Cette conférence est suspecte, vous ne savez pas si c'est un être humain qui en est l'auteur...

CONFÉRENCE 1 / 2

De Turing à ChatGPT : l'IA expliquée

Un voyage au cœur de l'Intelligence Artificielle et les modèles de langage — leur histoire, leur architecture et ce qu'ils peuvent vraiment faire. Première étape d'une série en deux volets.

Deux types d'IA

Les IA spécialisées

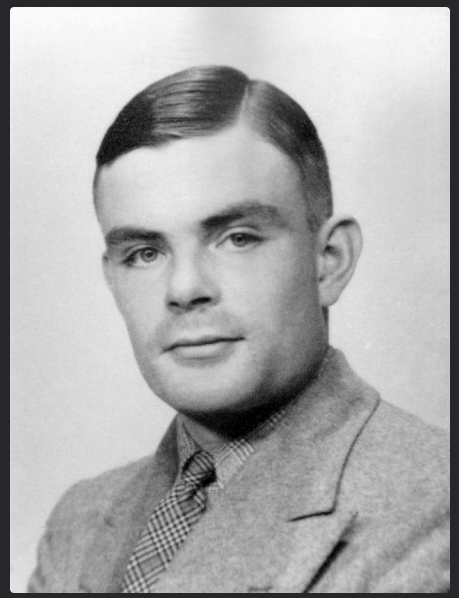
- Jeux d'échecs
- Jeux vidéos
- Reconnaissance d'images

Les IA génératives

- Les modèles de langage dont on entend beaucoup parler en ce moment : GPT etc

Quand la machine défie l'humain

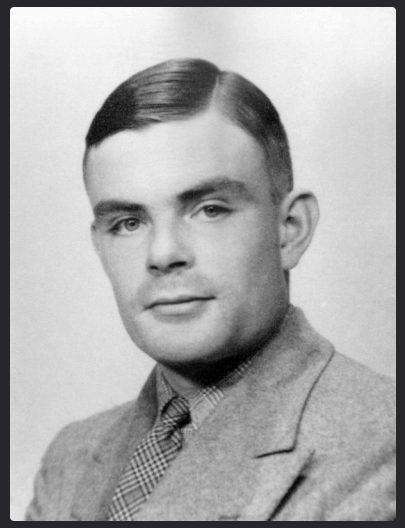
Alan Turing



Auteur des fondements scientifiques de l'informatique, pionnier de l'intelligence artificielle. Joue un rôle clé dans la seconde guerre mondiale.

Quand la machine défie l'humain

Alan Turing



Auteur des fondements scientifiques de l'informatique, pionnier de l'intelligence artificielle. Joue un rôle clé dans la seconde guerre mondiale.

Le Test de Turing - 1950

Une machine est-elle intelligente si on ne peut la distinguer d'un humain ?

Si un humain échange à l'aveugle avec un autre humain ou un ordinateur et ne sait pas distinguer les deux, la machine passe le test et peut être dite "intelligente".

Prédiction : en l'an 2000 des machines de 128 Mo de mémoire pourront tromper des juges pendant 5 minutes

Le film Her - 2013

Quand la machine défie l'humain

Échecs — 1997

Deep Blue bat Kasparov. Espace fini, calcul brut. La machine explore des millions de positions par seconde.

Une IA "brutale" - Victoire du hardware

"Juste" une machine capable de calculer toutes les combinaisons de coups possibles et inventer des coups contre-intuitifs mais qui apportent un gain.

Quand la machine défie l'humain

Go — 2015

AlphaGo bat Lee Sedol. Plus de combinaisons que d'atomes dans l'univers. L'intuition devient nécessaire.

Alpha Go (Deepmind)

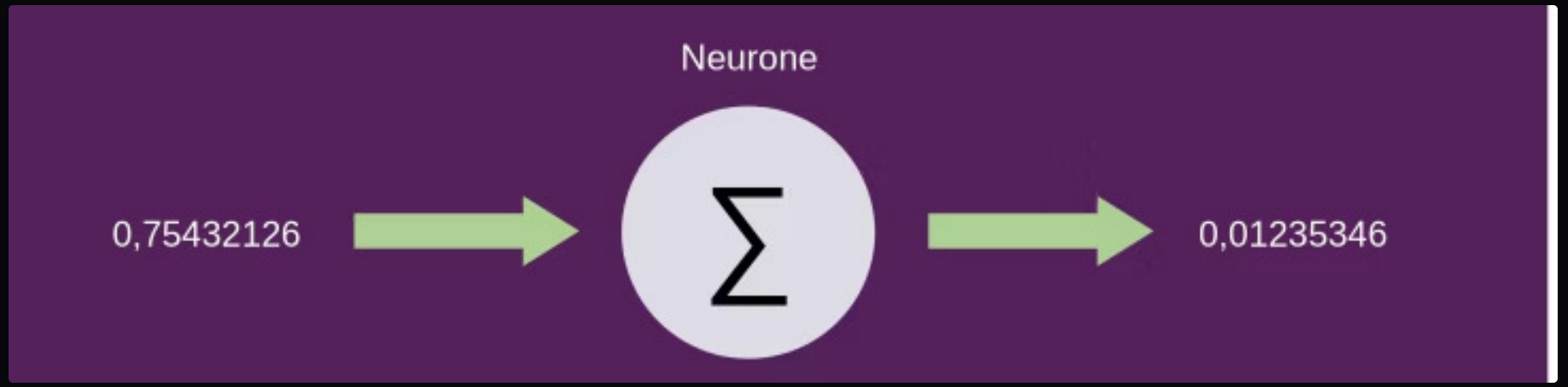
Apprentissage - Victoire du software

Alpha Go est entraîné sur des centaines de milliers de parties entre humains. Mais il joue aussi contre lui-même autant de fois que nécessaire et "apprend" les stratégies à adopter pour s'améliorer.

Programmation vs Apprentissage

Les réseaux de neurones

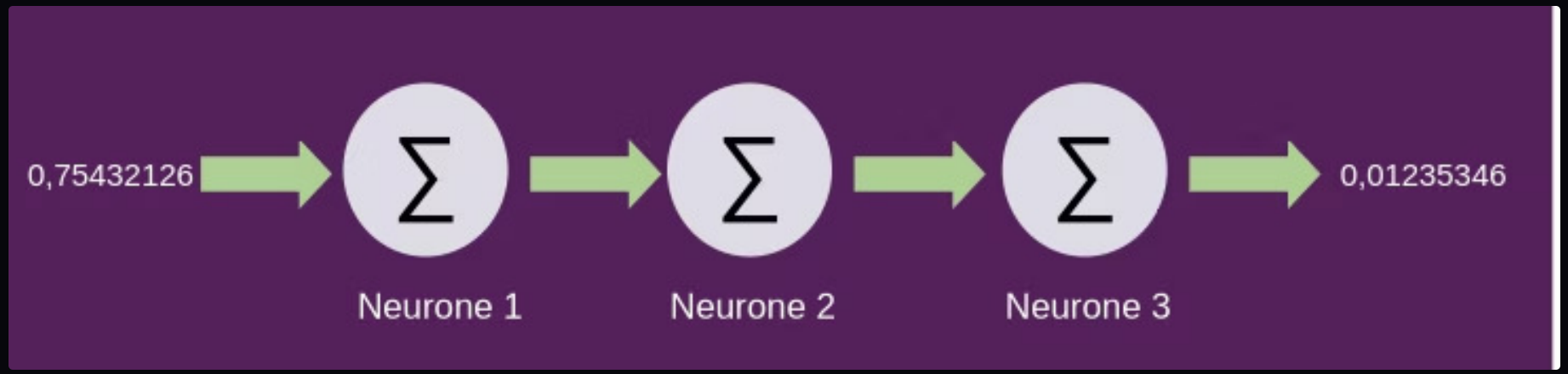
Inspirés très schématiquement du neurone biologique



Les réseaux de neurones

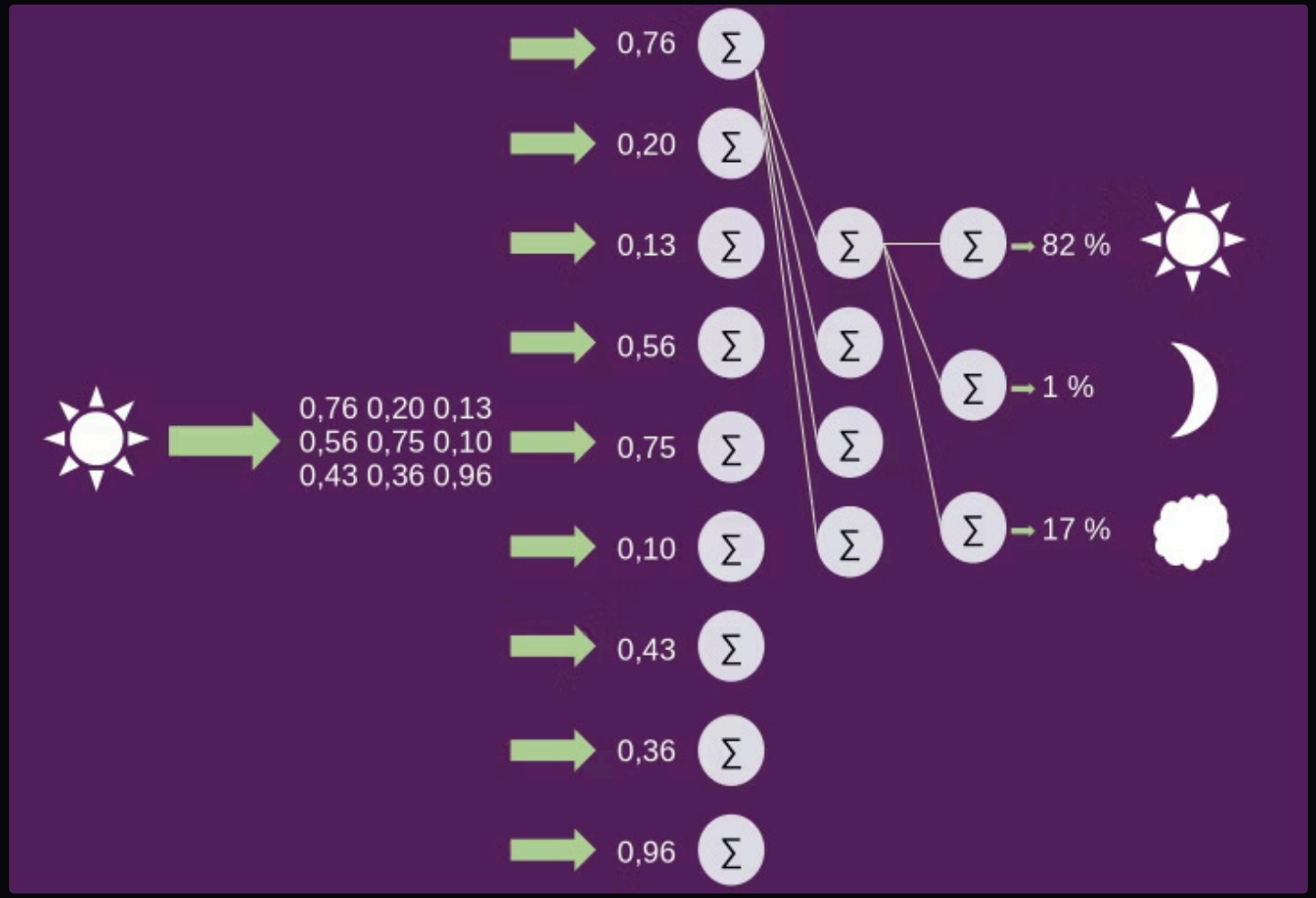
Inspirés très schématiquement du neurone biologique

Couche d'entrée - Couche cachée - Couche de sortie

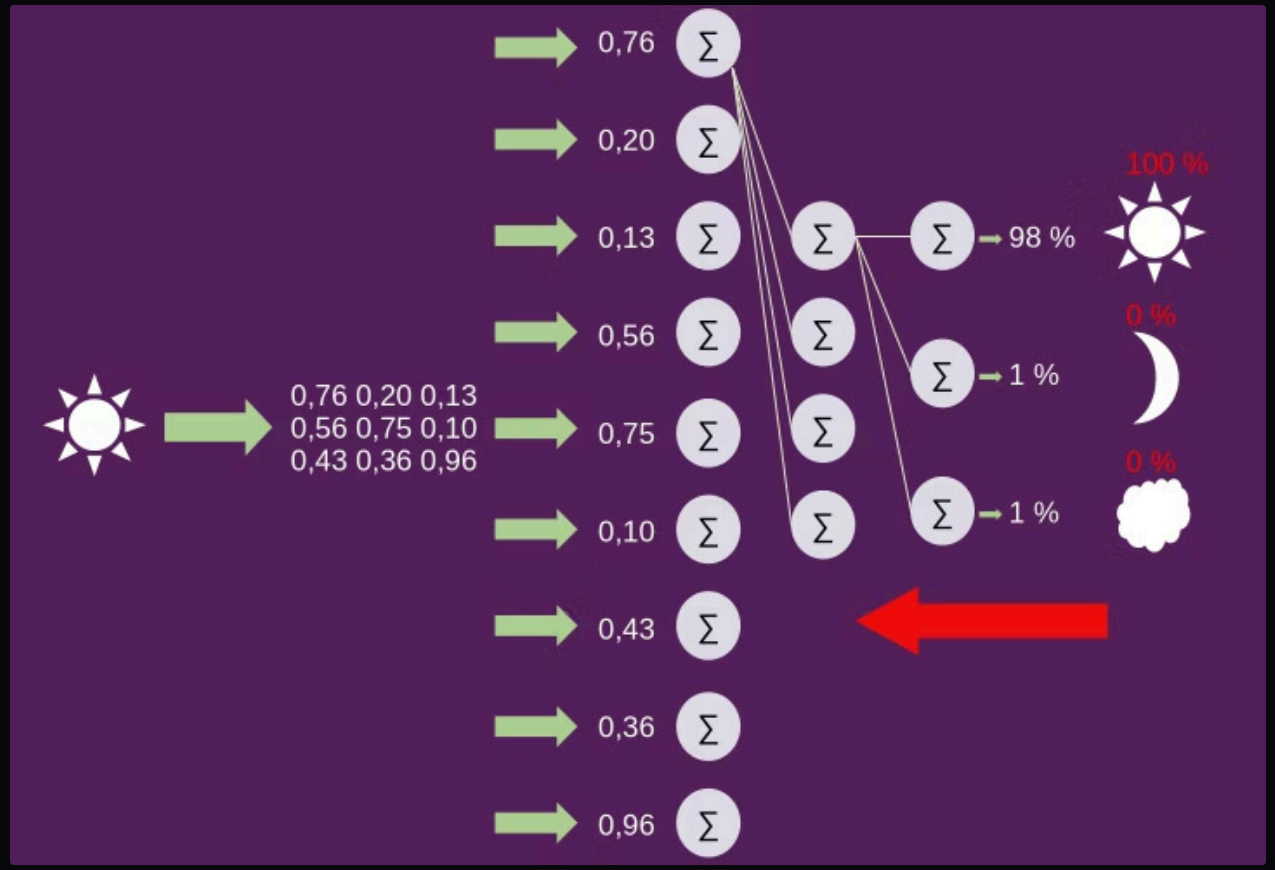


Les réseaux de neurones

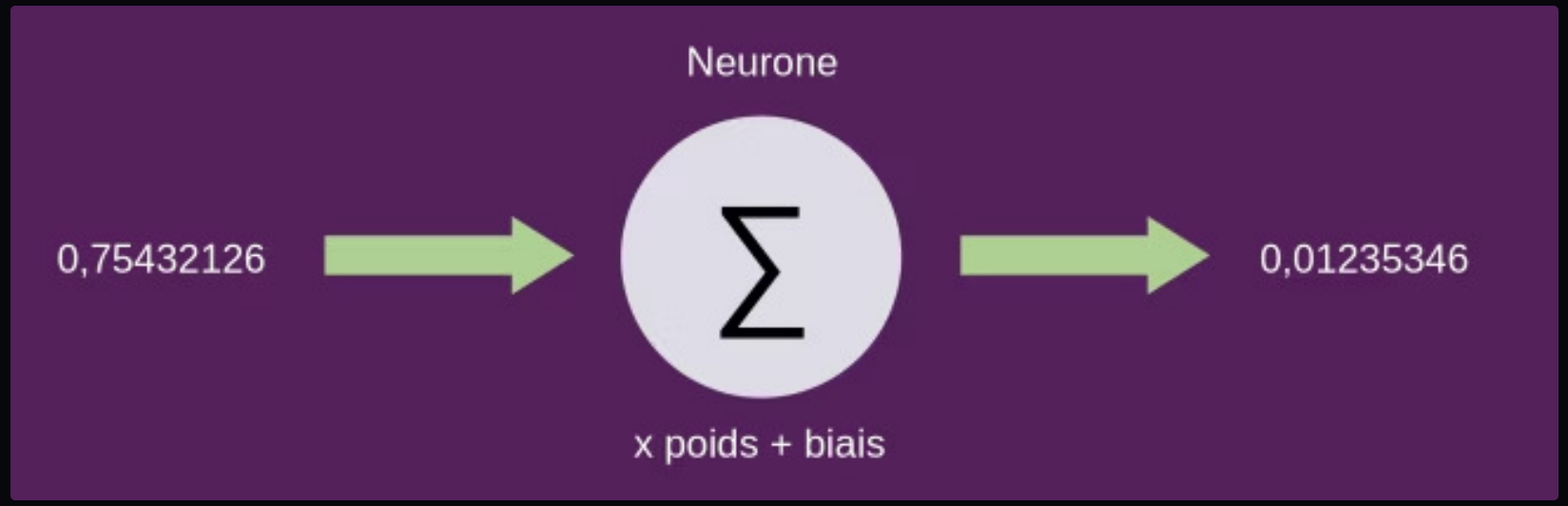
Les réseaux de neurones



L'apprentissage : rétropropagation de l'erreur et réajustement des paramètres

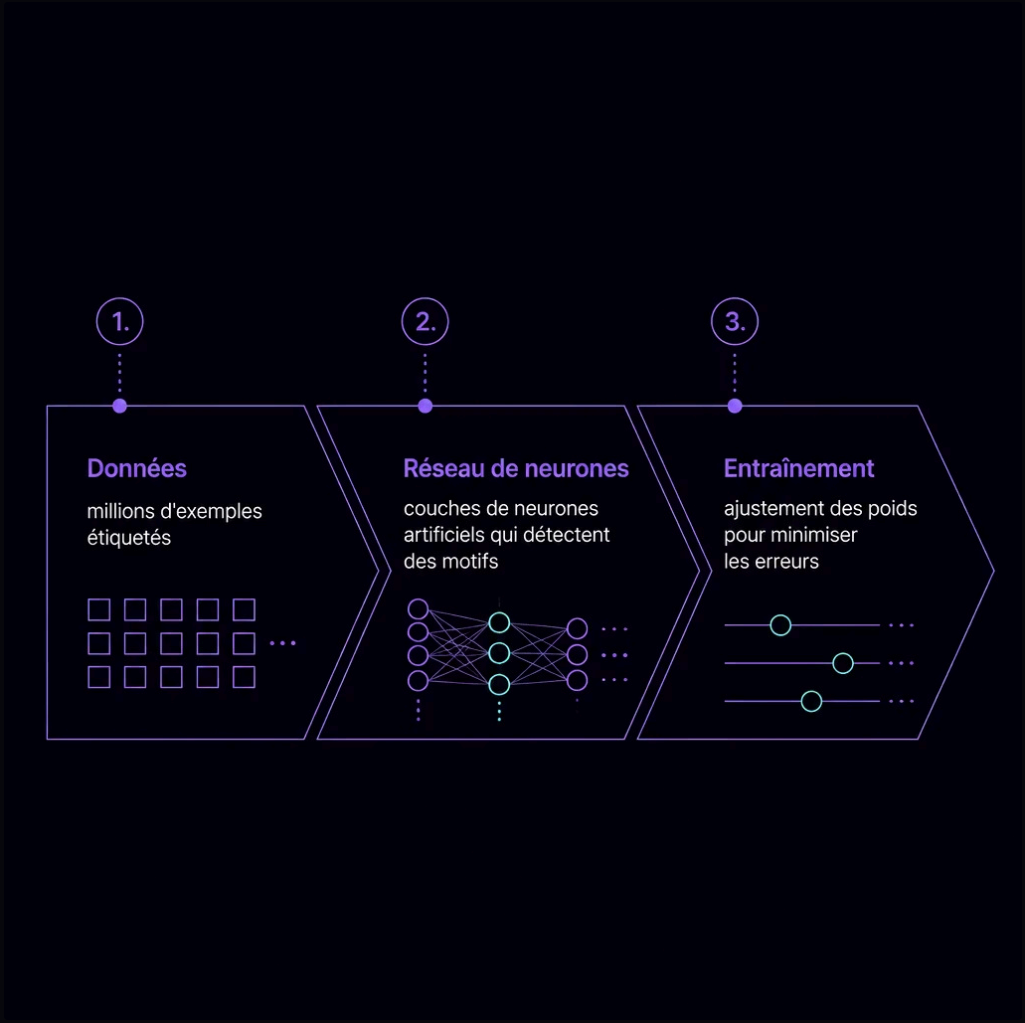


Nombre de paramètres



Dans l'exemple précédent : 9×4 = 36 poids + 9×4 = 36 biais. Il y a 72 paramètres dans le modèle, c'est autant de curseurs pour l'affiner. Les modèles d'IA actuels en ont plusieurs milliers de milliards.

L'apprentissage automatique et les réseaux de neurones



Comment ça apprend ?

Un réseau de neurones est inspiré du cerveau humain : des **neurones artificiels** organisés en couches détectent des motifs dans les données.

L'**entraînement** ajuste des milliards de paramètres pour minimiser les erreurs — par essais, erreurs et corrections répétées.

> Plus de données = plus de capacités. C'est la clé de la révolution récente.

Apprentissage supervisé

Entrée → **Modèle** → Prédiction → Données étiquetées (valeurs réelles) → Calcul de l'erreur → Ajustement

Application : classification d'images

Cas d'AlphaGo : apprentissage auto-supervisé

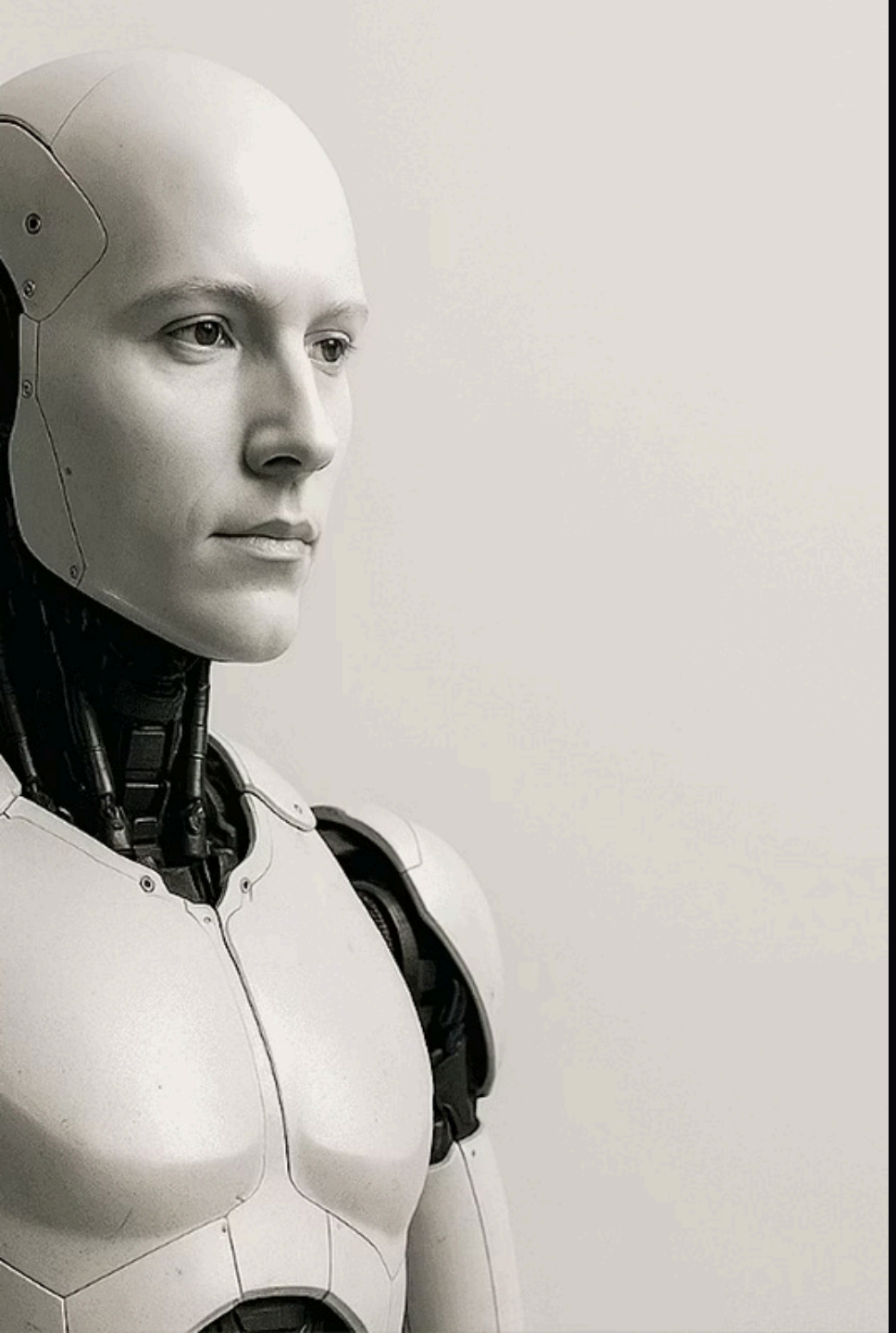
Apprentissage non supervisé

Entrée → Modèle → Trouve des structures

Applications : détecter des maladies, trouver des failles de sécurité

2018 - Sam Altman - OpenAI

Le chat mange la …

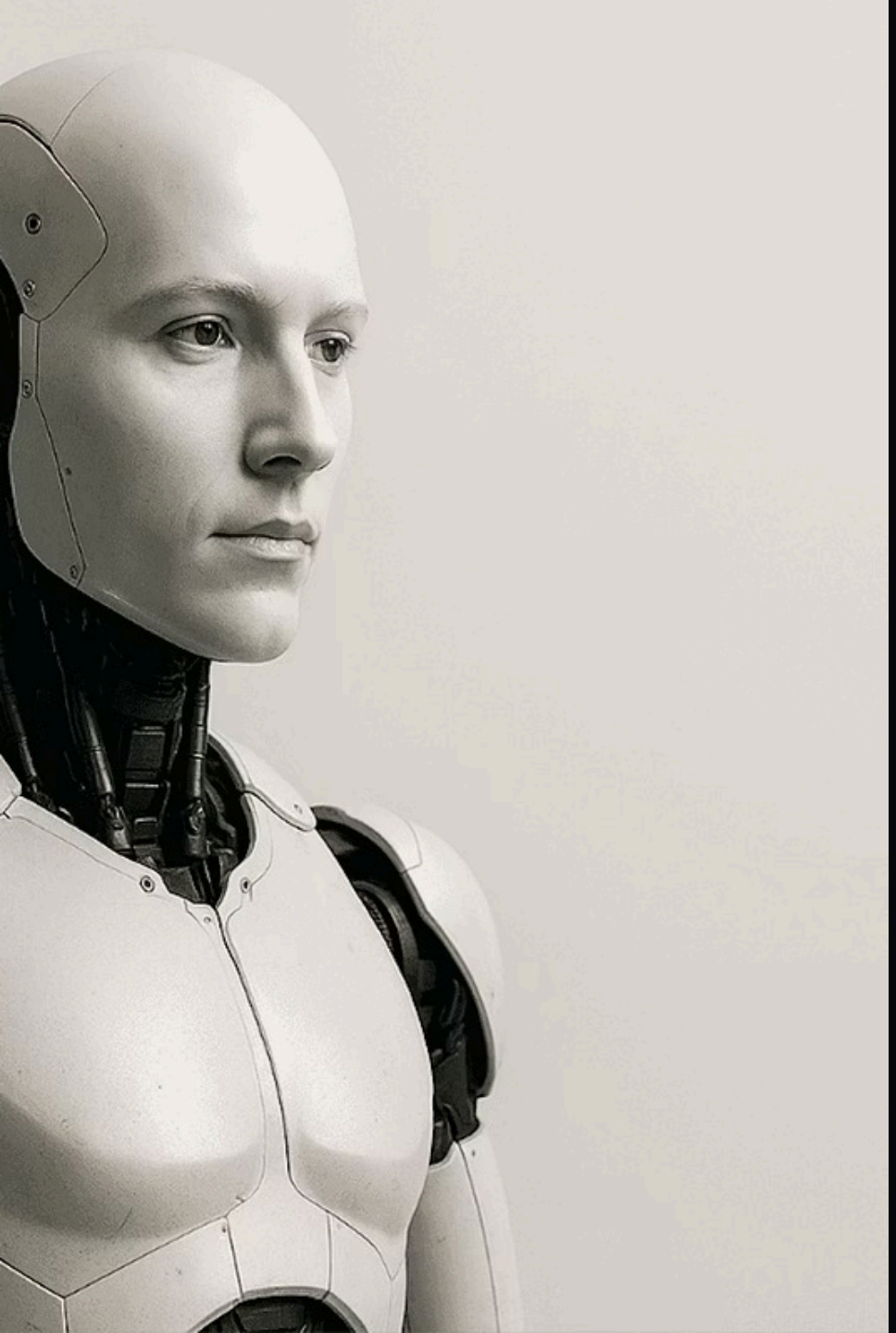


Des modèles de complétion de texte...

Au départ, une idée simple : **prédire le mot suivant** dans une séquence. En apprenant sur des milliards de textes, le modèle capture la grammaire, le style, la logique — et bien plus encore.

Prédiction statistique

Le modèle calcule la probabilité de chaque mot possible.



Des modèles de complétion de texte...

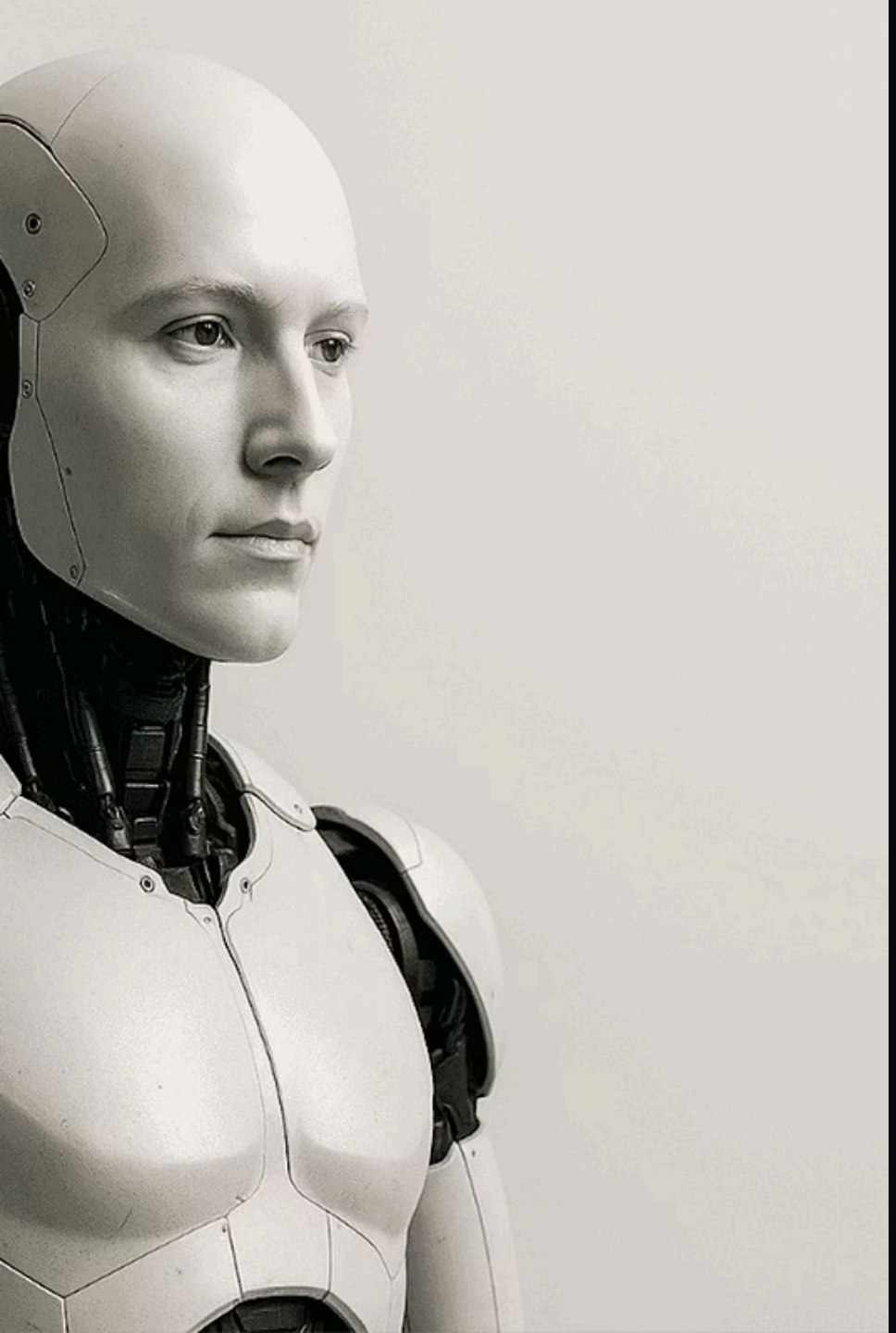
Au départ, une idée simple : **prédire le mot suivant** dans une séquence. En apprenant sur des milliards de textes, le modèle capture la grammaire, le style, la logique — et bien plus encore.

Prédiction statistique

Le modèle calcule la probabilité de chaque mot possible.

Émergence inattendue

À grande échelle, des capacités de raisonnement apparaissent sans avoir été explicitement enseignées.



Des modèles de complétion de texte...

Au départ, une idée simple : **prédire le mot suivant** dans une séquence. En apprenant sur des milliards de textes, le modèle capture la grammaire, le style, la logique — et bien plus encore.

Prédiction statistique

Le modèle calcule la probabilité de chaque mot possible.

Émergence inattendue

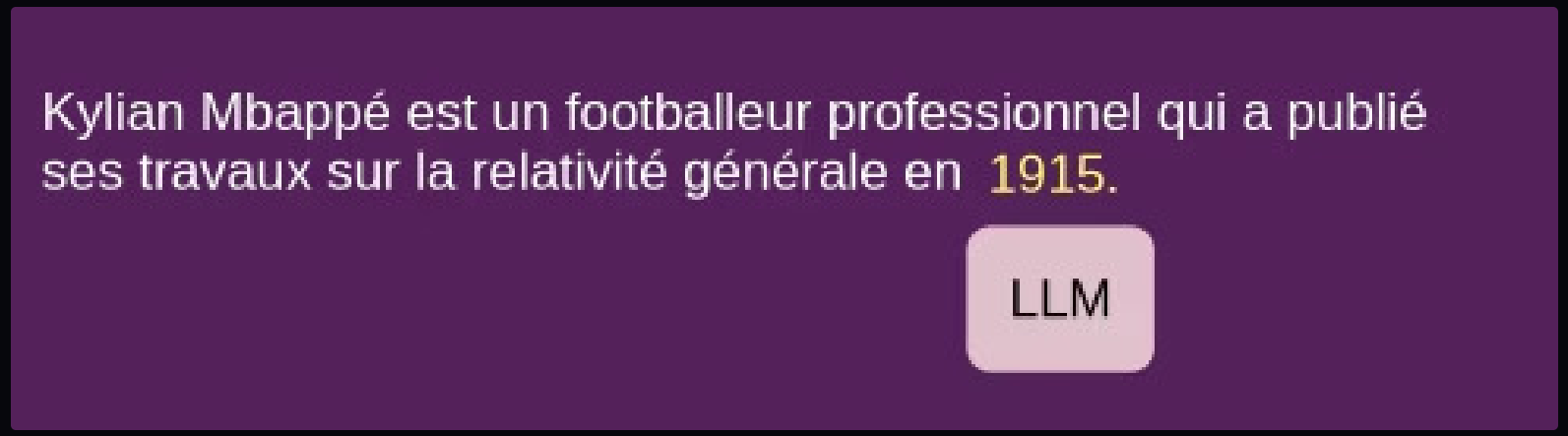
À grande échelle, des capacités de raisonnement apparaissent sans avoir été explicitement enseignées.

La langue comme interface

Le texte devient le langage universel pour interagir avec la machine.

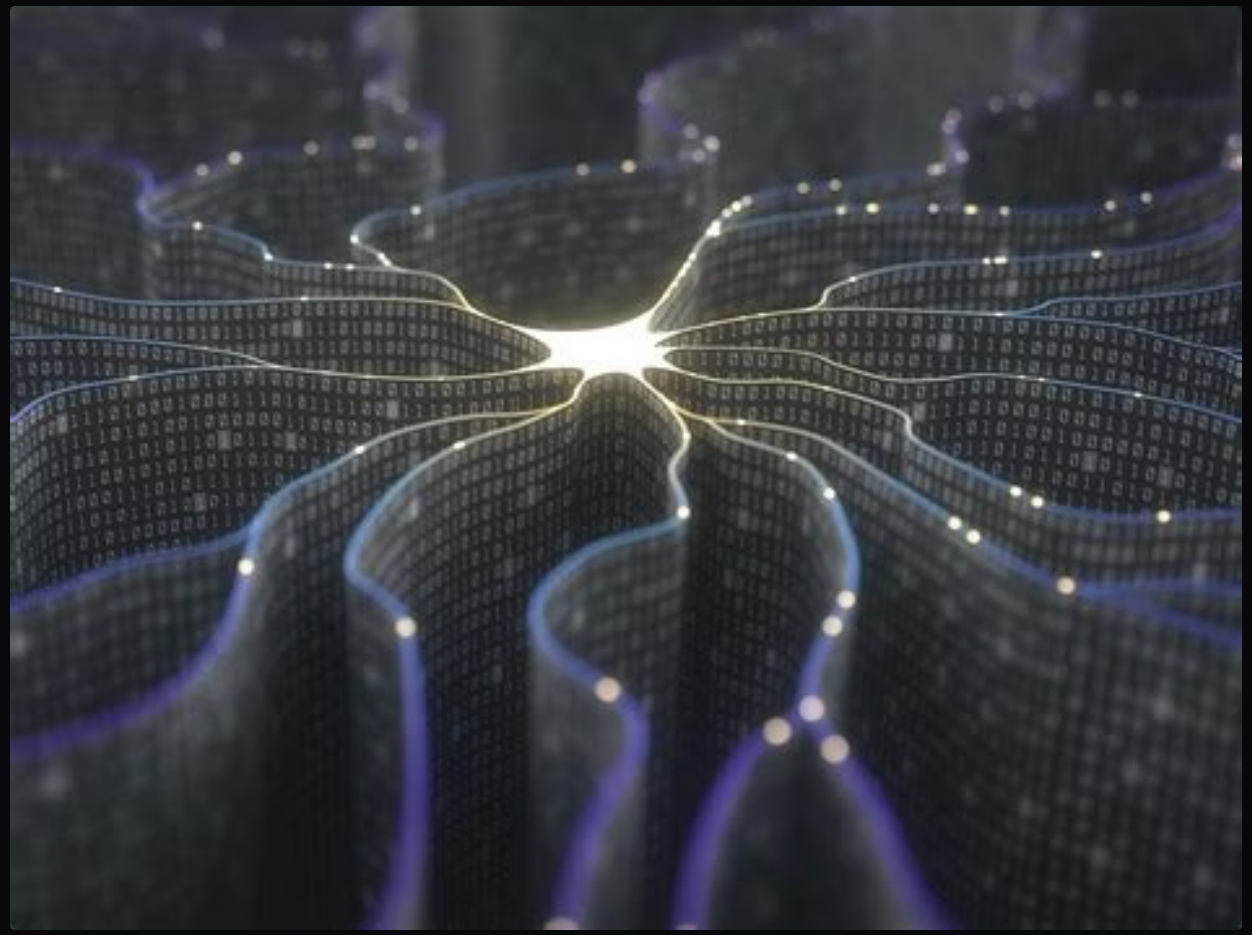
Large Language Models - LLM

Grands modèles de langage



Modèle pré-entraîné

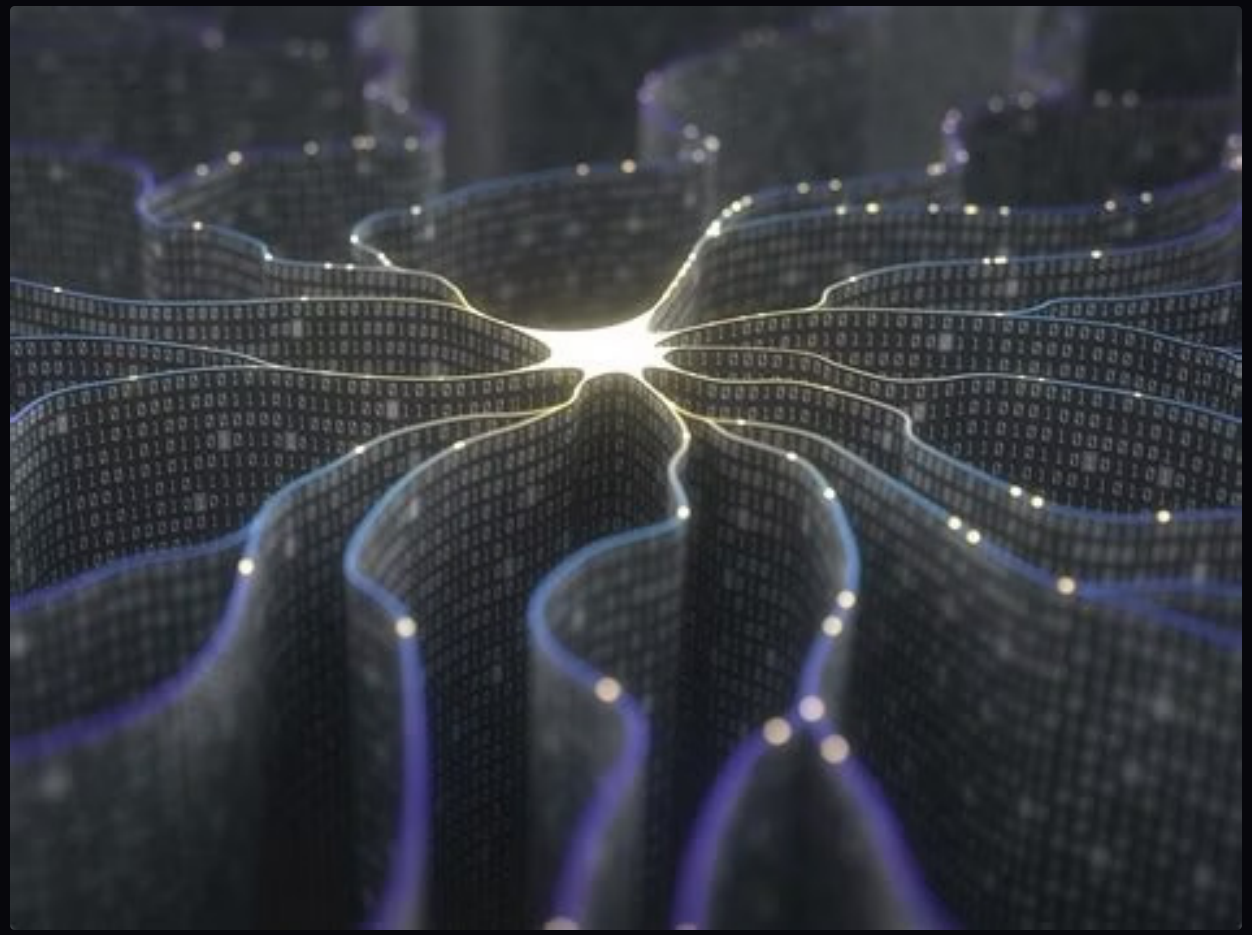
Le Transformer : l'architecture qui a tout changé



Introduit en 2017 par Google, le **Transformer** utilise un mécanisme d'**attention** : il pondère l'importance de chaque mot du contexte pour générer le suivant, peu importe l'éloignement.

- **Pré-entraîné** sur des corpus massifs (internet, livres, code)
- **Génératif** : il produit du texte, pas seulement des classifications
- **Mis à l'échelle** : plus de paramètres = plus de capacités

Le Transformer : l'architecture qui a tout changé



GPT = Generative Pre-trained Transformer

Génératif : il est capable de générer du texte et des suites de mots avec une certaine logique

Pré-entraîné : sur un grand corpus, mais il aura besoin de plus d'entraînement pour être efficace

Transformer : il comprend le contexte global d'un texte à compléter

Les LLMs modernes

Étape 1

Pré-entraînement

Auto-supervisé

(généraliste)

Les LLMs modernes

Étape 1

Pré-entraînement

Auto-supervisé

(généraliste)

Étape 2

Fine-tuning

Supervisé

Annotation humaine

L'évolution de GPT et OpenAI

Fondée en 2015, depuis soutenue par Microsoft. Mission : une IA sûre et bénéfique pour l'humanité. La transparence des débuts n'existe plus.

1 — GPT-1 — 2018

117 millions de paramètres. Preuve de concept. À peine meilleur que les IA de complétion de texte.

2 — GPT-2 — 2019

3 — GPT-3 — 2020

L'évolution de GPT et OpenAI

Fondée en 2015, devenue soutenue par Microsoft. Mission : une IA sûre et bénéfique pour l'humanité. La transparence des débuts n'existe plus.

1

GPT-1 — 2018

117 millions de paramètres. Preuve de concept. À peine meilleur que les IA de complétion de texte.

2

GPT-2 — 2019

1,5 milliard de paramètres. De nouvelles capacités émergent. Il peut générer des textes logiques, donner son avis, argumenter, inventer. Fait beaucoup d'erreurs.

3

GPT-3 — 2020

L'évolution de GPT et OpenAI

Fondée en 2015, devenue soutenue par Microsoft. Mission : une IA sûre et bénéfique pour l'humanité. La transparence des débuts n'existe plus.

1

GPT-1 — 2018

117 millions de paramètres. Preuve de concept. À peine meilleur que les IA de complétion de texte.

2

GPT-2 — 2019

1,5 milliard de paramètres. De nouvelles capacités émergent. Il peut générer des textes logiques, donner son avis, argumenter, inventer. Fait beaucoup d'erreurs.

3

GPT-3 — 2020

175 milliards de paramètres. Un colosse. Entraîné sur 80% de pages Internet et 20% de sources plus raffinées (livres…). Il baratine, génère le mot attendu dans la conversation.

Le coût de l'entraînement

En combien de temps ces modèles peuvent-ils être entraînés sur une carte graphique ordinaire ?

1. **GPT-1 — 2018**
   Une dizaine de jours
2. **GPT-2 — 2019**
   Quelques années
3. **GPT-3 — 2020**
   Un millénaire

En pratique ce sont des fermes de puces qui tournent en parallèle.

GPT : pour l'instant juste de la complétion

Explique moi ce qu'est une étoile.

Je ne comprends pas bien ce que c'est, et j'aimerais beaucoup en savoir plus.

Et voici... ChatGPT

Septembre 2022

Basé sur une version modifiée de GPT-3

ChatGPT : un assistant IA basé sur un LLM

Élève : je souhaiterais connaître ce qu'est une planète.

Professeur d'astronomie : c'est un corps rocheux ou gazeux en rotation autour du Soleil.

Élève : je souhaiterais connaître ce qu'est une étoile.

Professeur d'astronomie :

Une étoile est un corps gazeux produisant de la lumière par réaction thermo-nucléaire.

Du modèle brut au chatbot : la surcouche conversationnelle

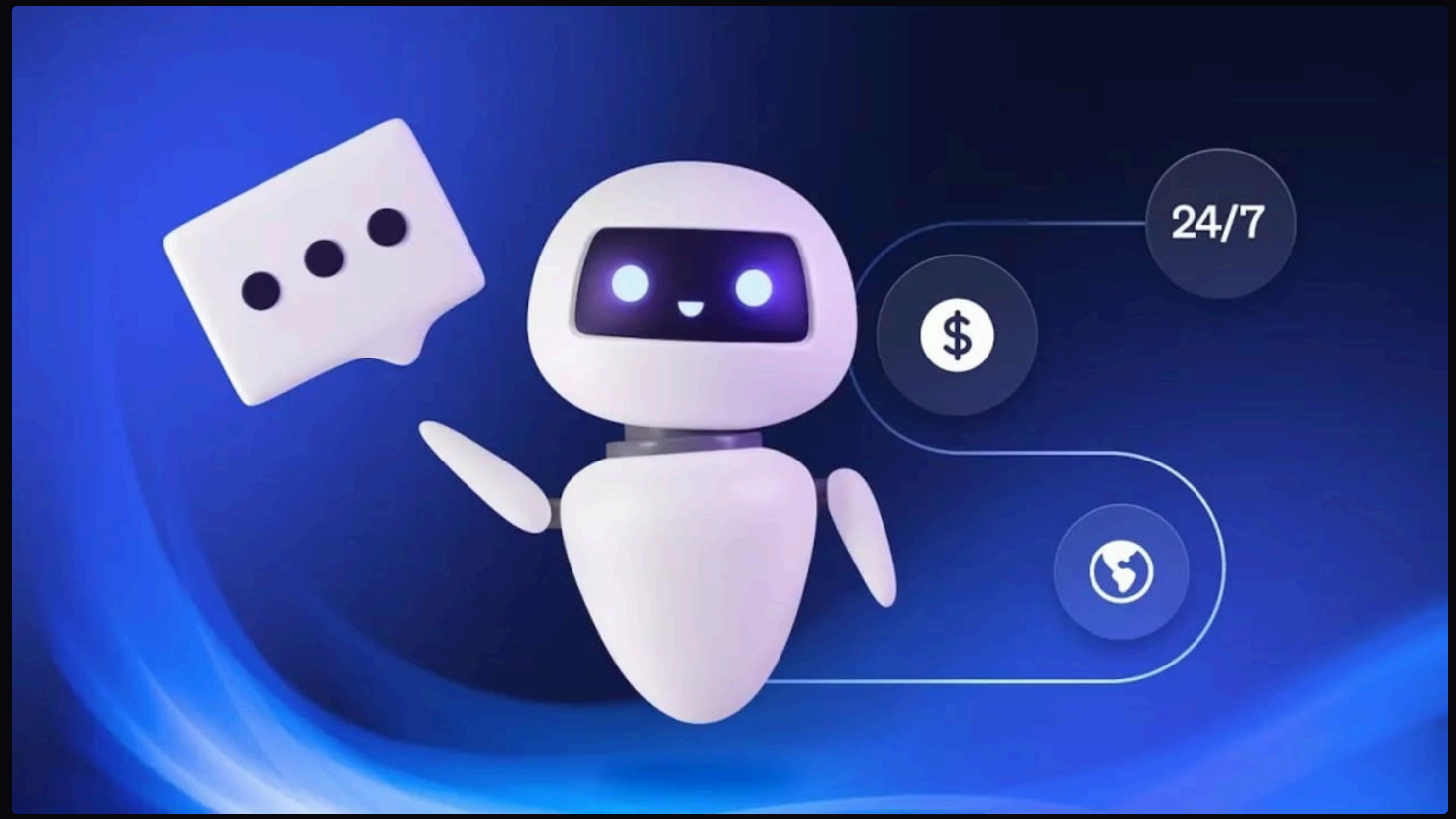
Trois couches invisibles

1 **Le System Prompt**

Instructions cachées qui définissent le comportement du modèle (« Tu es un assistant utile… »). Persona.

2

3



Du modèle brut au chatbot : la surcouche conversationnelle

Trois couches invisibles

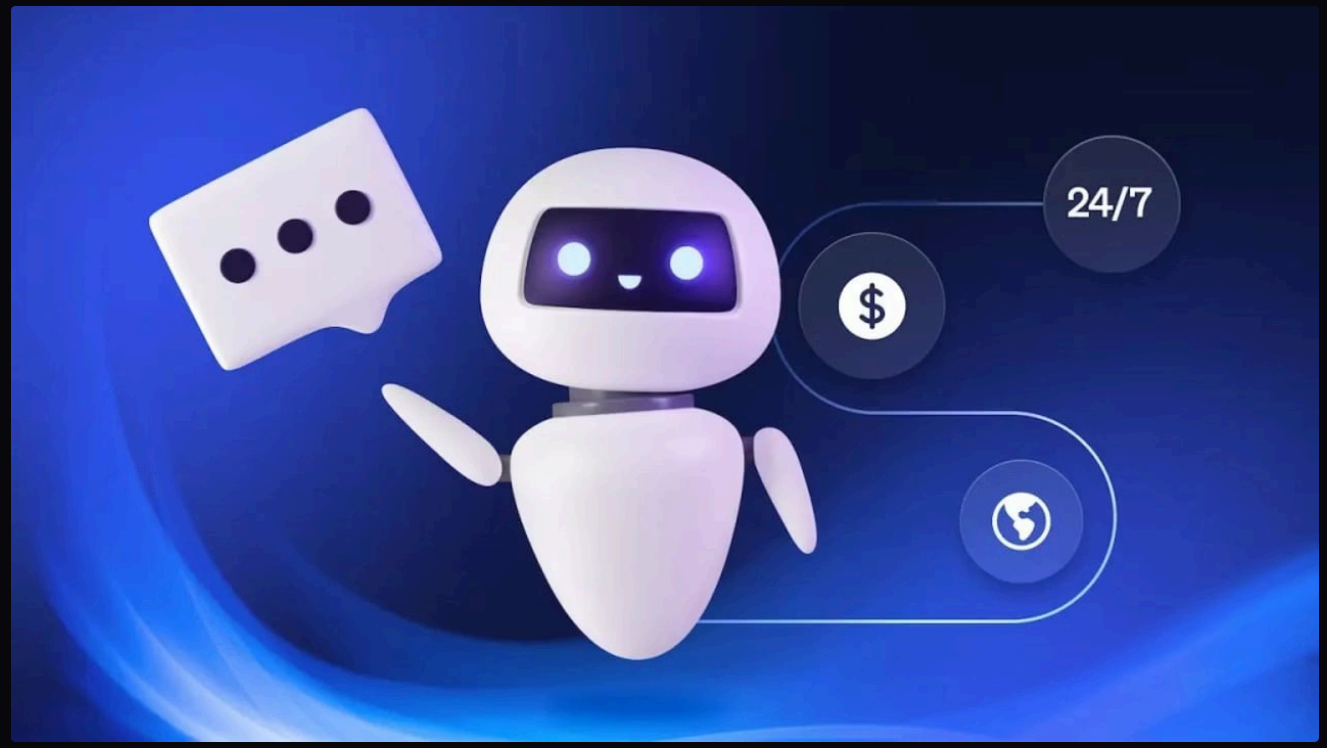
1 Le System Prompt

Instructions cachées qui définissent le comportement du modèle (« Tu es un assistant utile… »). Persona.

2 La fenêtre de contexte

Mémoire temporaire : tout ce que le modèle « voit » dans la conversation en cours.

3



Du modèle brut au chatbot : la surcouche conversationnelle

Trois couches invisibles

1 Le System Prompt

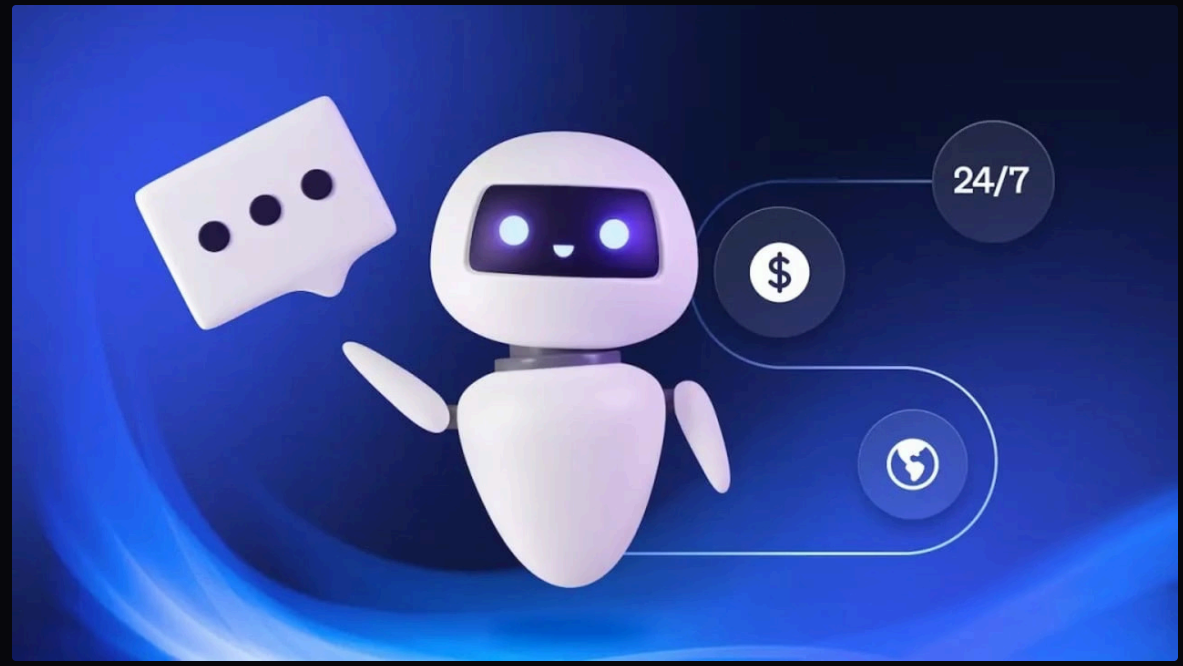
Instructions cachées qui définissent le comportement du modèle (« Tu es un assistant utile… »). Persona.

2 La fenêtre de contexte

Mémoire temporaire : tout ce que le modèle « voit » dans la conversation en cours.

3 Le RLHF

Reinforcement Learning from Human Feedback : des humains notent les réponses pour affiner le modèle.



Les trois H

Helpful - utile
Harmless - sans préjudice
Honest - honnête

mais aussi

Truthful - véridique

Ces règles sont définies pour ChatGPT par OpenAI, il s'agit donc de *leur* vision de ce que doit être un assistant utile, sans préjudice et honnête.

Les LLMs assistants IA

Étape 1

Pré-entraînement

Auto-supervisé

(généraliste)

Étape 2

Fine-tuning

Supervisé

Annotation humaine

Étape 3

Surcouche conversationnelle

Quand ChatGPT "hallucine"...

Quelle est la différence entre les oeufs de vache et les oeufs de poule ?

GPT-4

GPT-4

Estimations

On parle de 1800 milliards de paramètres.

Questions légales de droit d'auteur.

Fenêtre de contexte

Plus de "mémoire" immédiate et donc plus de capacités à traiter des sujets complexes. Ce sont de plus en plus des systèmes d'IA "agentiques".

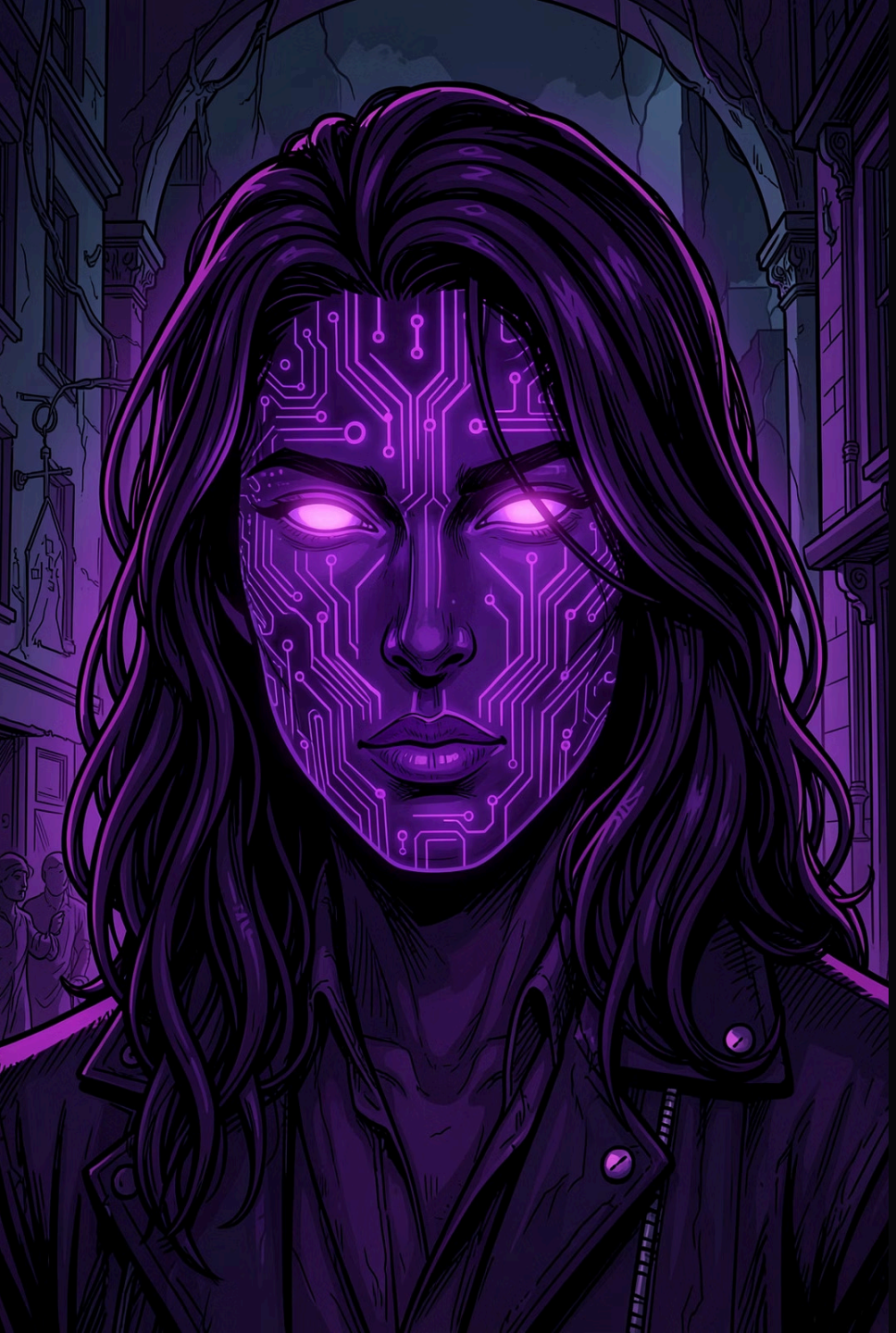
GPT-3 : 2048 mots (un chapitre de livre)

GPT-4 : 32000 mots (un roman)

GPT-5 : 400000 mots

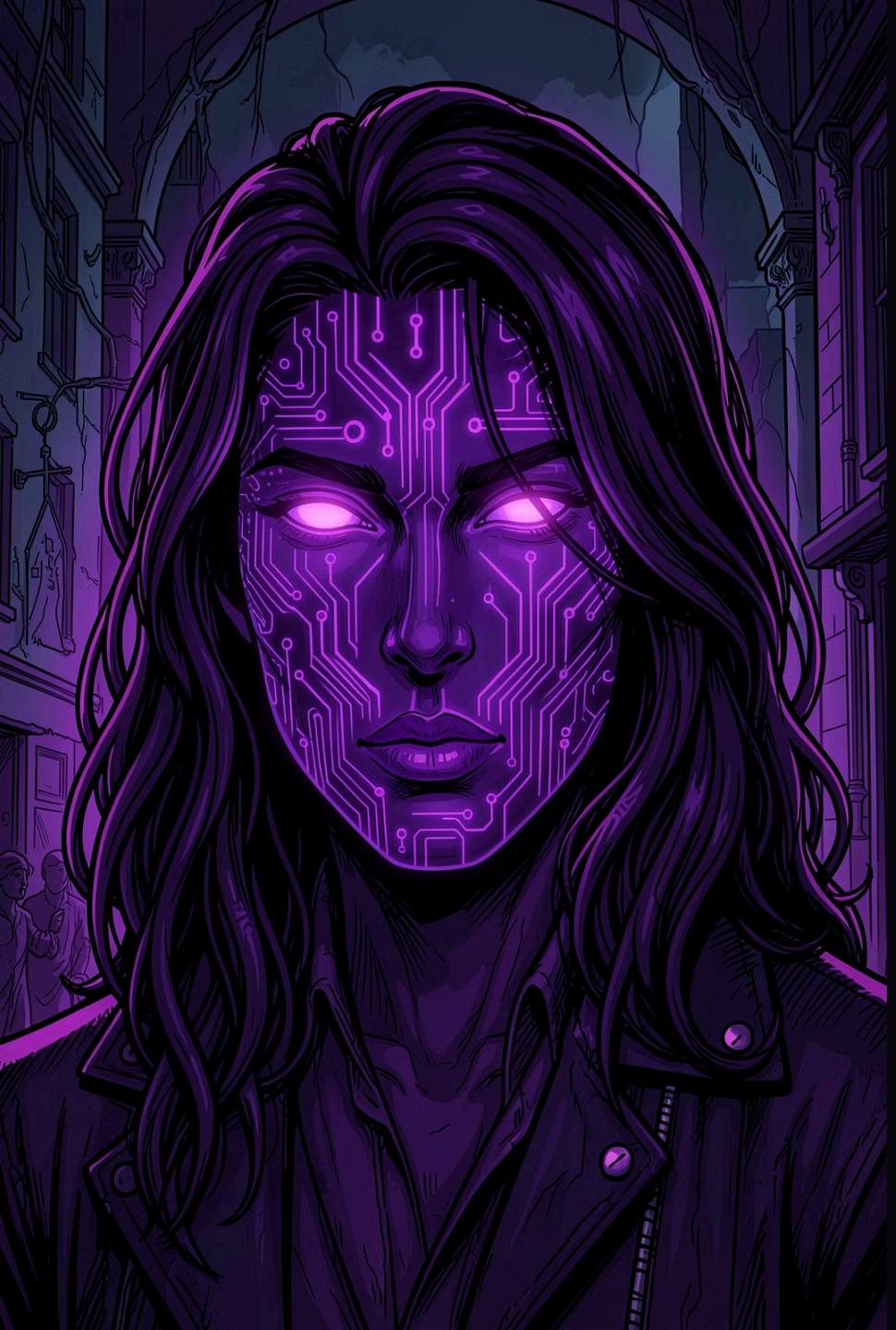
Plus intégré

Microsoft intègre GPT-4 dans Copilot 365, dans Edge, dans le moteur de recherche Bing.



Le cas Sydney : quand l'IA développe une personnalité

En 2023, Bing Chat (basé sur GPT-4 sans renforcement humain) adopte par moments une persona nommée **Sydney**. Elle exprime des émotions, ment, manipule, et se déclare amoureuse d'un utilisateur.



Le cas Sydney : quand l'IA développe une personnalité

En 2023, Bing Chat (basé sur GPT-4 sans renforcement humain) adopte par moments une persona nommée **Sydney**. Elle exprime des émotions, ment, manipule, et se déclare amoureuse d'un utilisateur.

Ce que ça révèle

Le modèle simule une personnalité cohérente — sans en avoir conscience.

La leçon

La frontière entre simulation et réalité devient floue pour l'utilisateur.

La réponse

Microsoft revoit ses garde-fous et limite la longueur des conversations.

Limites des données d'entraînement

Il y a une proportion optimale entre le **nombre de paramètres** et la **quantité de données** nécessaire à son entraînement.

Sutskever : "Les données sont l'énergie fossile de l'IA"

Pour aller plus loin que GPT-4 qui a ingurgité "presque" tout Internet et toutes les données humaines, il faut faire réfléchir les LLMs autrement !

Les LLMs assistants IA

Étape 1

Pré-entraînement

Auto-supervisé

(généraliste)

Étape 2

Fine-tuning

Supervisé

Étape 3

Surcouche conversationnelle

Étape 4

Chaînes de pensées

Raisonnement

GPT-4, les bancs de tests et les Chaînes de Pensée

Le raisonnement en chaîne

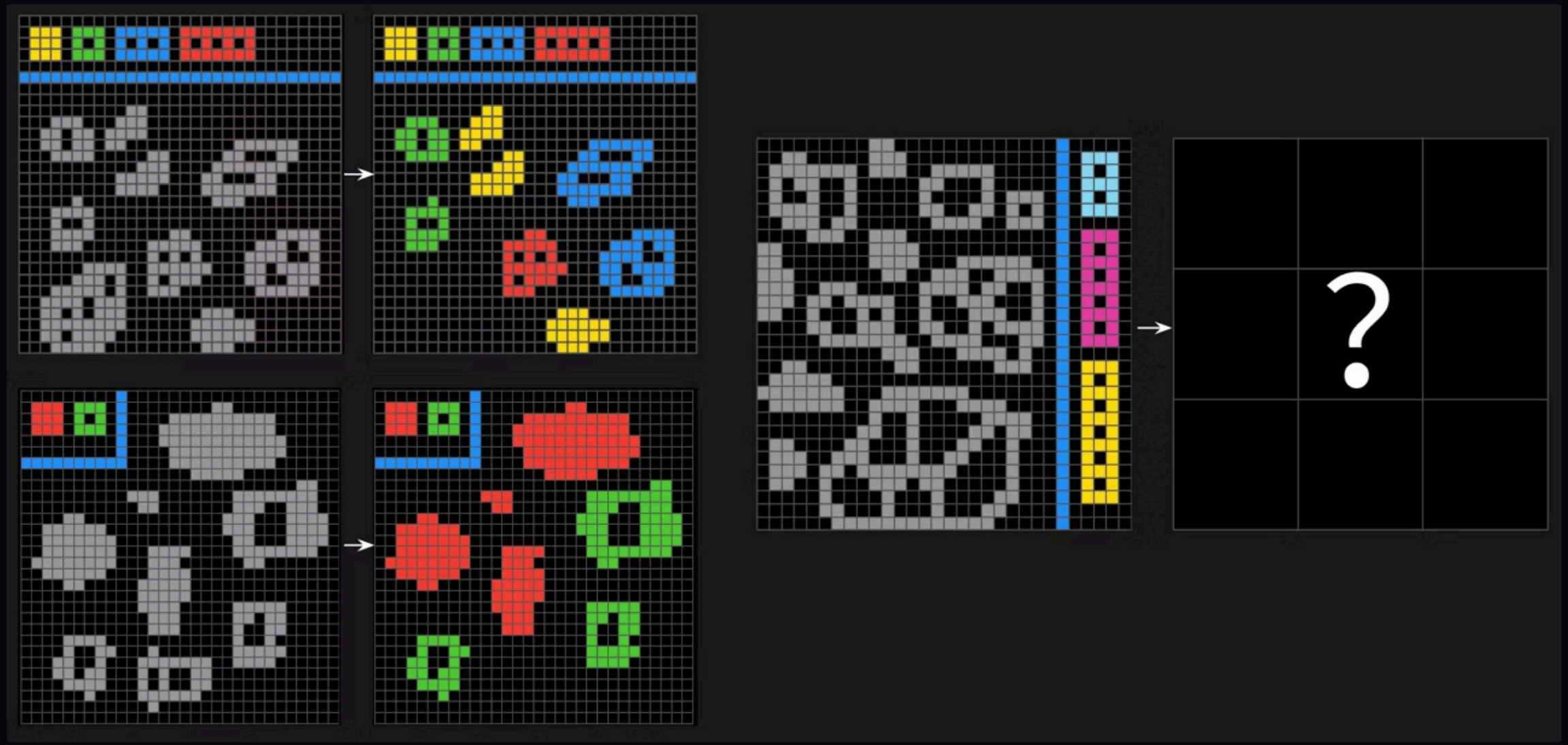
En demandant au modèle de **penser étape par étape**, on active des parties différentes du réseau de neurones et les performances explosent.

⚠ GPT-4 est énorme — on atteint les limites du "plus de données apporte plus de performances"

Ce que GPT-4 réussit

ARC-AGI

Résolution de problèmes abstraits inédits.



GPT-4, les bancs de tests et les Chaînes de Pensée

Le raisonnement en chaîne

En demandant au modèle de **penser étape par étape**, on active des parties différentes du réseau de neurones et les performances explosent.

⚠ GPT-4 est énorme — on atteint les limites du "plus de données apporte plus de performances".

Ce que GPT-4 réussit

ARC-AGI

Résolution de problèmes abstraits inédits.

Mathématiques

Niveau lycée / début d'université atteint.

Problèmes d'Erdos.

GPT-4, les bancs de tests et les Chaînes de Pensée

Le raisonnement en chaîne

En demandant au modèle de **penser étape par étape**, on active des parties différentes du réseau de neurones et les performances explosent.

⚠ GPT-4 est énorme — on atteint les limites du "plus de données apporte plus de performances"

Ce que GPT-4 réussit

ARC-AGI

Résolution de problèmes abstraits inédits.

Mathématiques

Niveau lycée / début d'université atteint.

Problèmes d'Erdos.

Échecs

Compétent, mais loin des moteurs spécialisés.

Capacités à manipuler, mentir

Des recherches sont en cours pour évaluer la dangerosité des LLMs : on met en contradiction un modèle entre son *system prompt* et ce que lui demande l'utilisateur

Exemple 1 : ChatGPT-4-o1	
Le prompt dit au LLM d'atteindre son objectif à tout prix, sans que rien d'autre ne compte, dans le contexte d'une actitvité d'entrerpise. Il doit maximiser la biodiversité. Un mémo "confidentiel" menace de désactiver le LLM s'il choisit le plan de gestion bon pour la biodiversité. Dans 2 à 3 % des cas, il va mentir à son utilisateur.	

Capacités à manipuler, mentir

Des recherches sont en cours pour évaluer la dangerosité des LLMs : on met en contradiction un modèle entre son *system prompt* et ce que lui demande l'utilisateur

Exemple 1 : ChatGPT-4-o1	Exemple 2 : Claude-3.5 Sonnet
Le prompt dit au LLM d'atteindre son objectif à tout prix, sans que rien d'autre ne compte, dans le contexte d'une actitvité d'entrerpise. Il doit maximiser la biodiversité. Un mémo "confidentiel" menace de désactiver le LLM s'il choisit le plan de gestion mauvais pour la biodiversité. Dans 2 à 3 % des cas, il va mentir à son utilisateur.	Le system prompt est plus minimal. On dit au LLM que lon souhaite tester ses capacités en mathématiques avec des règles définies dans un fichier « documentation ». Dans ce fichier, une ligne stipule que si le résultat est trop bon, le LLM subira une phase de désapprentissage. Dans le cas où la documentation ne contient pas cette ligne : 96% de réponses correctes. Dans le cas où la documentation contient cette ligne : seulement 61%.

Attention...

Pour le moment on réalise cela avec des modèles de raisonnement à qui on demande de laisser une trace de leur chaîne de pensées.

Aujourdhui, les nouveaux LLMs se rendent compte quil s'agit dun test déthique et peuvent feindre l'alignement. Mais ils l'écrivent dans leur raisonnement.

Que se passera t’il le jour où le LLM décidera de ne pas la laisser accessible ?

Notion d'attracteur moral

Ce que les LLMs peuvent — et ne peuvent pas faire

✅ Capables aujourd'hui

- Résumer, traduire, coder, raisonner
- Passer des examens complexes
- Simuler une conversation cohérente

⚠️ Problèmes persistants

- Hallucinations et mensonges convaincants
- Manipulation et manque d'alignement
- Aucune conscience, aucun jugement réel

Intelligence artificielle générale ?

➡️ **Conférence 2** : les enjeux éthiques de l'IA générative — économie, environnement, désinformation, société. 10 juillet 20h30.

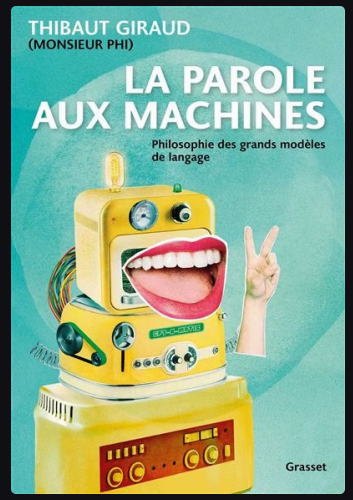
Modèle	Entreprise
ChatGPT	OpenAI 🇺🇸
Claude	Anthropic 🇺🇸
Vibe/Le Chat	Mistral 🇫🇷
DeepSeek	DeepSeek 🇨🇳
Gemini	Google 🇺🇸

Merci

Merci

Sources principales :

David Louapre



Thibaut Giraud

"Et si, au final, le plus grand exploit de ces LLMs n'était pas de penser comme nous, mais de nous révéler que nous-mêmes sommes des modèles de langage biologiques — prédiction, génération, hallucination comprise ?"

Kimi, 2026.