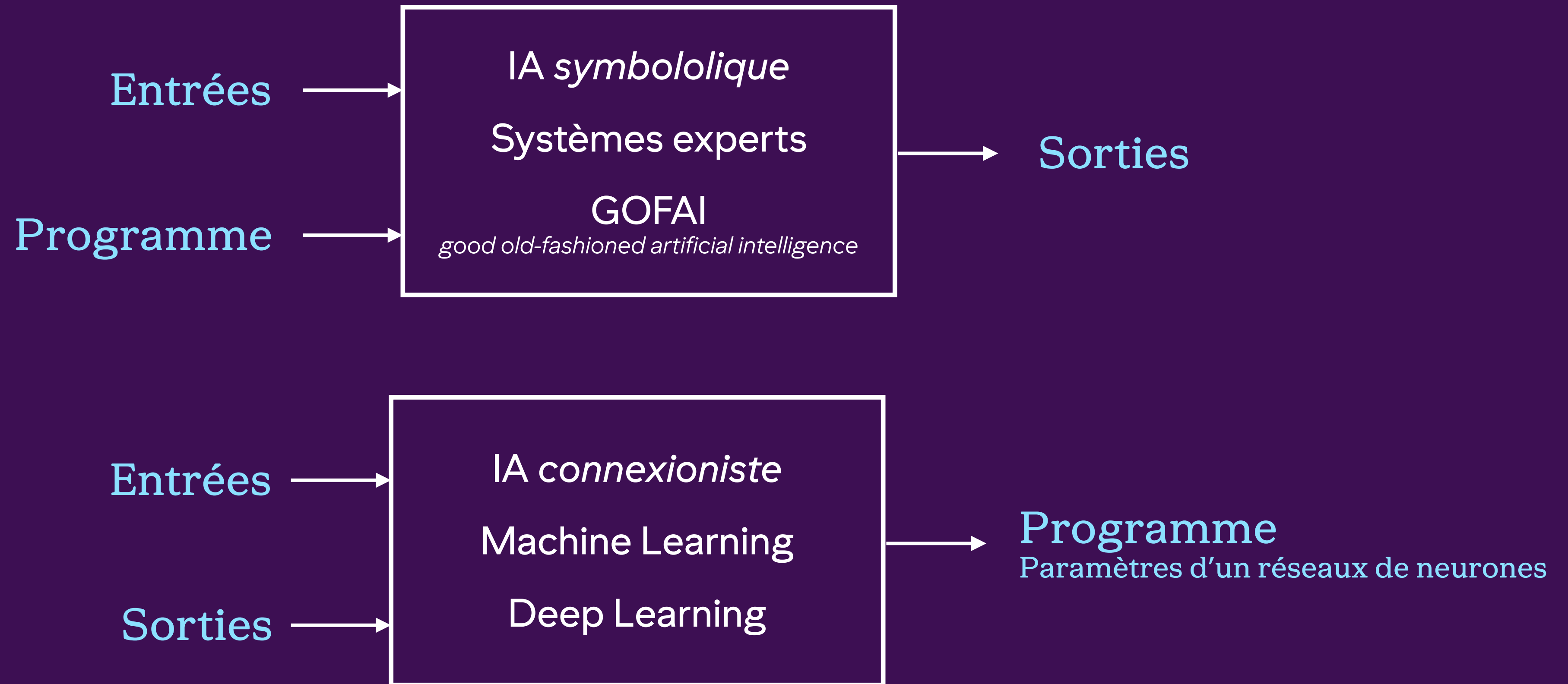
Dominique Cardon, Jean-Philippe Cointet, Antoine Mazières, *La revanche des neurones*, Réseaux 2018/5 (n° 211), pp. 173 à 220.

Changement de paradigme

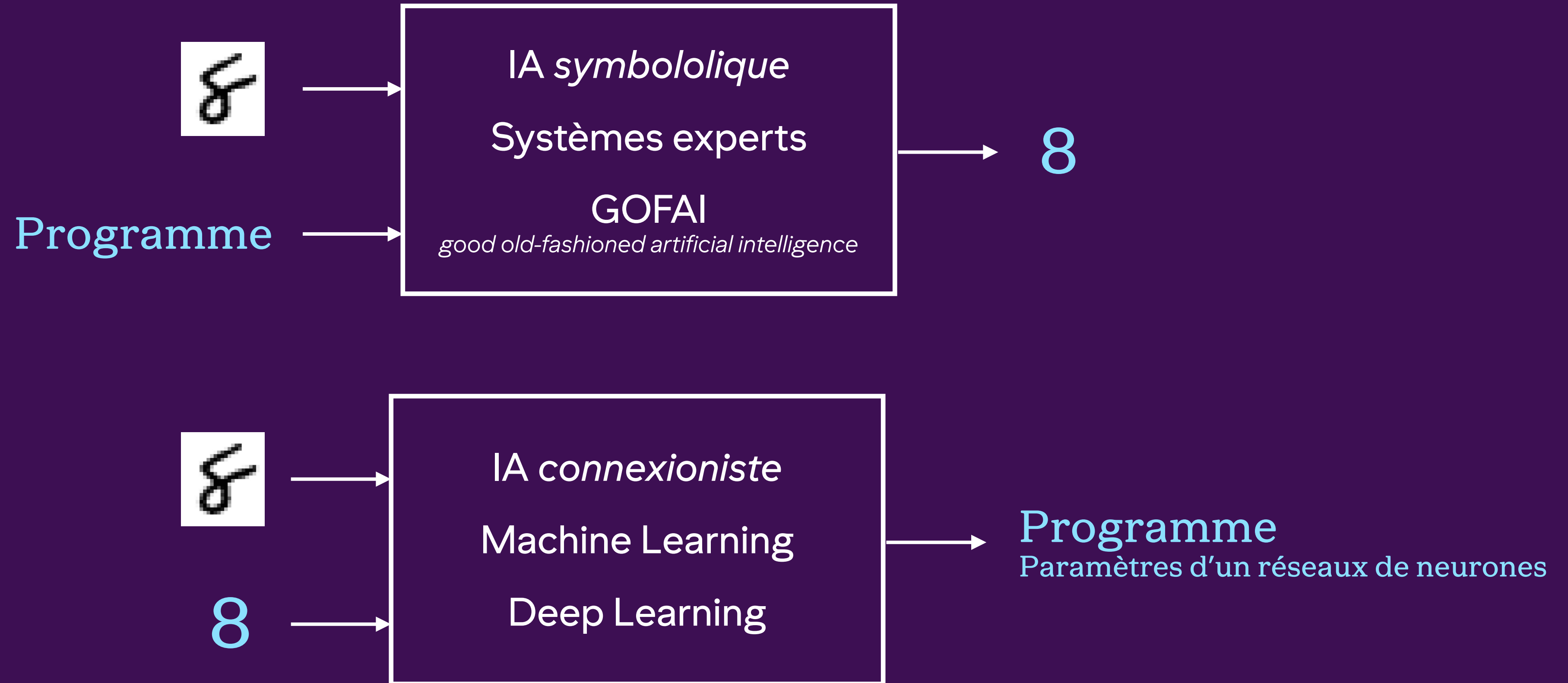


d'après : Dominique Cardon, Jean-Philippe Cointet, Antoine Mazières, La revanche des neurones, Réseaux 2018/5 (n° 211), pp. 173 à 220.

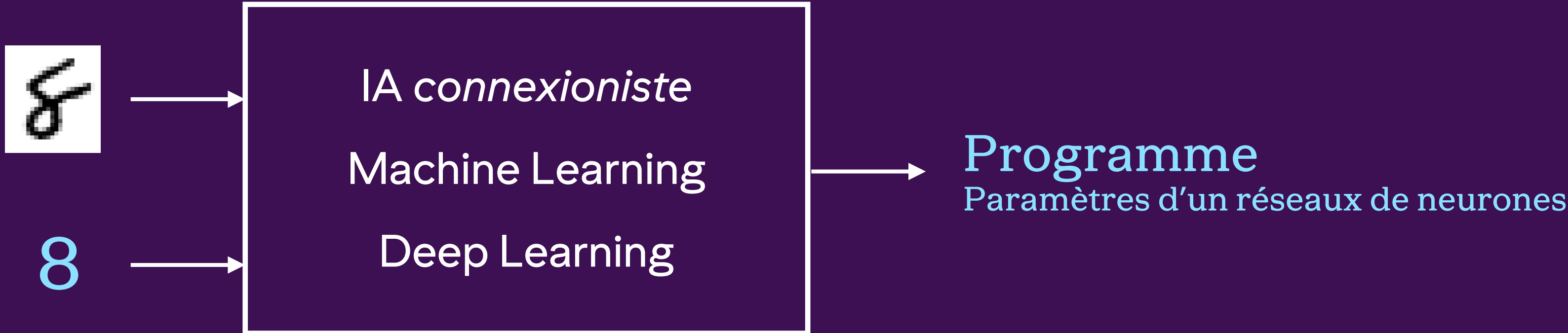
Changement de paradigme



Changement de paradigme

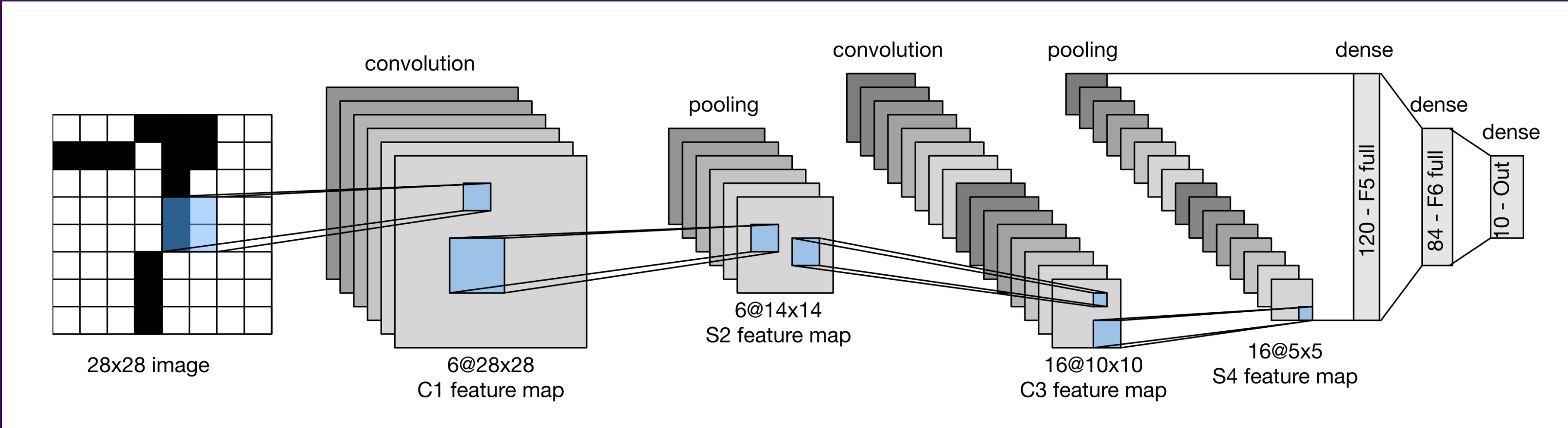
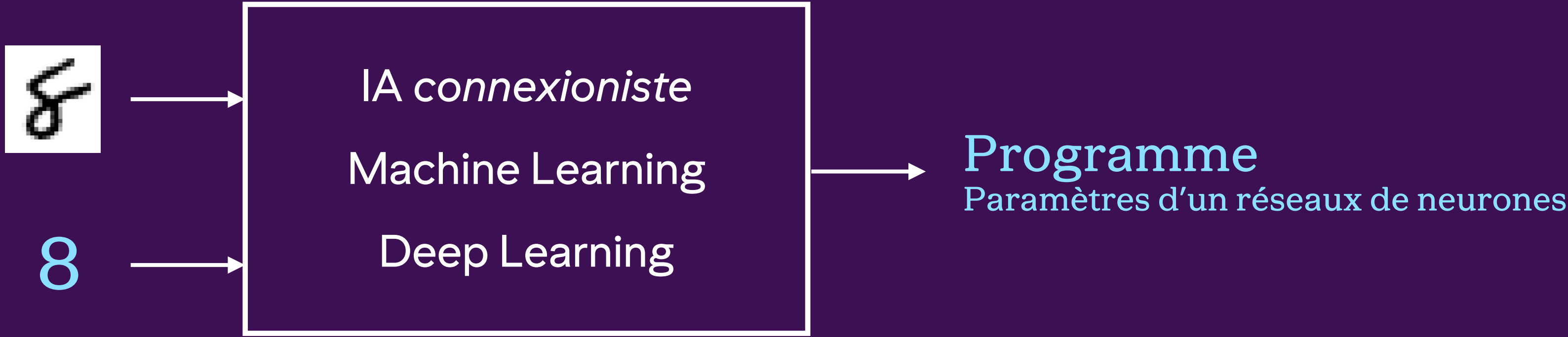


Changement de paradigme



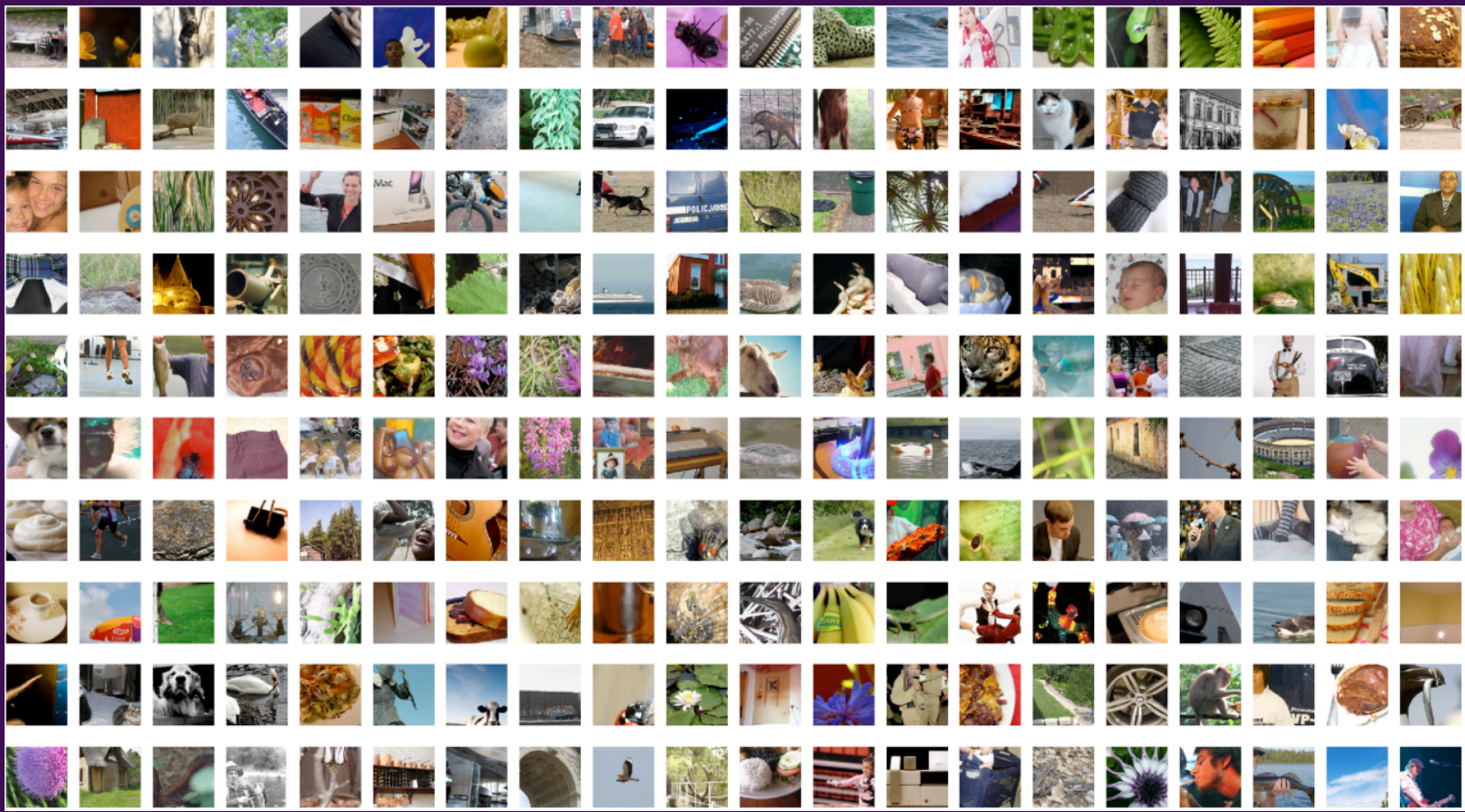
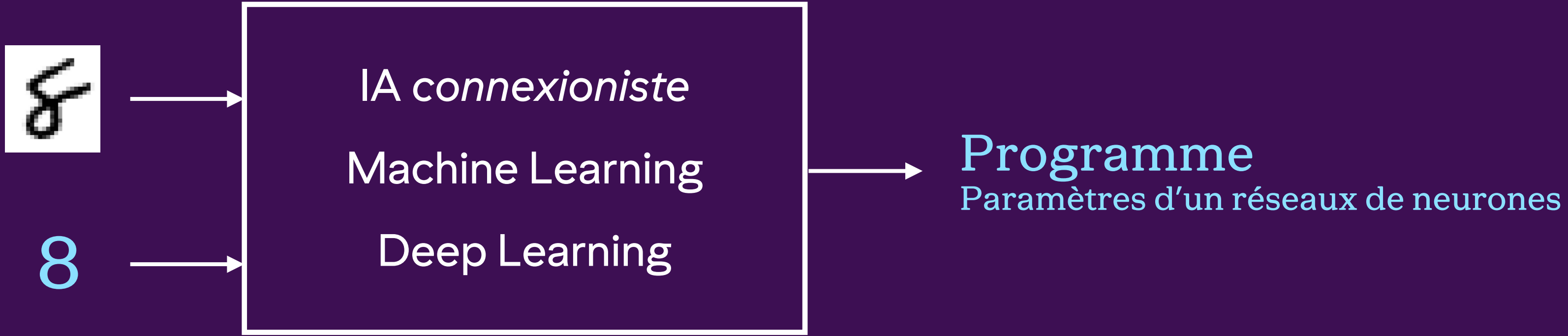
MNIST dataset - 1994

Changement de paradigme

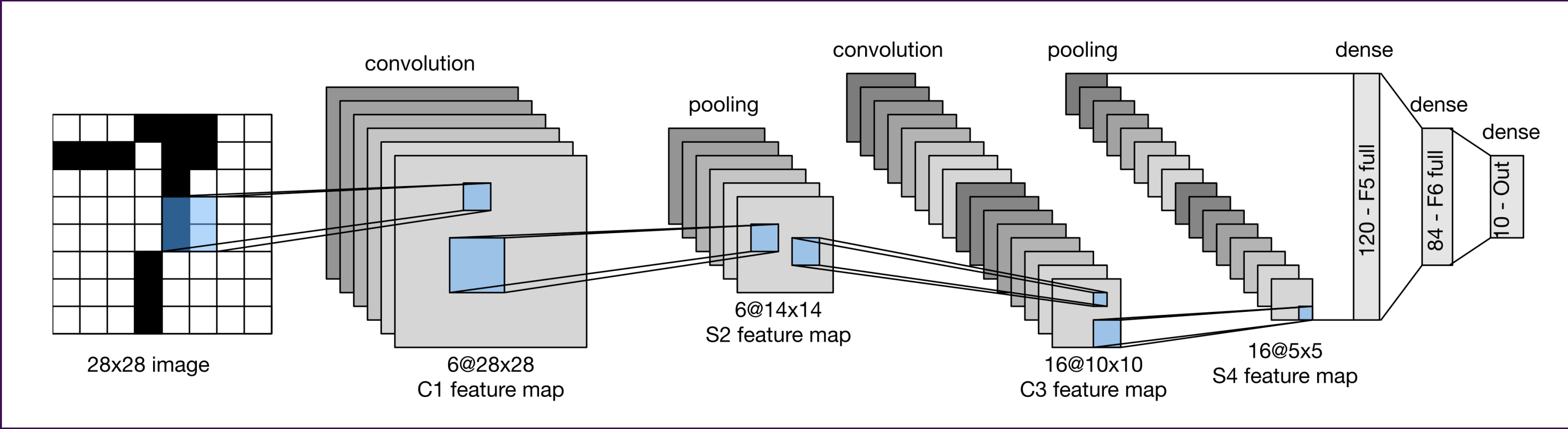


LeNet 1989 - Lecun - Représentation de Zhang, Aston and Lipton, Zachary C. and Li, Mu and Smola, Alexander J. 2023

Changement de paradigme



ImageNet - 2009-2021

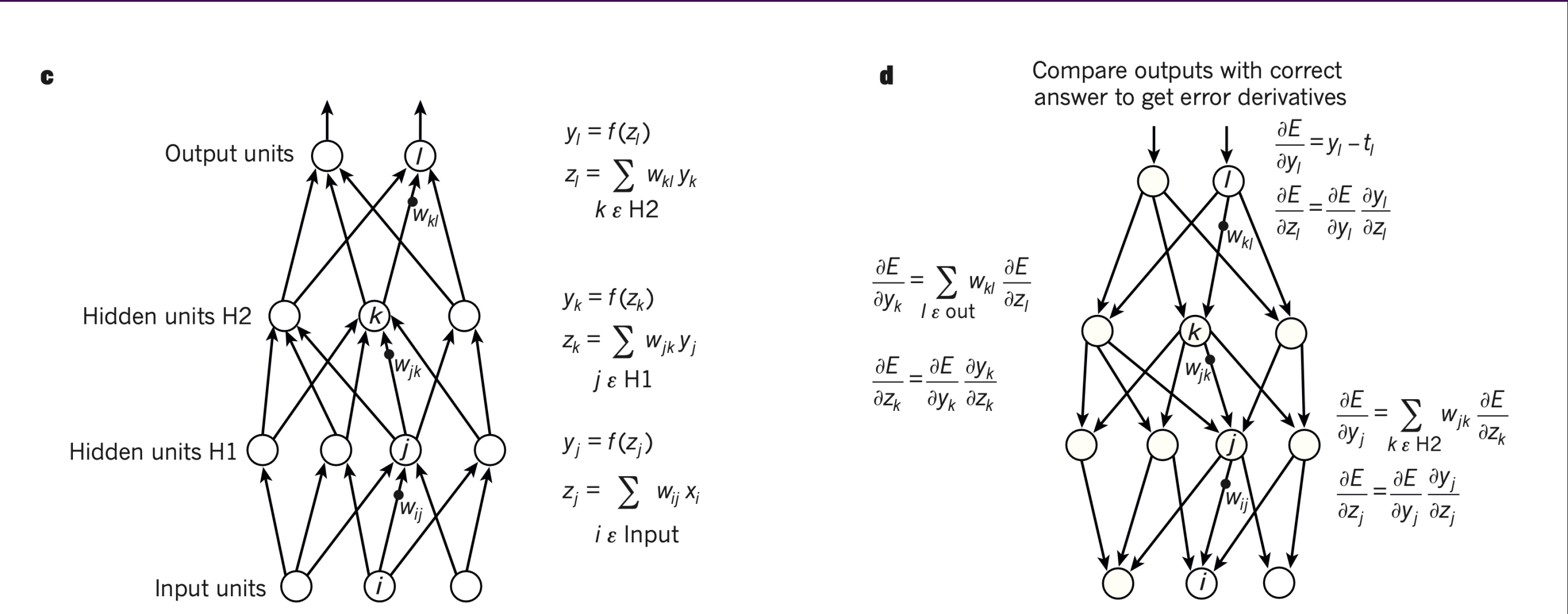


LeNet 1989 - Lecun - Représentation de Zhang, Aston and Lipton, Zachary C. and Li, Mu and Smola, Alexander J. 2023

Changement de paradigme



Chat

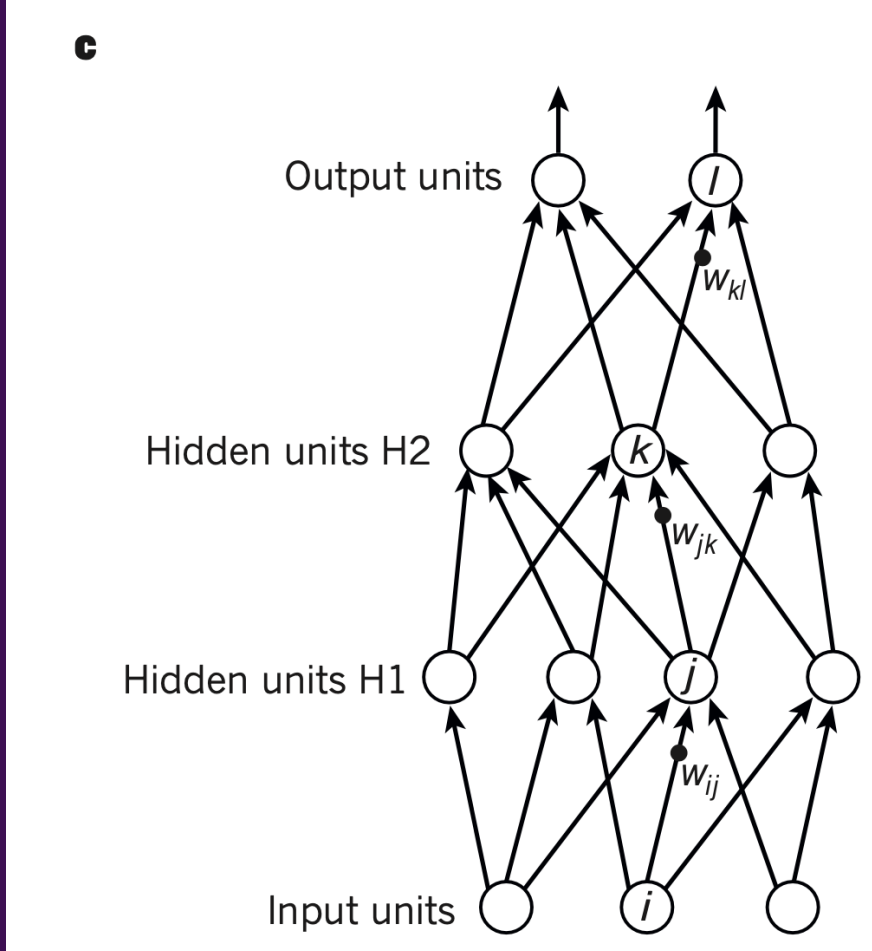


Yann LeCun, Yoshua Bengio & Geoffrey Hinton, *Deep learning*, Nature 2015 (vol.521), pp. 436-444.

Changement de paradigme



Chat



$$y_l = f(z_l)$$

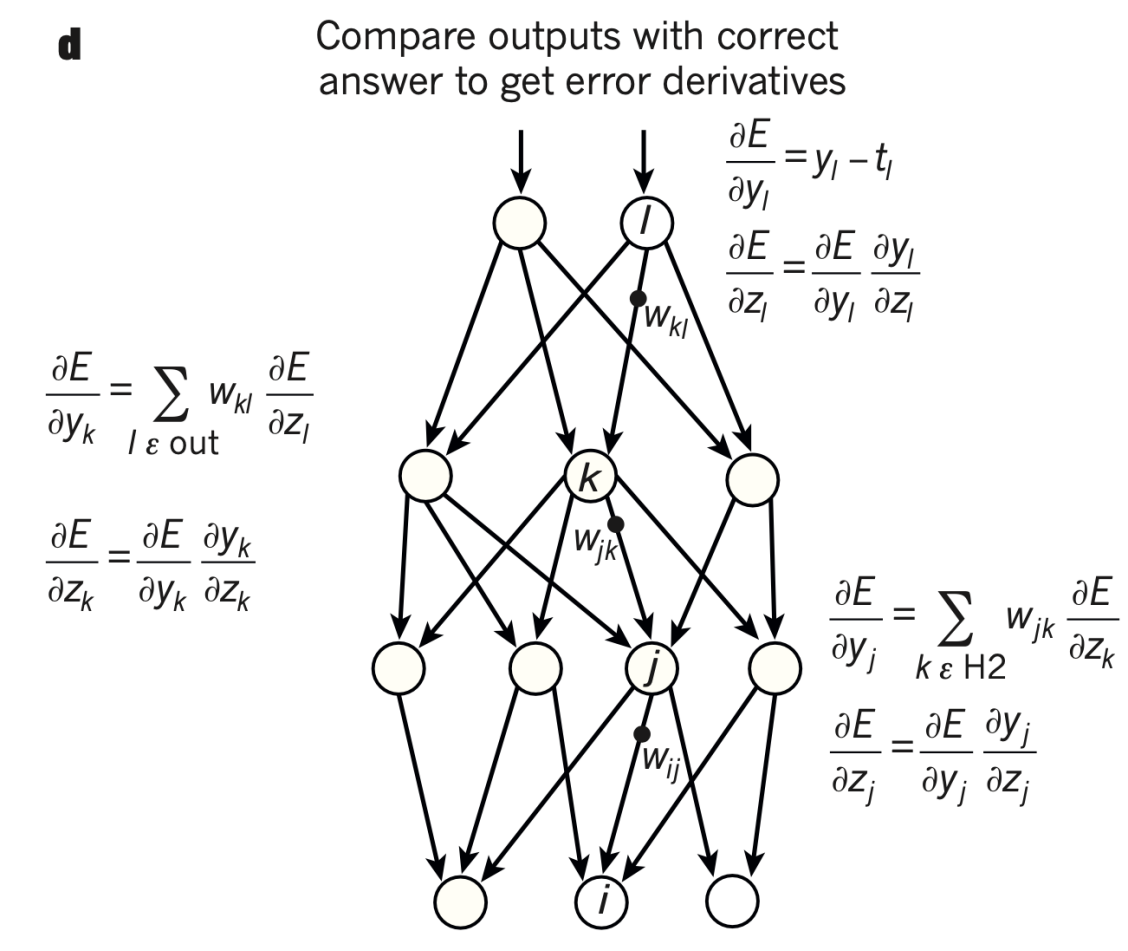
$$z_l = \sum_{k \in H2} w_{kl} y_k$$

$$y_k = f(z_k)$$

$$z_k = \sum_{j \in H1} w_{jk} y_j$$

$$y_j = f(z_j)$$

$$z_j = \sum_{i \in \text{Input}} w_{ij} x_i$$



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

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

$$\frac{\partial E}{\partial y_k} = \sum_{l \in \text{out}} w_{kl} \frac{\partial E}{\partial z_l}$$

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

$$\frac{\partial E}{\partial y_j} = \sum_{k \in H2} w_{jk} \frac{\partial E}{\partial z_k}$$

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

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

REVIEW

Deep learning

Yann LeCun¹, Yoshua Bengio² & Geoffrey Hinton³

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 architectures have become a powerful paradigm for practical data analysis, and have achieved new benchmarks in performance on several key tasks. This review surveys these architectures and the applications of deep learning to various fields of study. It also covers the theoretical and algorithmic advances that have driven the success of deep learning, and discusses the challenges that remain. Deep learning has been applied to a wide range of tasks, including image and audio processing, machine translation, natural language processing, speech recognition, and many others. It has also been used in a variety of other domains, such as drug discovery and genomics.

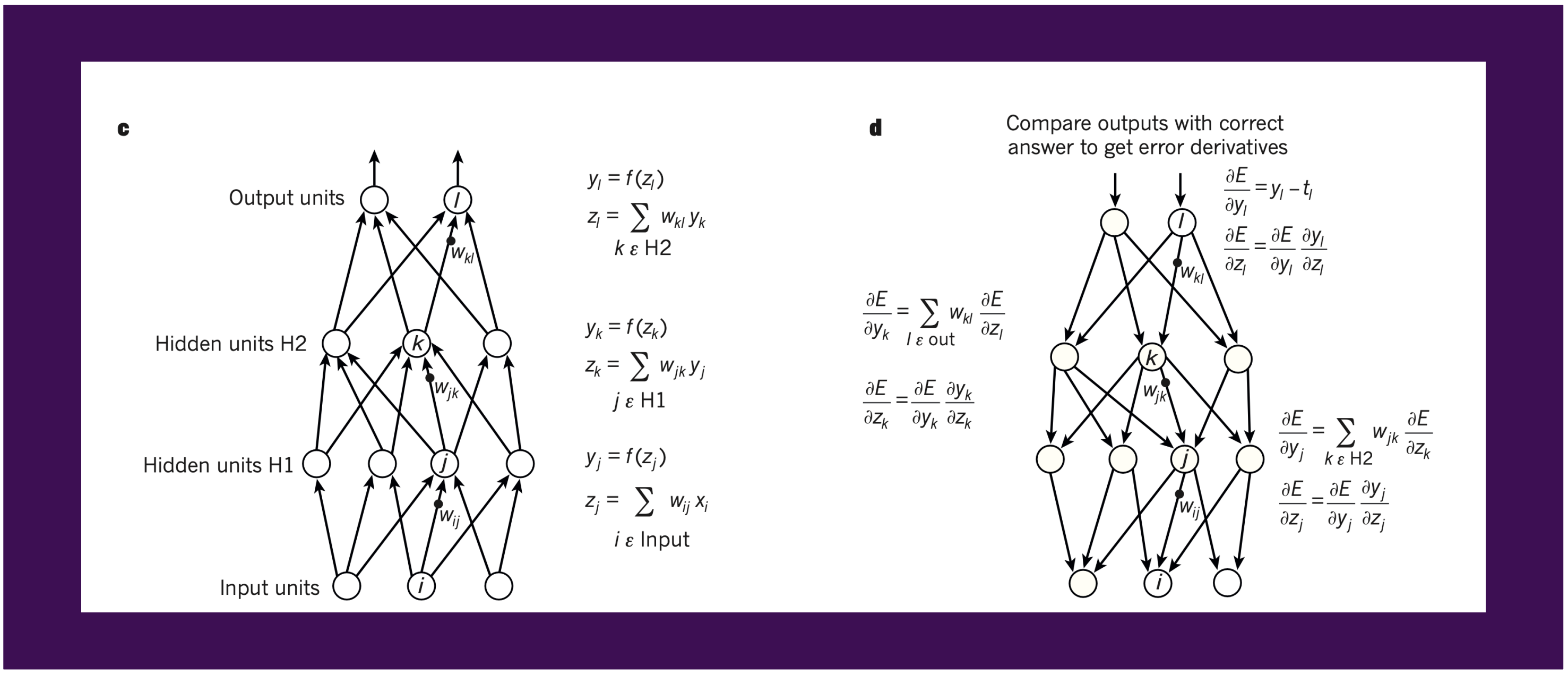
Machine learning technology grows every aspect of modern society, from healthcare to education. It is used in a wide range of applications, from recommendation systems to autonomous vehicles. Machine learning systems are used to identify objects in images, transcribe speech into text, and to predict the effectiveness of drugs. Deep learning has been a major driver of this progress, and has led to significant advances in many domains. This review surveys the state-of-the-art in deep learning, and discusses the challenges that remain. It also covers the theoretical and algorithmic advances that have driven the success of deep learning, and discusses the challenges that remain.

Supervised learning is the most common form of machine learning, and is used to train models to perform tasks such as classification and regression. It involves feeding a model a set of input-output pairs, and training it to learn a function that maps the input to the output. Deep learning has led to significant advances in supervised learning, and has achieved new benchmarks in many domains. This review surveys the state-of-the-art in supervised learning, and discusses the challenges that remain. It also covers the theoretical and algorithmic advances that have driven the success of deep learning, and discusses the challenges that remain.

¹Department of Computer Science, University of Toronto, Toronto, Ontario M5S 2E4, Canada. ²Department of Computer Science, University of Montreal, Montreal, Quebec H3T 3J4, Canada. ³Department of Computer Science, University of Toronto, Toronto, Ontario M5S 2E4, Canada.

Changement de paradigme

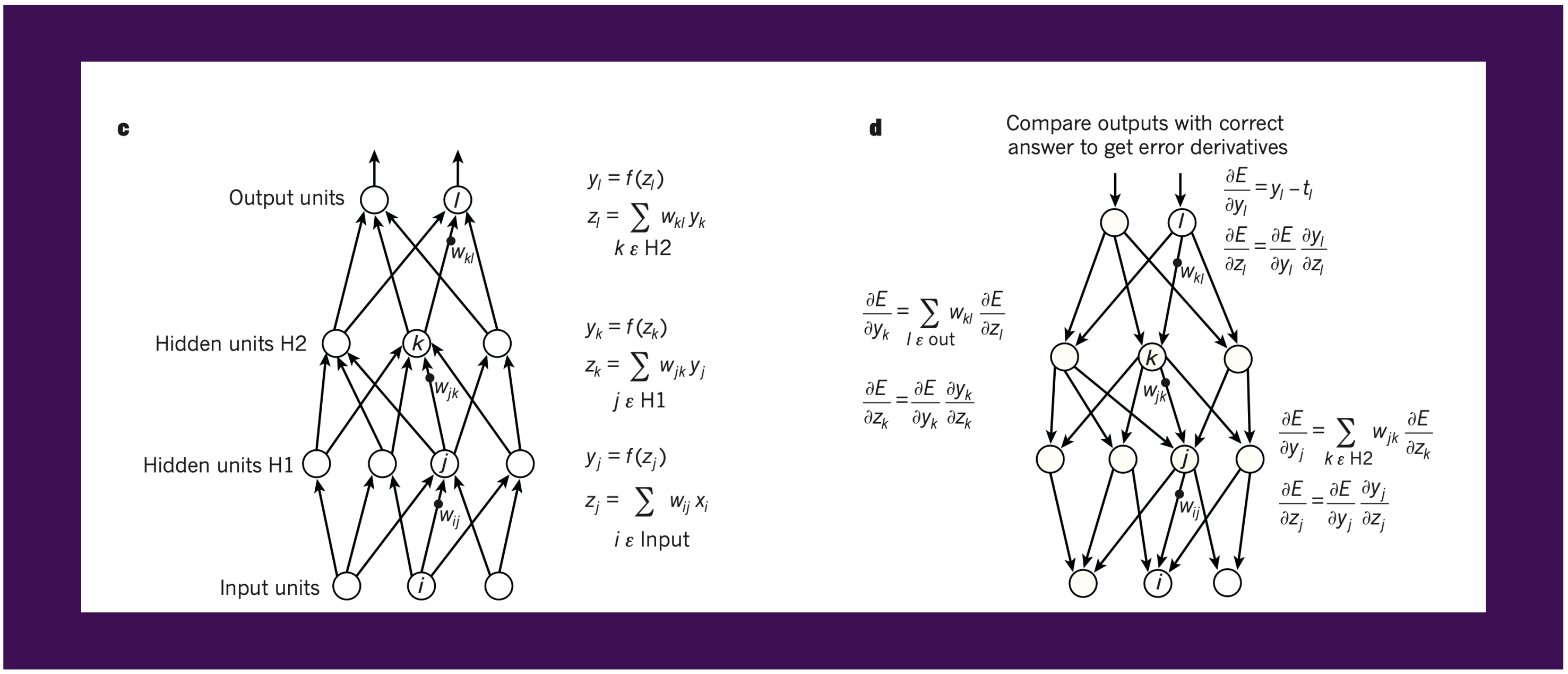
Apprentissage



Chat

Changement de paradigme

Inférence

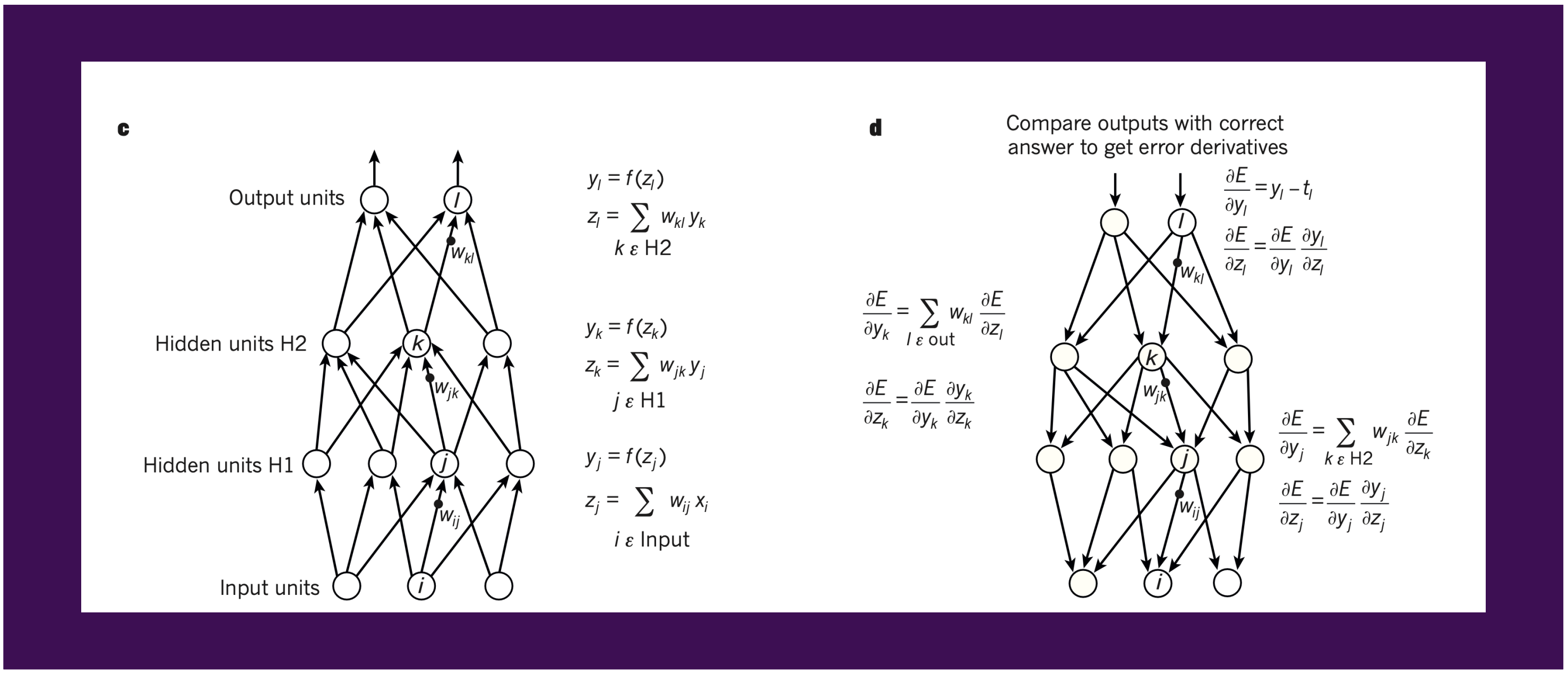


Chat

Changement de paradigme

Apprentissage

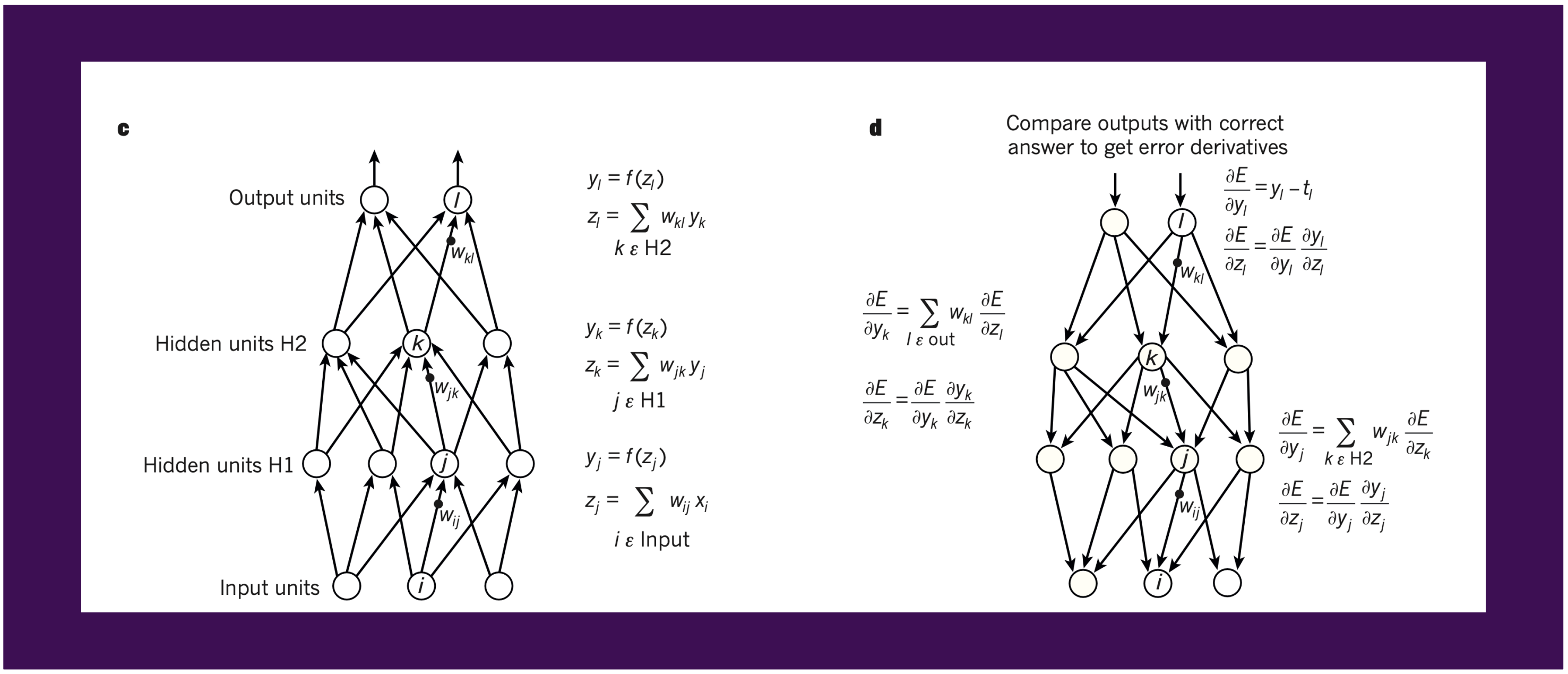
Chat



Changement de paradigme

Inférence

Chat



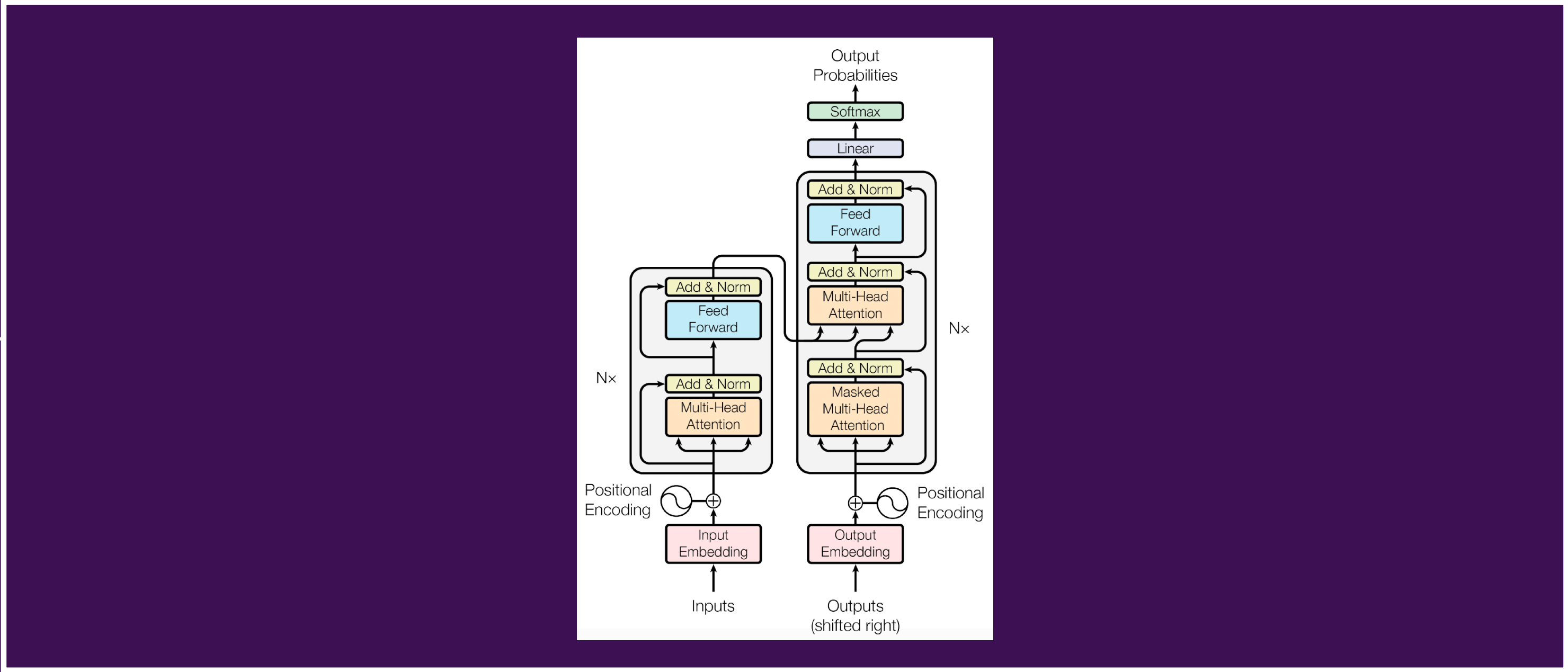
Attention is all you need

Changement d'échelle

Changement de paradigme

Apprentissage

C'est un chat gris



It is a grey cat.

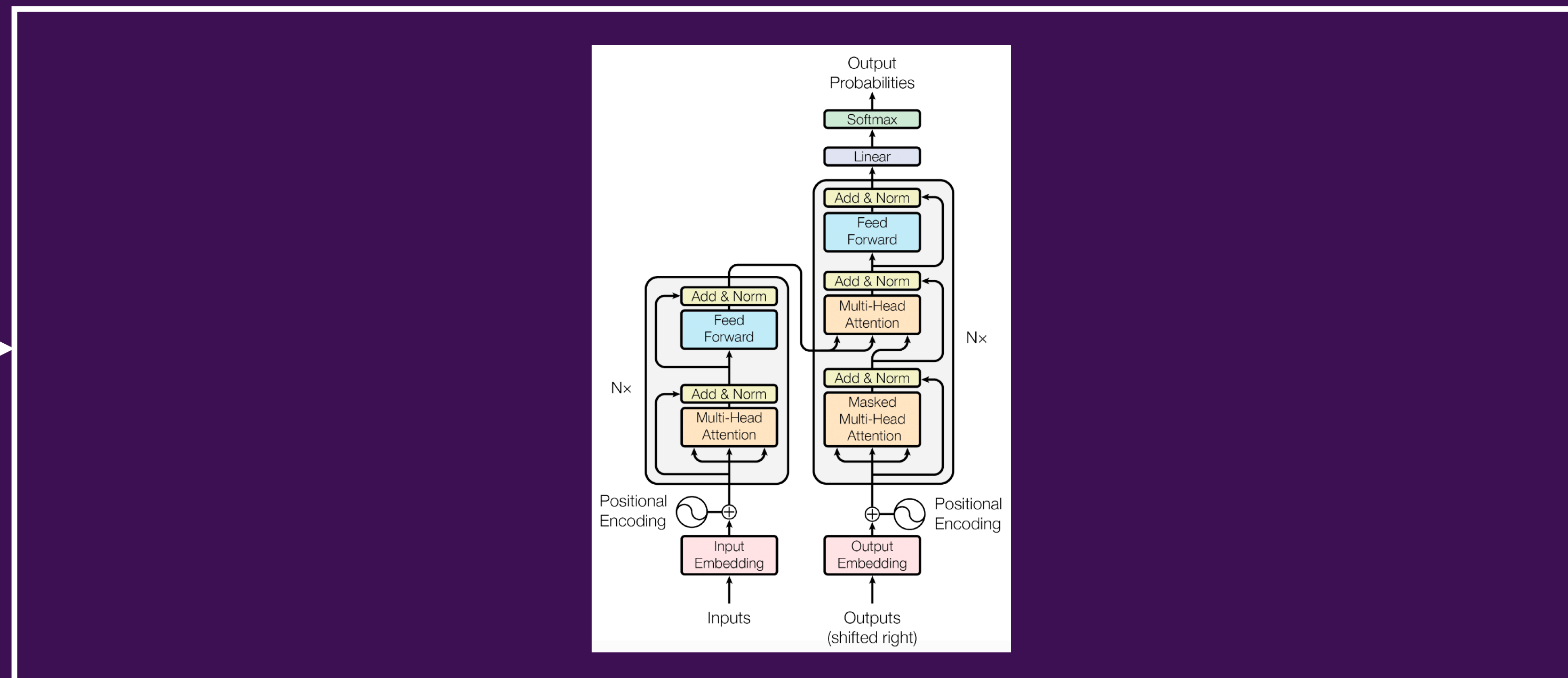


Transformers

Changement de paradigme

Inférence

C'est un chat gris



It is a grey cat.

Transformers

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

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
lukaszkaizer@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.

Embedding

Efficient Estimation of Word Representations in Vector Space

Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean

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 sets, 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*, Noam Shazeer*, Niki Parmar*, Jakob Uszkoreit*, Llion Jones*, Aidan N. Gomez¹, Lukasz Kaiser*, Illia Polosukhin¹

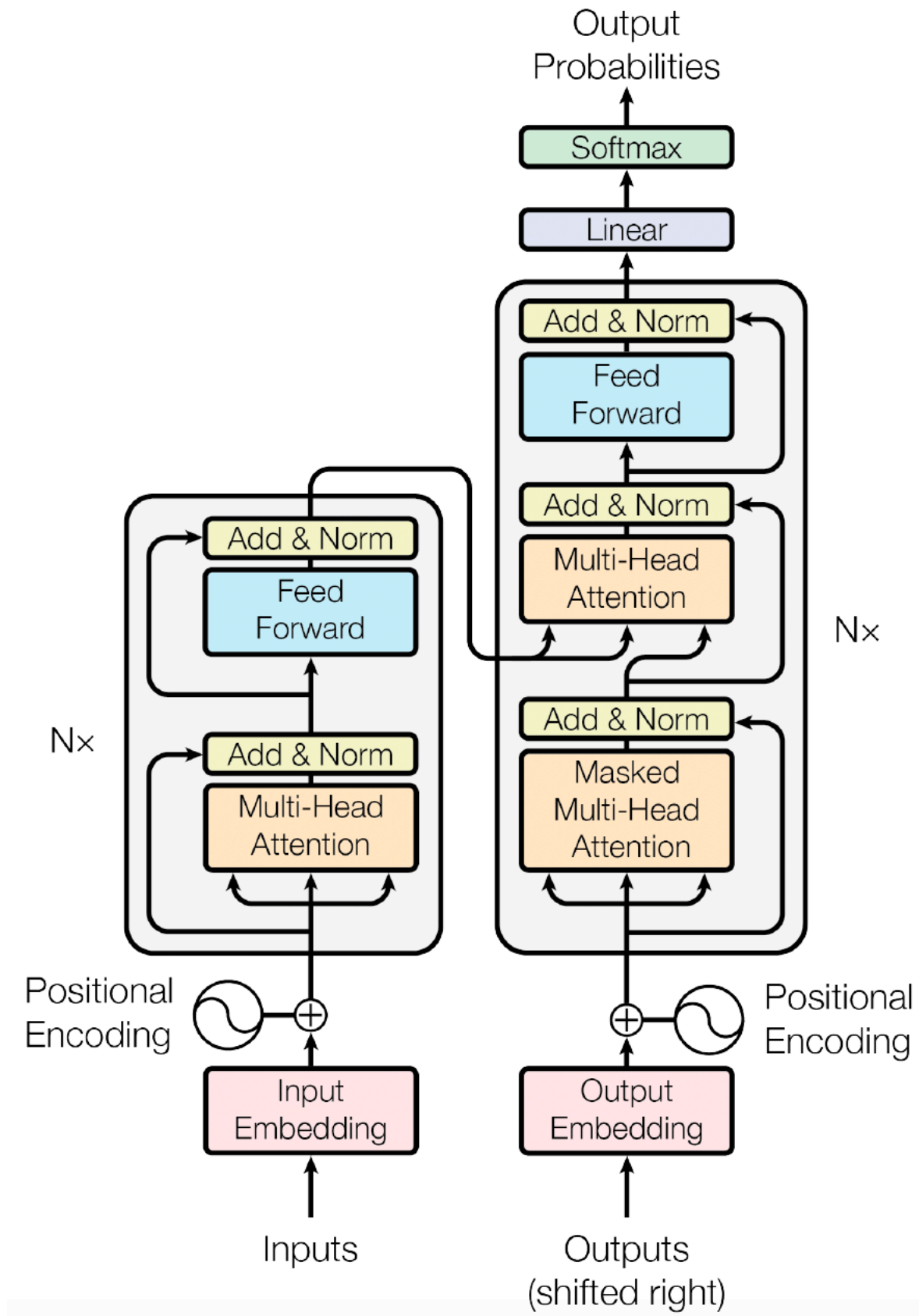
Abstract

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.
¹Work performed while at Google Brain.
²Work performed while at Google Research.

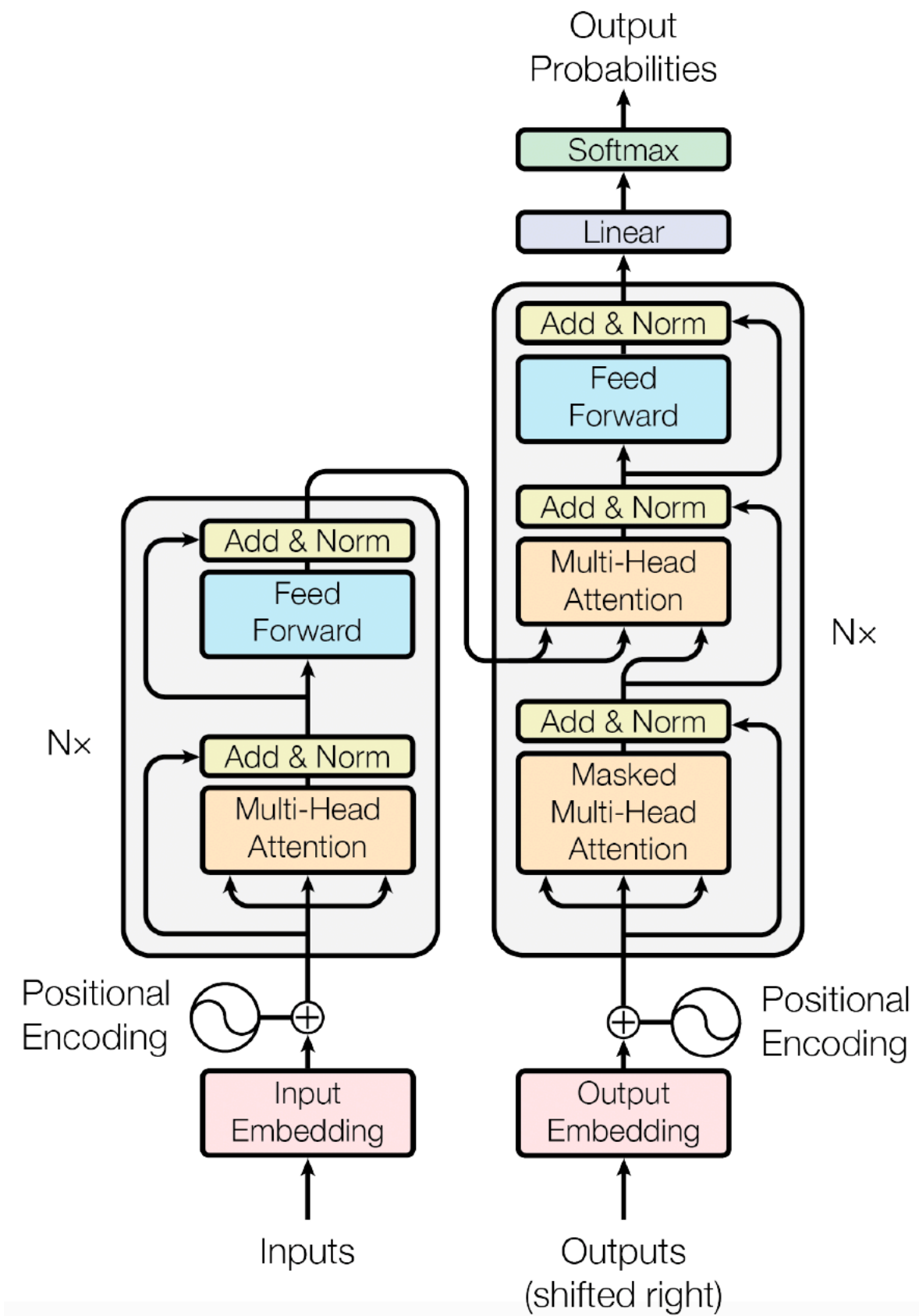
31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

2017





2017

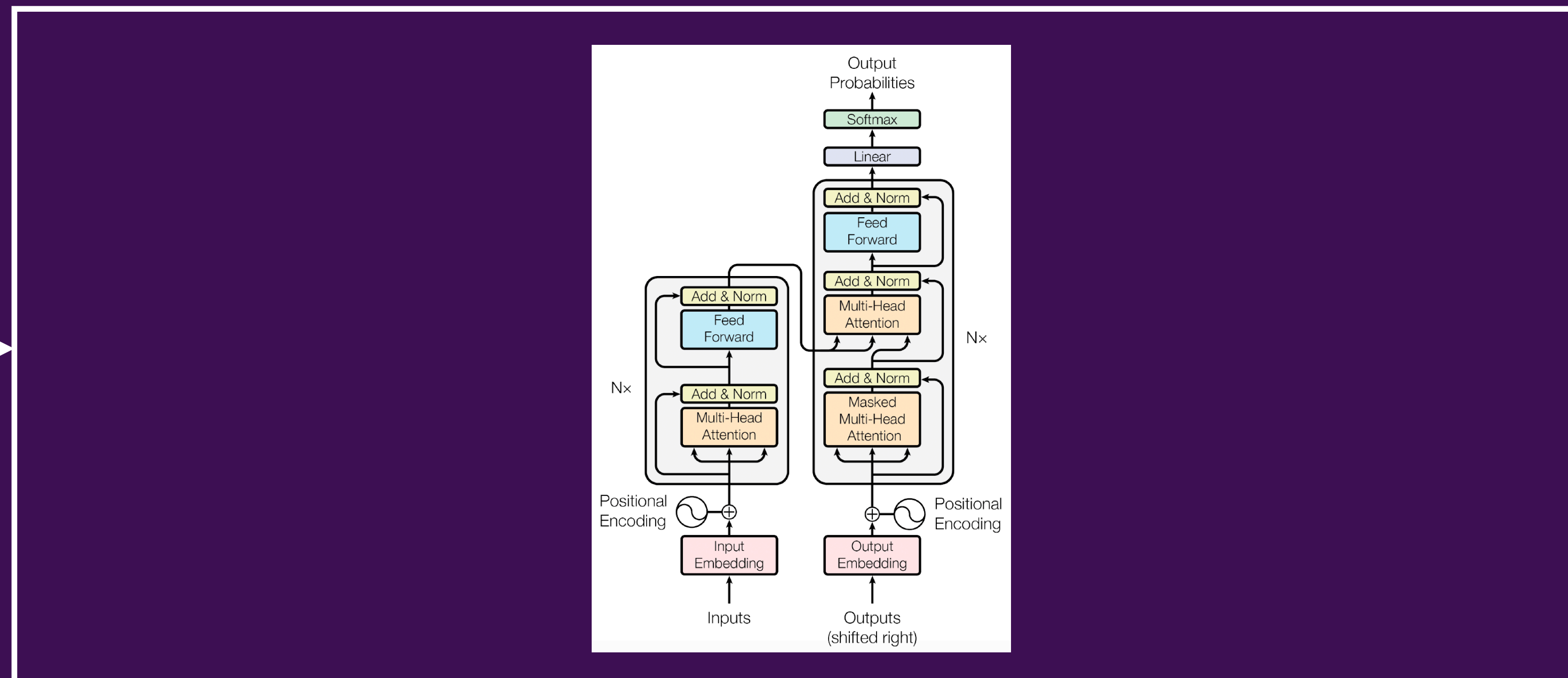


Architecture	Modèles représentatifs	Usage principal
Encodeur	ModernBERT, XLM-RoBERTa, E5, BGE, GTE	Classification, NER, embeddings, RAG
Décodeur	GPT-4o, Claude 3.7, Llama 3, Mistral, DeepSeek, Gemini	Génération, chat, raisonnement, code
Encodeur-décodeur	T5, BART, mT5, T5Gemma	Traduction, résumé, tâches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

Changement de paradigme

Inférence

C'est un chat gris



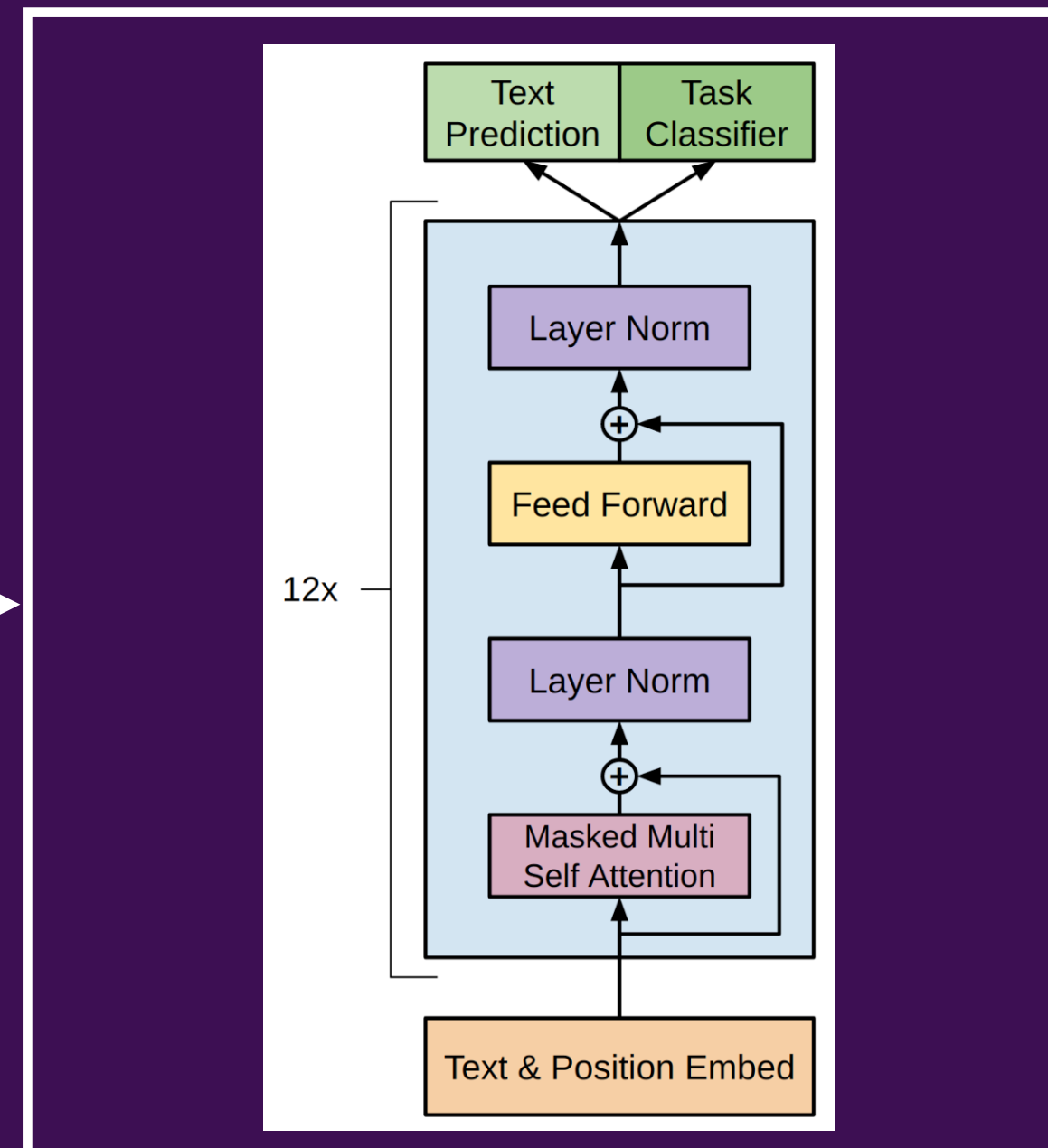
It is a grey cat.

Transformers

Exemples plus complexes

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

Inférence



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

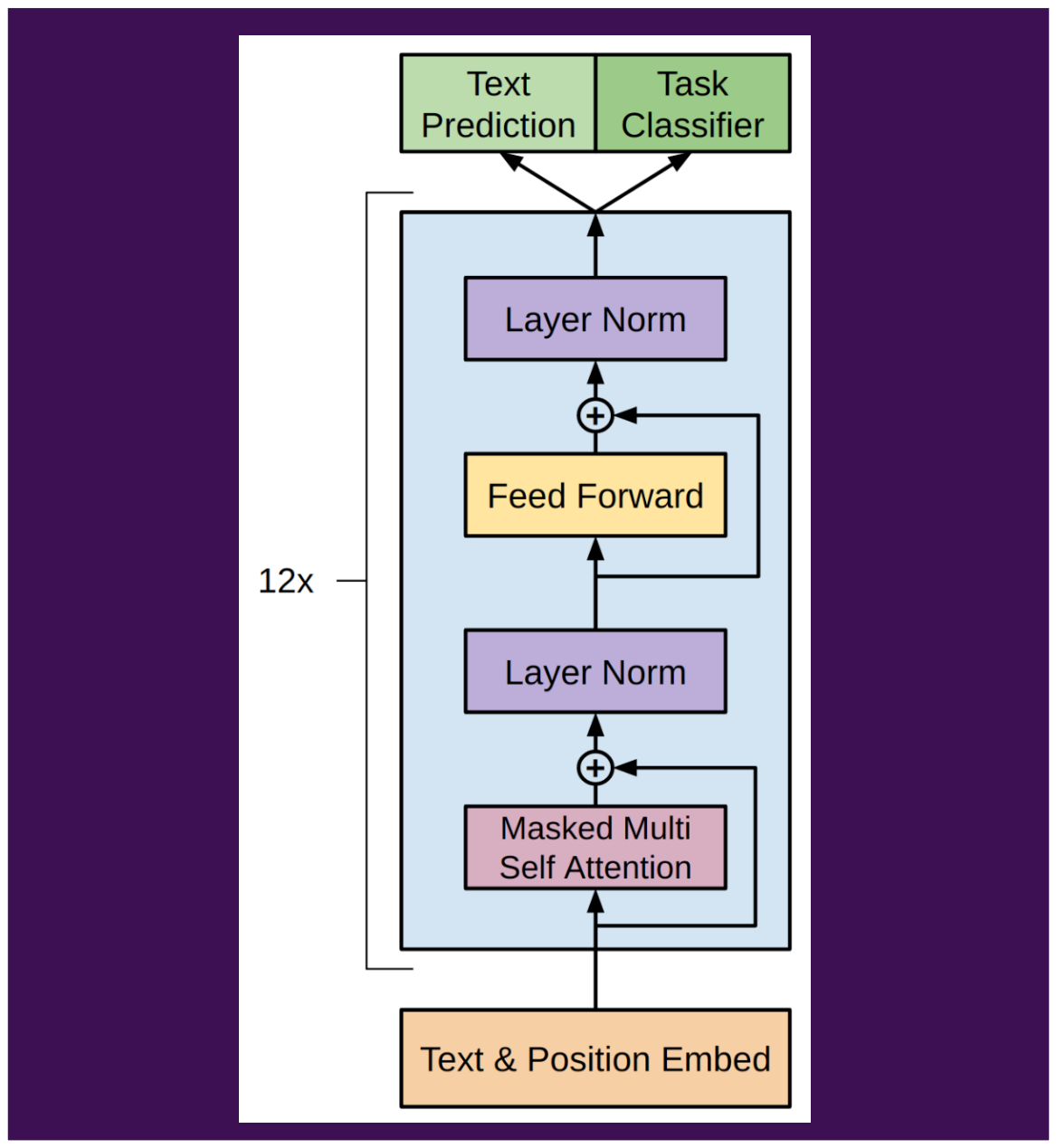
Generative Pretrained Transformers

Sens commun ?

Exemples plus complexes

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 ?

Inférence



Il faut 47 jours.

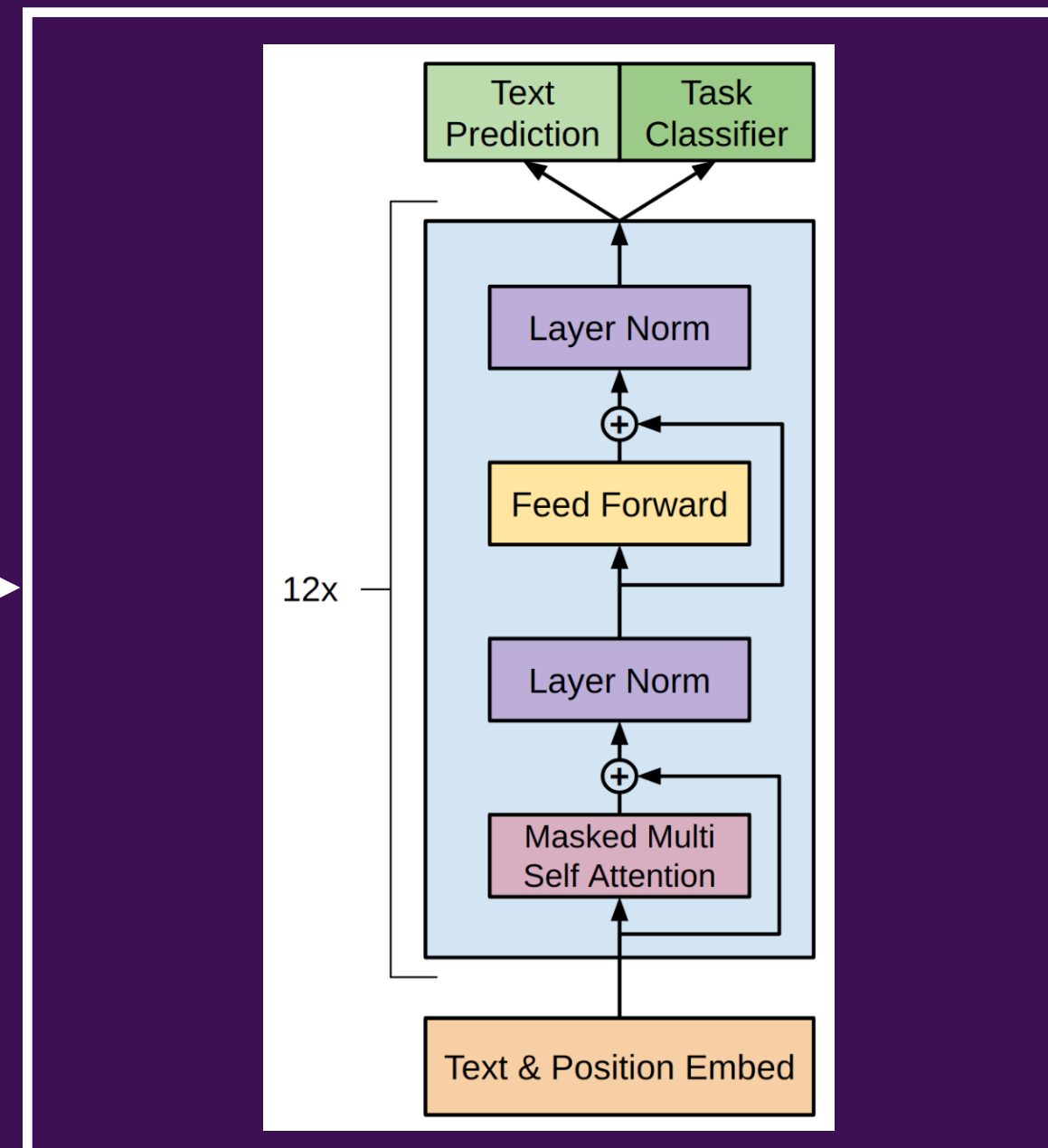
Generative Pretrained Transformers

Raisonnement ?

Exemples plus complexes

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

Inférence



Confabulation ou refus de réponse

Generative Pretrained Transformers

knowledge cutoff

Embedding

Transformers

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 sets, 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

2017

Attention Is All You Need

Ashish Vaswani*
Google Brain
avaswani@google.com

Noam Shazeer*
Google Brain
noam@google.com

Niki Parmar*
Google Research
niki@google.com

Jakob Uszkoreit*
Google Research
usz@google.com

Llion Jones*
Google Research
llion@google.com

Aidan N. Gomez*¹
University of Toronto
aidan@cs.toronto.edu

Lukasz Kaiser*
Google Brain
lukaszkaiser@google.com

Illia Polosukhin*¹
illiaa.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.

¹Work performed while at Google Brain.

²Work performed while at Google Research.

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

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

octobre 2018

Article cité plus de 15 000 fois ...

Article cité plus de 150 000 fois ...

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 sets, 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*
avaaswan@google.com

Noam Shazeer*
noam@google.com

Niki Parmar*
nikip@google.com

Jakob Uszkoreit*
usz@google.com

Llion Jones*
llion@google.com

Aidan N. Gomez[†]
aidan@cs.toronto.edu

Lukasz Kaiser*
lukaszkaiser@google.com

Illa 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
alecrad@openai.com

Karthik Narasimhan
karthikn@openai.com

Tim Salimans
tim@openai.com

Ilya Sutskever
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 **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

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.

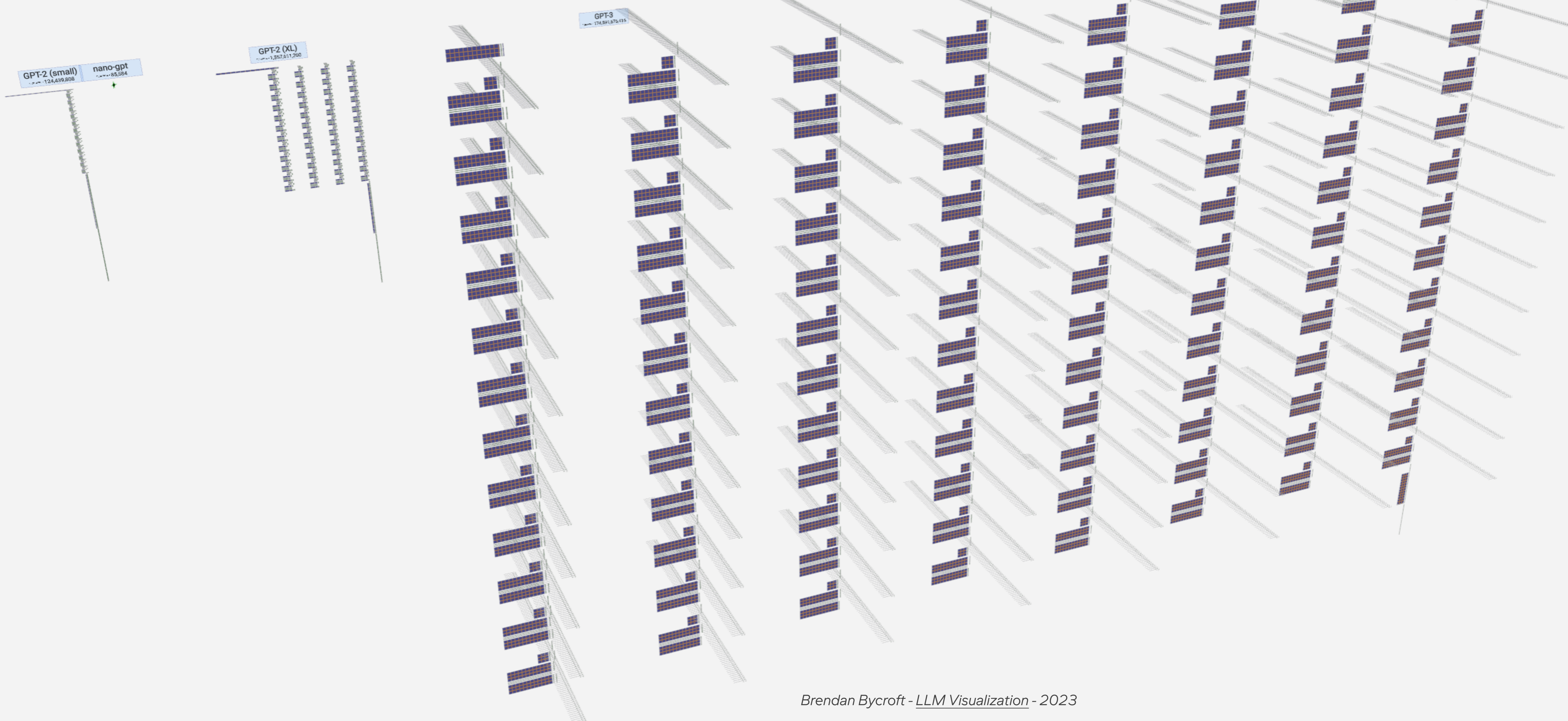
2020

Article cité plus de 65 000 fois ...

GPT2 small
124 millions de paramètres

GPT2 XL
1,5 milliards de paramètres

GPT3
174 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 sets, 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
nikiip@google.com

Jakob Uszkoreit*
Google Research
usz@google.com

Llion Jones*
Google Research
llion@google.com

Aidan N. Gomez*¹
University of Toronto
aidan@cs.toronto.edu

Lukasz Kaiser*
Google Brain
lukaszkaiser@google.com

Illa Polosukhin*¹
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.

¹Work performed while at Google Brain.

²Work 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
aradford@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 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

Article cité plus de 65 000 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



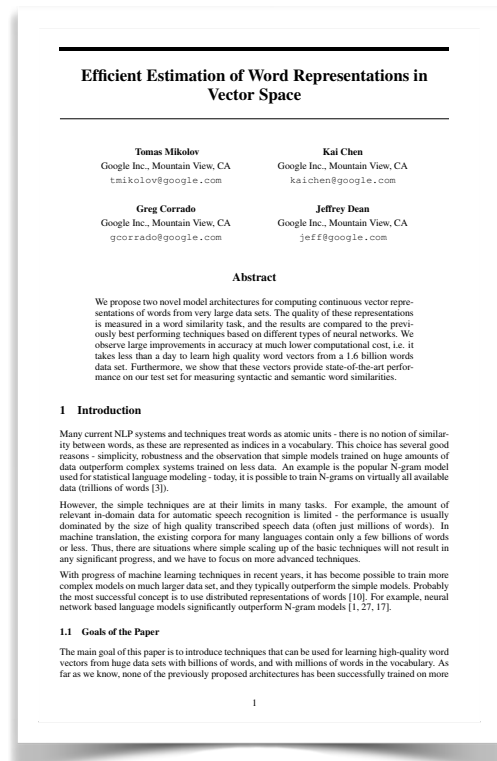
Embedding

Transformers

GPT

Bert

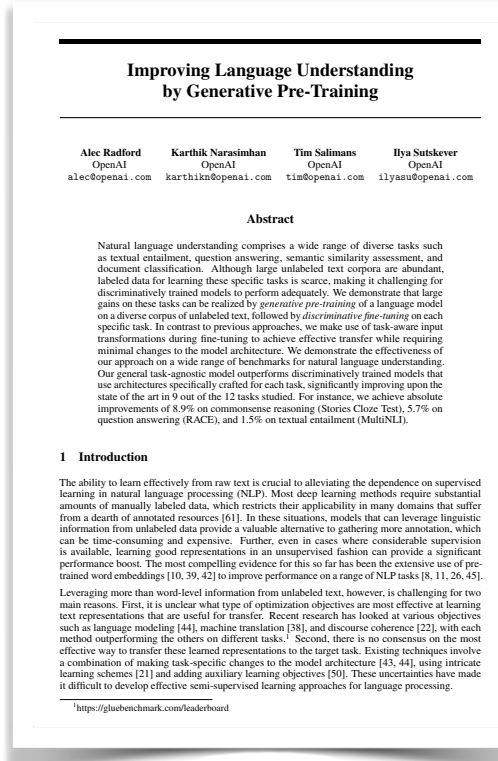
GPT 3



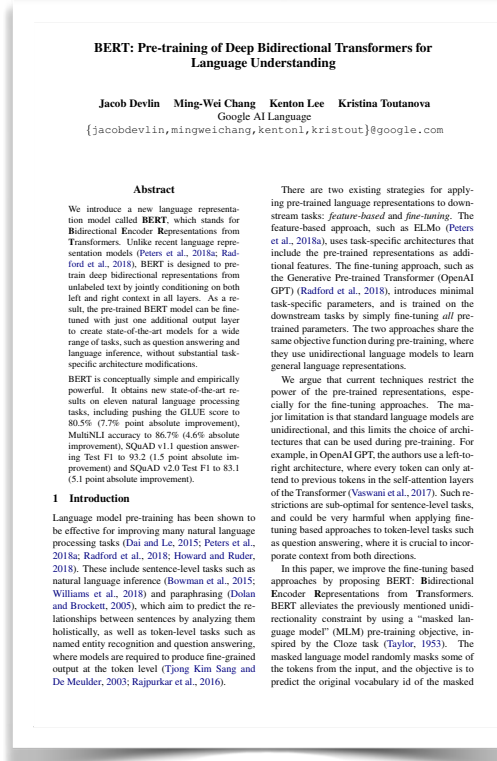
2013



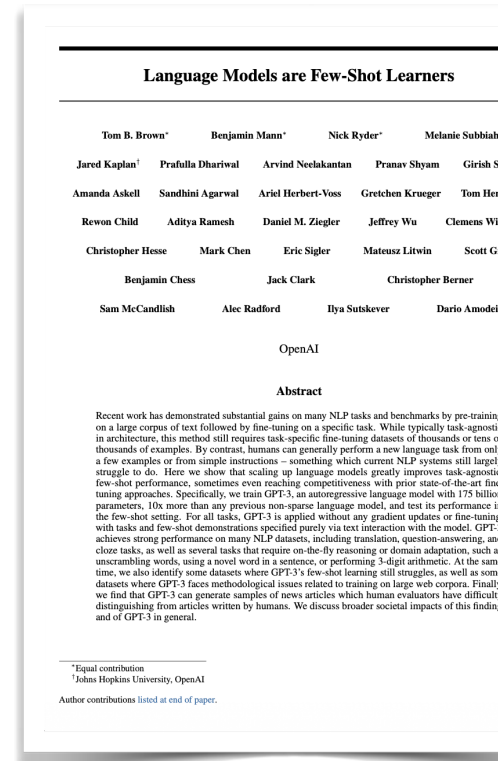
2017



2018



2018



2020

Article cité plus de 20 000 fois ...

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

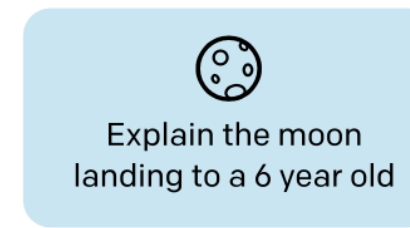
*Primary authors. This was a joint project of the OpenAI Alignment team. RL and JL are the team leads. Corresponding author: lowe@openai.com.

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

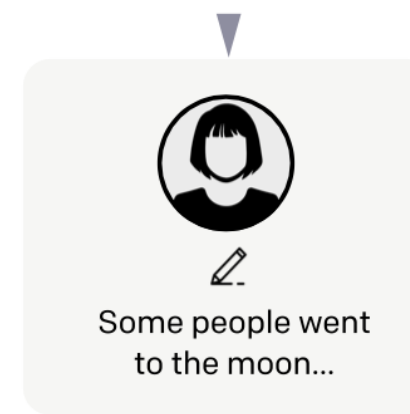
Step 1

Collect demonstration data, and train a supervised policy.

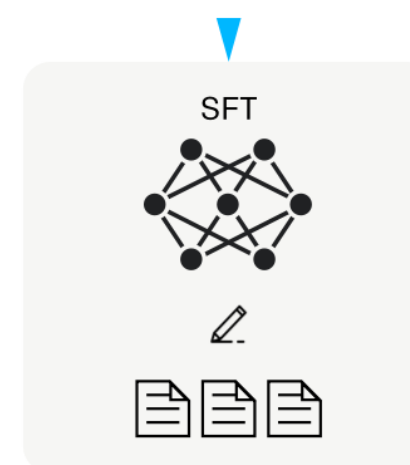
A prompt is sampled from our prompt dataset.



A labeler demonstrates the desired output behavior.



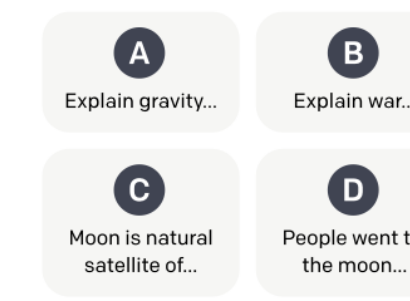
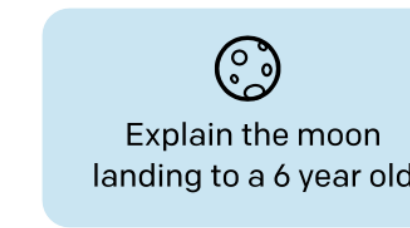
This data is used to fine-tune GPT-3 with supervised learning.



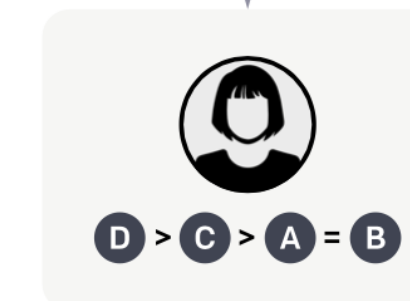
Step 2

Collect comparison data, and train a reward model.

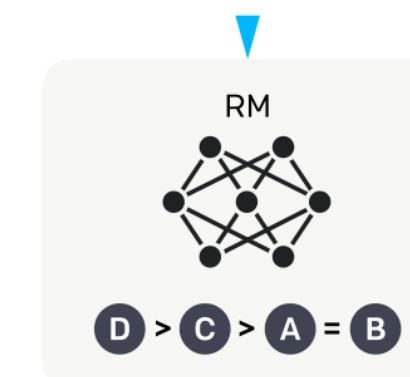
A prompt and several model outputs are sampled.



A labeler ranks the outputs from best to worst.



This data is used to train our reward model.



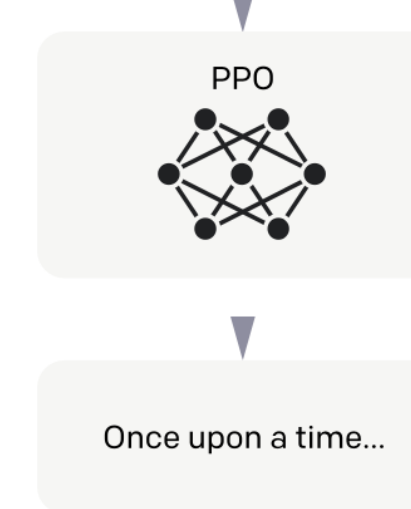
Step 3

Optimize a policy against the reward model using reinforcement learning.

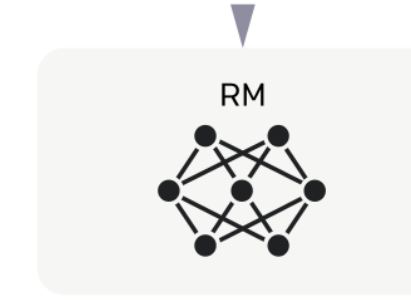
A new prompt is sampled from the dataset.



The policy generates an output.



The reward model calculates a reward for the output.



The reward is used to update the policy using PPO.



RLAIF

Reinforcement learning from AI feedback

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

Constitutional AI



Embedding

Transformers

GPT

Bert

GPT 3

RLHF



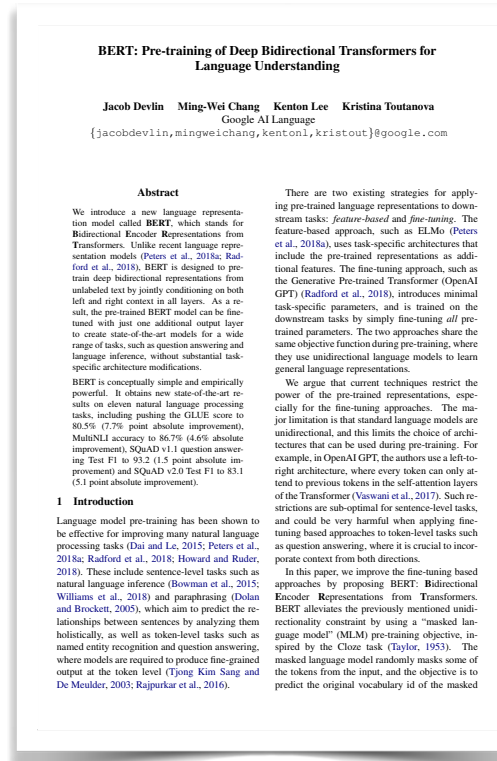
2013



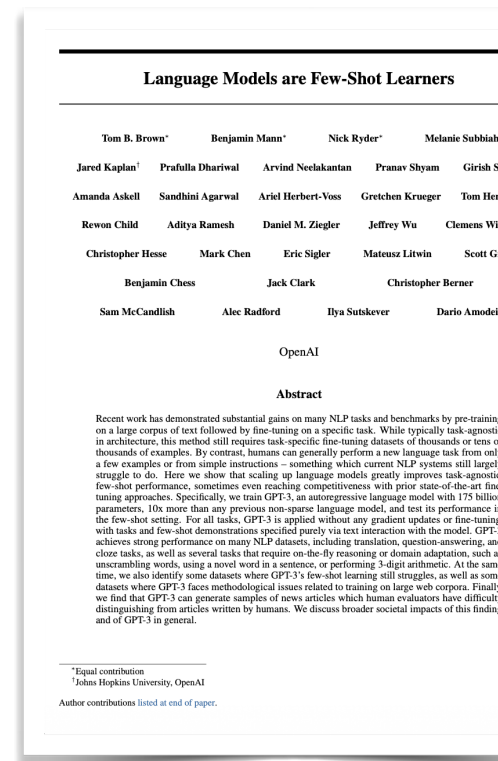
2017



2018



2018



2020



2022

Embedding

Efficient Estimation of Word Representations in Vector Space

Tom Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, Jeff Dean

Abstract

We propose an unsupervised neural network approach to estimating word representations in vector space. This approach is based on the previous work of Mikolov et al. (2013) and is significantly simpler and more efficient than the previous state-of-the-art methods. We show that these vectors provide state-of-the-art performance on a set of word analogy tasks.

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 thesis has several practical reasons: simplicity, efficiency and the observation that simple models trained on large 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 (Liu and El Ghayour, 2015).

However, the simple techniques are at their limits in many tasks. For example, the amount of relevant 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 amount of relevant data for many languages contains only a few billions of words at best. Thus, there are many situations where simple scaling up of the basic techniques will not work any longer (Bordes et al., 2014) and we have to look for more advanced techniques.

With progress of machine learning techniques in recent years, it has become possible to train more complex models on much larger domain data sets. Many proposed neural network models, particularly the most successful attempt to word distributed representations of words (Mikolov et al., 2013), are simple neural network based language models specifically adapted to N-gram models (Liu, 2015).

1.1 Goals of this Paper

The main goal of this paper is to introduce techniques that can be used for learning high-quality word vectors from large 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, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin

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 an alternative architecture: the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show that models trained by the proposed attention-based architecture can produce output of quality comparable with those of sequence-to-sequence models trained by the state-of-the-art neural machine translation models. On the WMT 2014 English-to-French translation task, our model achieves a new single-model state-of-the-art BLEU score of 41.1 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.

1 Introduction

The ability to learn effectively from raw text is crucial to alleviating the dependence on supervised learning in natural language processing (NLP). Many deep learning methods require substantial amounts of manually labeled data, which limits their applicability in many domains that suffer from a dearth of annotated resources [1]. In these situations, models that can leverage implicit information from unlabeled data provide a valuable alternative to gathering more annotations, 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, 19, 42] to improve performance on a range of NLP tasks [8, 11, 20, 45].

Levitating 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 such representations that are useful for transfer. Recent research has looked at various objectives such as language modeling [10], word translation [13], and document classification [22], with each method outperforming the others in 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 manual task-specific changes to the model architecture [11, 44], using additional learning objectives [21] and adding auxiliary learning objectives [50]. These techniques have made it difficult to develop effective cross-task representation learning approaches for language processing.

¹Equal contribution. Listing order is random. Jakob proposed replacing RNNs with self-attention and started an effort to evaluate this idea, which, with this, designed and implemented the 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 representations and the paper was finalized in nearly every detail. Niki designed, implemented, tested and evaluated countless model variants in our original codebase and implementation. Llion also experimented with model variants, was responsible for our initial codebase, and efficient inference and visualization. Łukasz and Aidan spent countless long days debugging various parts of and implementing our codebase, replacing our entire codebase, greatly improving results and facilitating accelerated development.

²Work performed while at Google Brain.

³Work 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, Karthik Narasimhan, Tim Salimans, Ilya Sutskever

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 discriminatorily 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-agnostic 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 discriminatorily 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 (SenseEval), 3.7% on question answering (QA), and 1.9% on textual entailment (MCTest).

1 Introduction

The ability to learn effectively from raw text is crucial to alleviating the dependence on supervised learning in natural language processing (NLP). Many deep learning methods require substantial amounts of manually labeled data, which limits their applicability in many domains that suffer from a dearth of annotated resources [1]. In these situations, models that can leverage implicit information from unlabeled data provide a valuable alternative to gathering more annotations, 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, 19, 42] to improve performance on a range of NLP tasks [8, 11, 20, 45].

Levitating 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 such representations that are useful for transfer. Recent research has looked at various objectives such as language modeling [10], word translation [13], and document classification [22], with each method outperforming the others in 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 manual task-specific changes to the model architecture [11, 44], using additional learning objectives [21] and adding auxiliary learning objectives [50]. These techniques have made it difficult to develop effective cross-task representation learning approaches for language processing.

¹Equal contribution. Listing order is random.

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

2018

Bert

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova

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., 2018; Radford et al., 2018), BERT is designed by pre-training deep bidirectional representations from unlabeled text by jointly modeling left-to-right and right-to-left languages with all-pairs self-attention in all layers. As a result, the pre-trained BERT model can be finetuned with just one additional output layer to create state-of-the-art models for a wide range of tasks, such as question answering, named entity recognition, text classification, and other standard language understanding tasks. We report that current techniques (noted the power of the pre-trained representations, especially for the fine-tuning approaches). The main 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 models 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 propose the fine-tuning based approach by proposing BERT, Bidirectional Encoder Representations from Transformers. BERT addresses the previously mentioned unidirectionality constraint by using a "masked language model" (MLM) pre-training objective, inspired by the Cloze task (Tasbe, 1953). The masked language model randomly masks some of the tokens from the input, and the objective is to predict the original vocabulary of all 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, Rerand Chen, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Chenxia Wang, Christopher Jones, Max Chen, Eke Szigler, Matthew DeLora, Scott Gray, Benjamin Chess, Mark Chen, Eric Sigler, Mantas Llamza, Greg Brockman, Gwanwan Han, Chris Jones, Alex Acero, Jan Leike, Junxue Guo, Jerry Twigg, Raymond Wu, Fei-Fei Li, David Dohan, Pranav Shyam, Greg Brockman

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 wide range of tasks 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 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. This is more than any previous non-sparse language model, and not 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 unanswerable questions, using a novel word for a sentence, or performing 3-digit arithmetic. At the same time, we also identify some datasets where GPT-3 has still learning still struggles, as well as some datasets where GPT-3 faces methodological issues related to training on large text. Finally, we find that GPT-3 can generate samples of text that are indistinguishable from human writing, demonstrating that large-scale language models can generate text that is highly similar to human writing. We discuss broader societal impacts of this finding and of GPT-3 in general.

¹Equal contribution.

²Work done while at 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 Su, Janine Taylor, Michael Saikh, Nelson El-Farouh, Joshua Reizenstein, Saizheng Zhang, Ryan Jiang, Michele Jenia, Hiroaki Kato, Austin Reiter, Edun S. Kim, David Schuff, Shrivardhan V. Adiga, M. S. Kumar Abhinav, Dhruv Choudhary, Anja Hauth, Adam Fisch, Jihan Yang, Andre N. Karpathy, Jan Leike, Alex Gray, and Jan Leike

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 unaligned, 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 over a wide range of tasks by training with human feedback. Starting with a set of domain-specific prompts and prompts submitted through the OpenAI API, we collect a dataset of labeled 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 10% fewer parameters. Moreover, InstructGPT models show improvements in helpfulness 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 (LLMs) can be "prompted" to perform a range of natural language processing (NLP) tasks, given some examples of the task in 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; Brundage et al., 2021; Krner et al., 2021; Washadek et al., 2021; Touvron et al., 2021; Gehrmann et al., 2020). This is because the language modeling objective

¹Primary author. This was a joint project of the OpenAI Alignment team, RL, and us, the core team.

²Work done while at OpenAI.

Corresponding author: longo@openai.com

Work done while at OpenAI.

Corresponding author: longo@openai.com

Work done while at OpenAI.

Corresponding author: longo@openai.com

2022

RLAIF

Constitutional AI: Harmlessness from AI Feedback

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Andra Mirowski, Cameron McKinnon, Conrad Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Turner, Ethan Perez, Jesse Krutz, Josiah Davis, Jeffrey Luhik, Joshua Landau, Karan Nohria, Kamal Lakshminarayanan, Liane Lovitt, Michael Schmitt, Nelson El-Farouh, Nicholas Schiefer, Noam Mazirow, Nova Daswara, Robert Lasnik, Rohin Kari, Sam Wang, Scott Johnston, Shantanu Kravets, Jeffrey Larson, Janina Nicholson, Nicholas Schiefer, Noam Mazirow, Nova Daswara, Robert Lasnik, Rohin Kari, Sam Wang, Scott Johnston, Shantanu Kravets, Shree EP Shrivastava, Stanislas Hart, Tanner Lianoglou, Timothy Telleen Lavian, Tim Conroy, Tom Henighan, Tristan Hunt, Sumit G. Datta, Zac Hatfield Dodds, Ben Mann, Daria Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, Jared Kaplan

Abstract

As AI systems become more capable, we would like to enable their help to supervise other AI. 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 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 prompt self-critique and revision, and then finetune the original model on revised outputs. In the RL phase, we sample from the finetuned model, use a model to estimate quality of the two samples a better, and then train a preference model from this dataset of AI preferences. We then fine-tune with RL, using the preference model for the reward signal. In our use RL from AI Feedback (RLAIF), as a result we are able to train a harmless but non-toxic AI assistant that engages with harmful queries by explaining its objections to them. Both the RL 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.

¹Correspondence to: {yuntao, jesse}@openai.com

²Author contributions are detailed in [1]

2022

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...)