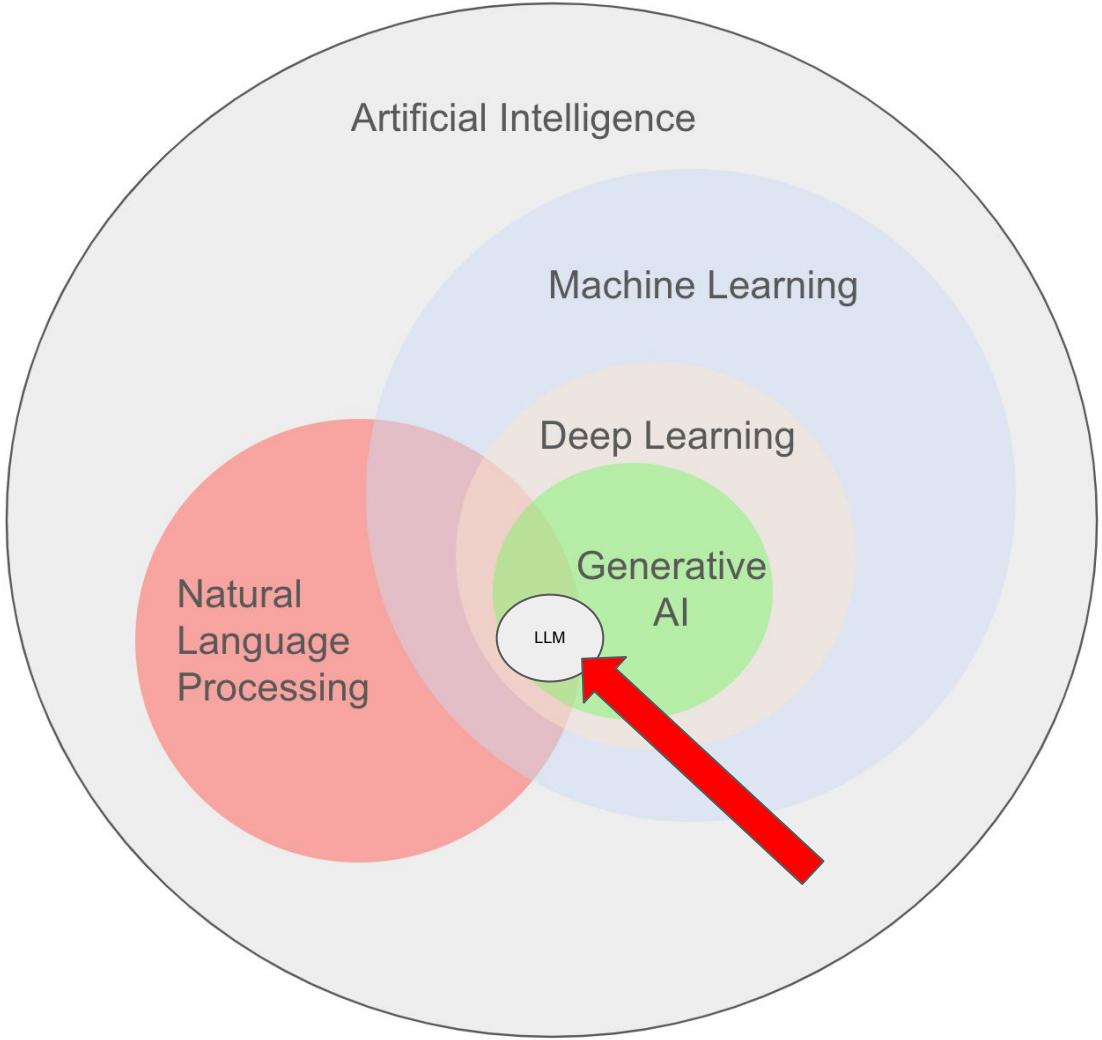
SmartMonday

7 Avril 2025

Large Languages Models (LLMs)

By Guillaume Wafflard
a.k.a crevetteboiii(i)*

Generate an image of a conference with geeks about large language models, transformers and attention mechanism. Ajoute plus de diversité parmi les participants (tout en restants geeks). L'ambiance est chill et les gens boivent des bières ou du club mate


Artificial Intelligence
Machine Learning
Deep Learning
Generative AI
LLM
Natural Language Processing

“Un **algorithme** est une suite d’instructions précises permettant de résoudre un problème ou d’effectuer une tâche de manière systématique. Il suit une logique déterministe : **à une entrée donnée correspond toujours une sortie prévisible**.

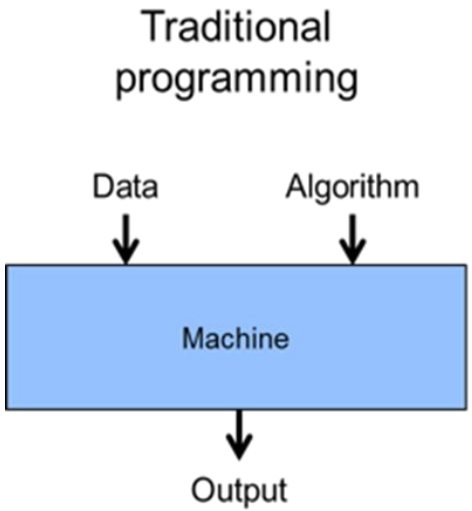
🖥️ **Exemple** : Un tri par insertion prend une liste de nombres en entrée et applique un ensemble de règles fixes pour les classer par ordre croissant.”

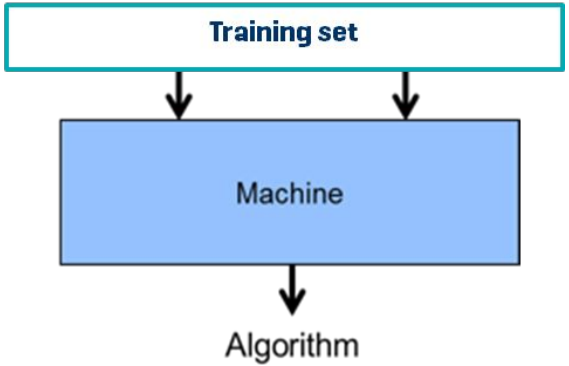
source: ChatGPT

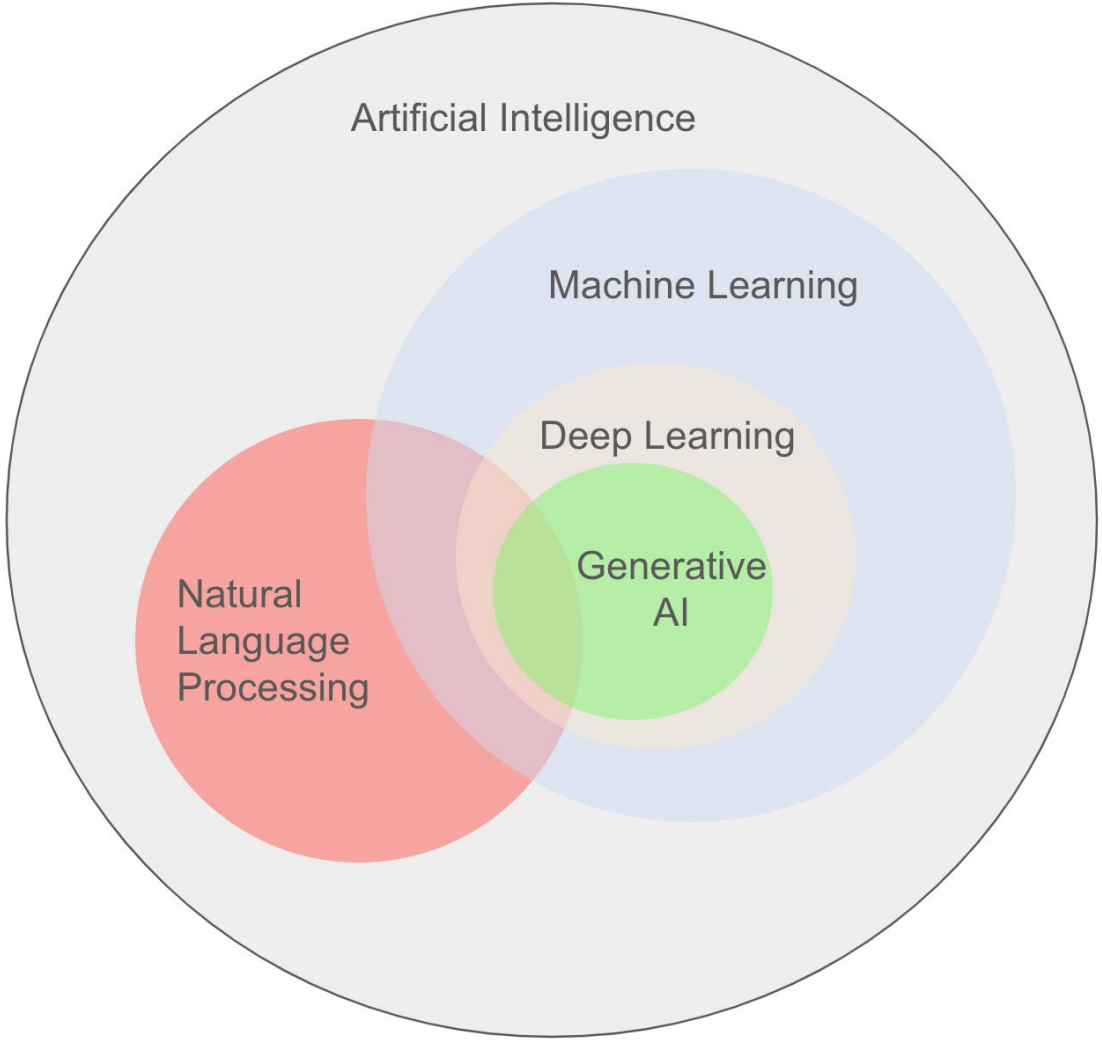
“L’**IA** est un ensemble de techniques permettant à une machine d’**imiter certaines capacités cognitives humaines**, comme l’apprentissage, la reconnaissance de modèles ou la prise de décision. Contrairement aux algorithmes classiques, l’IA **ne suit pas un ensemble fixe de règles**, mais apprend à partir de données.

🧠 **Exemple** : Un modèle de reconnaissance d’images n’a pas une règle explicite pour identifier un chat. Il a été entraîné sur des milliers d’images pour reconnaître les motifs caractéristiques (formes, textures, couleurs).”

source: ChatGPT


Traditional programming
Data
Algorithm
Machine
Output

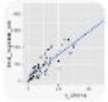

Machine learning
Training set
Machine
Algorithm


Artificial Intelligence
Machine Learning
Deep Learning
Generative AI
Natural Language Processing

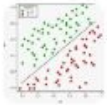
Ex d’algorithmes de Machine Learning

Top machine learning algorithms

From sources across the web



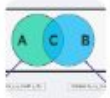
Linear regression



Logistic regression

Decision tree

Support Vector Machines



Naive Bayes

Random forest

Mean

KNN

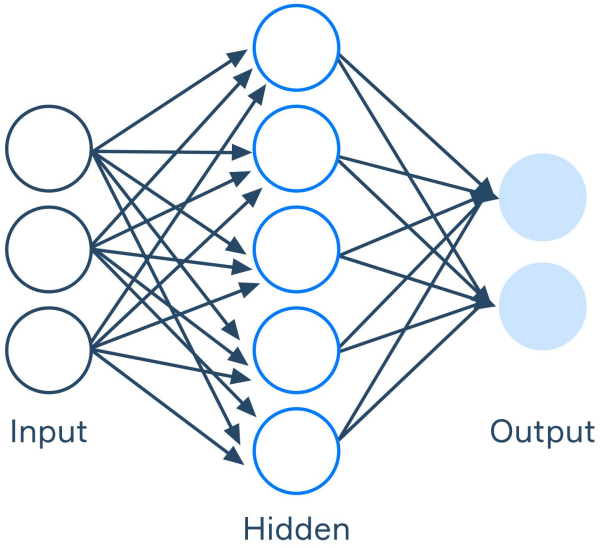
Neural network

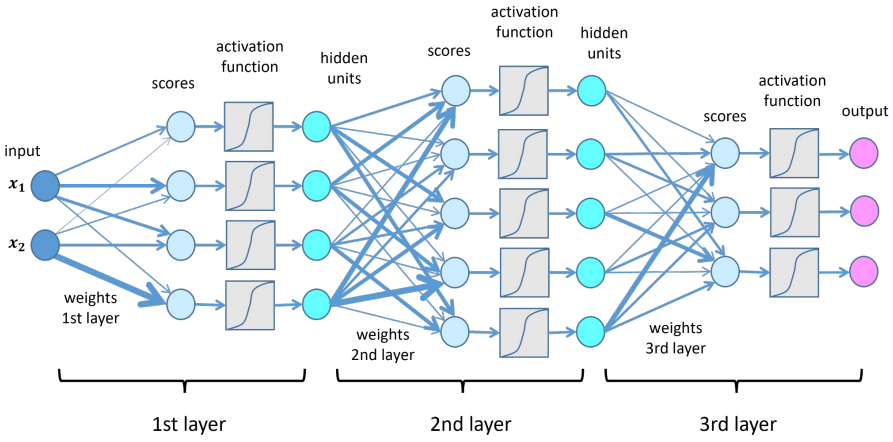
3 more

Feedback

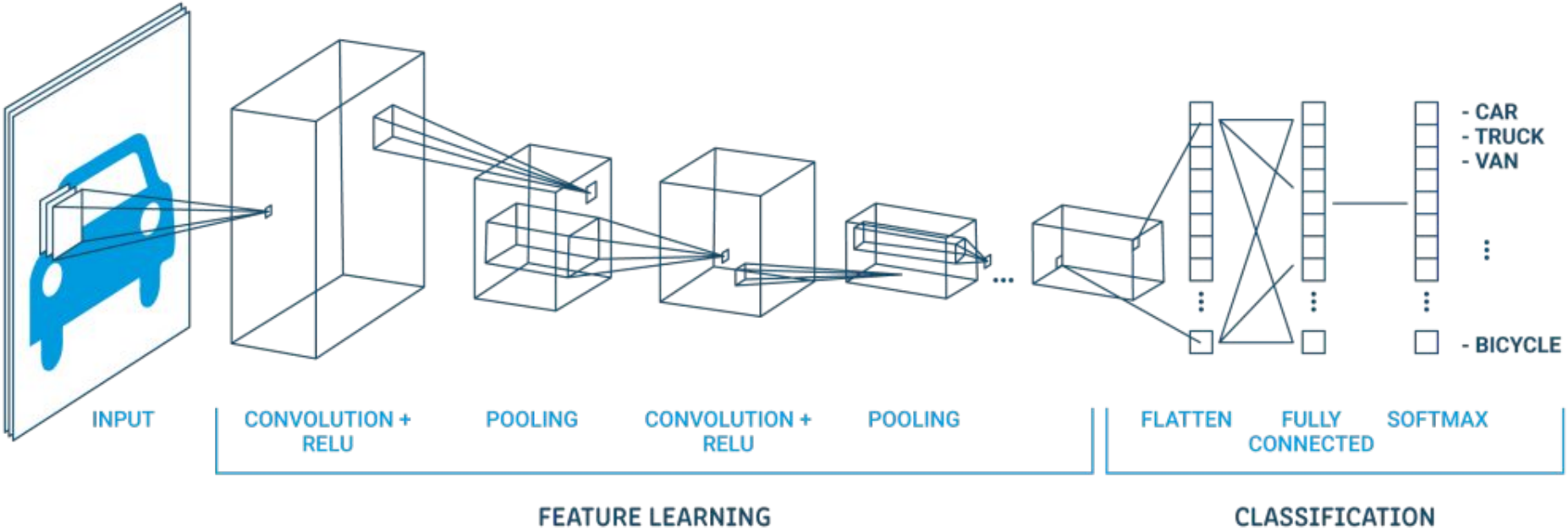
source: Google

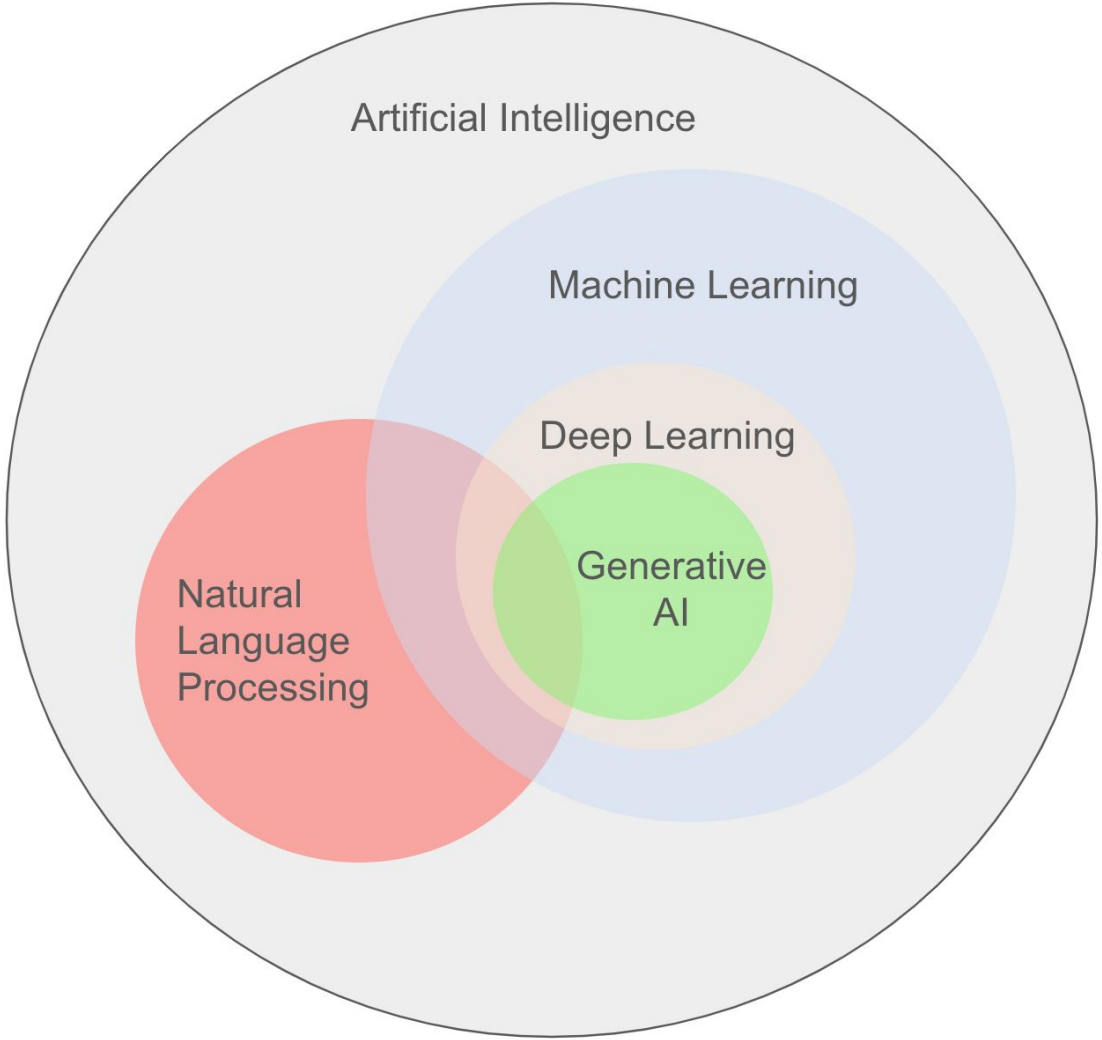
Neural Networks

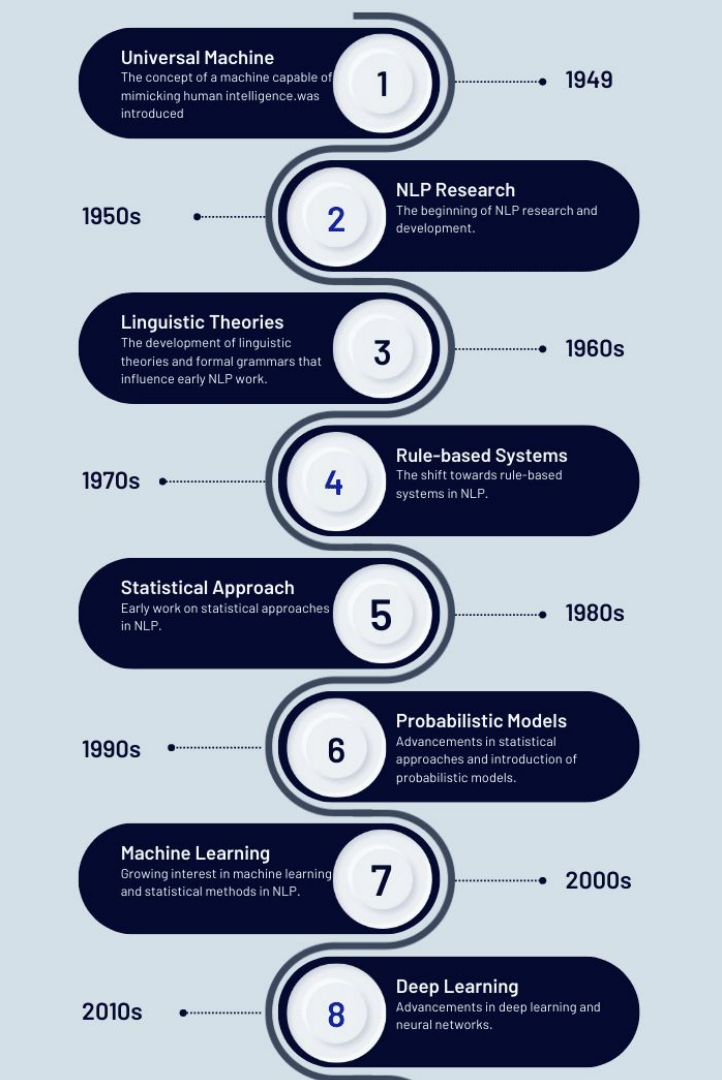


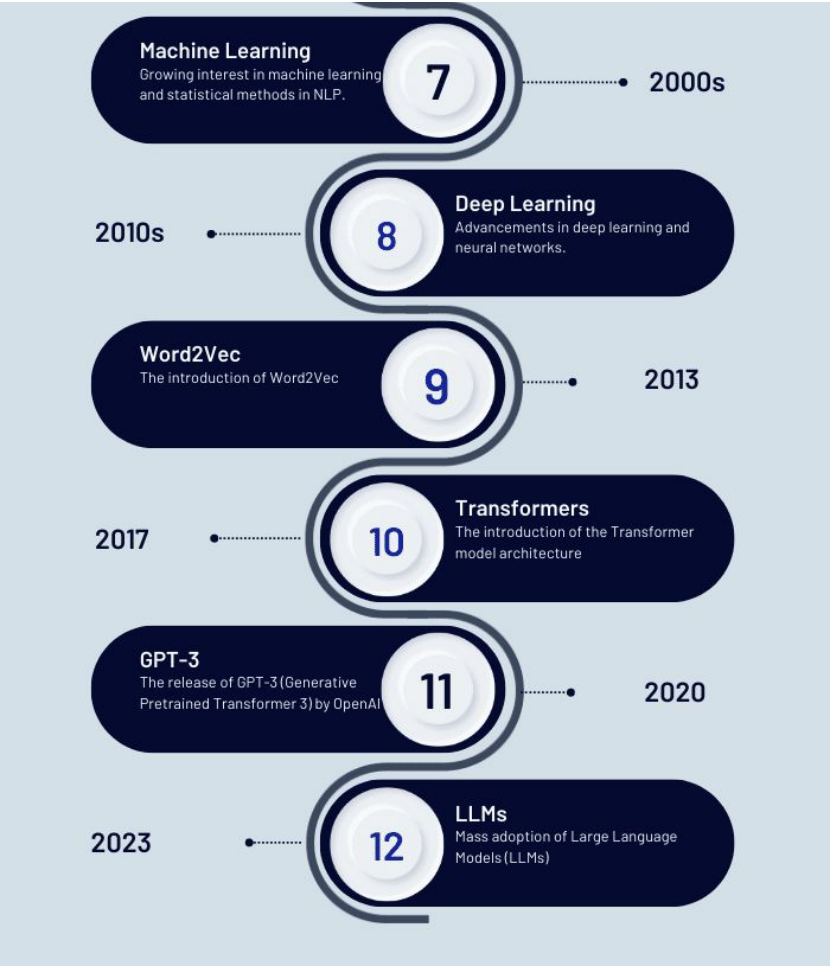


Deep Learning




Artificial Intelligence
Machine Learning
Deep Learning
Generative AI
Natural Language Processing


Universal Machine
The concept of a machine capable of mimicking human intelligence.was introduced
1
1949
1950s
2
NLP Research
The beginning of NLP research and development.
Linguistic Theories
The development of linguistic theories and formal grammars that influence early NLP work.
3
1960s
1970s
4
Rule-based Systems
The shift towards rule-based systems in NLP.
Statistical Approach
Early work on statistical approaches in NLP.
5
1980s
1990s
6
Probabilistic Models
Advancements in statistical approaches and introduction of probabilistic models.
Machine Learning
Growing interest in machine learning and statistical methods in NLP.
7
2000s
2010s
8
Deep Learning
Advancements in deep learning and neural networks.


Machine Learning
Growing interest in machine learning and statistical methods in NLP.
7
2000s
2010s
8
Deep Learning
Advancements in deep learning and neural networks.
Word2Vec
The introduction of Word2Vec
9
2013
2017
10
Transformers
The introduction of the Transformer model architecture
GPT-3
The release of GPT-3 (Generative Pretrained Transformer 3) by OpenAI
11
2020
2023
12
LLMs
Mass adoption of Large Language Models (LLMs)

```
  Welcome to
                    EEEEEE  LL      IIII   ZZZZZZ   AAAAA
                    EE      LL       II       ZZ   AA   AA
                    EEEEE   LL       II     ZZZ    AAAAAAA
                    EE      LL       II    ZZ      AA   AA
                    EEEEEE  LLLLLL  IIII ZZZZZZ    AA   AA

  Eliza is a mock Rogerian psychotherapist.
  The original program was described by Joseph Weizenbaum in 1966.
  This implementation by Norbert Landsteiner 2005.


ELIZA: Is something troubling you ?
YOU:   Men are all alike.
ELIZA: What is the connection, do you suppose ?
YOU:   They're always bugging us about something or other.
ELIZA: Can you think of a specific example ?
YOU:   Well, my boyfriend made me come here.
ELIZA: Is it important to you that your boyfriend made you come here ?
YOU:   He says I'm depressed much of the time.
ELIZA: I am sorry to hear that you are depressed.
YOU:   It's true. I am unhappy.
ELIZA: Can you explain what made you unhappy ?
YOU:
```

Siri

Use your voice to send messages, set reminders, search for information, and more.

```
Up a Tree                                    Score: 5        Moves: 10
>follow path
You must specify a direction to go.

>n
Forest Path
This is a path winding through a dimly lit forest. The path heads north-south
here. One particularly large tree with some low branches stands at the edge of
the path.

>climb tree
Up a Tree
You are about 10 feet above the ground nestled among some large branches. The
nearest branch above you is above your reach.
Beside you on the branch is a small bird's nest.
In the bird's nest is a large egg encrusted with precious jewels, apparently
scavenged by a childless songbird. The egg is covered with fine gold inlay,
and ornamented in lapis lazuli and mother-of-pearl. Unlike most eggs, this one
is hinged and closed with a delicate looking clasp. The egg appears extremely
fragile.

>take egg
Taken.

>
```

Context: Zork - Wikipedia
Play the game: https://github.com/RickyVimon/Zork/tree/master

Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com
Noam Shazeer* Google Brain noam@google.com
Niki Parmar* Google Research nikip@google.com
Jakob Uszkoreit* Google Research usz@google.com
Llion Jones* Google Research llion@google.com
Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu
Łukasz Kaiser* Google Brain lukaszkaiser@google.com
Illia Polosukhin* ‡ illia.polosukhin@gmail.com

Abstract

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

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

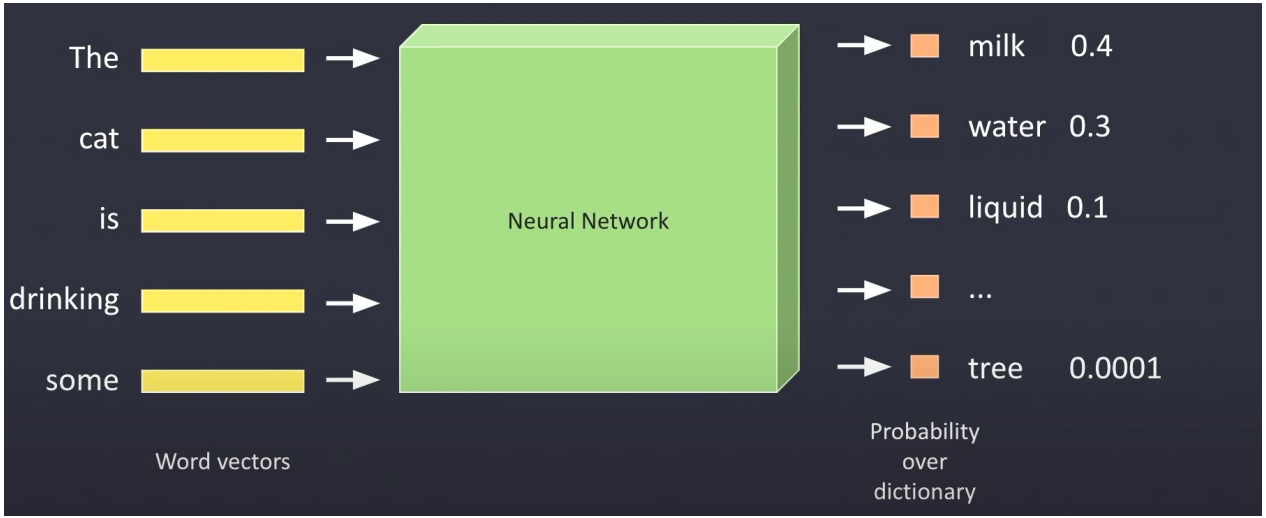
†Work performed while at Google Brain.

‡Work performed while at Google Research.

A. Vaswani et al.
“Attention is all you need” (2017)
https://arxiv.org/pdf/1706.03762
175.000+ citations

Qu’est-ce qu’un LLM ?

- Seule chose que sait faire un LLM → Prédire le prochain mot



Source: [Democratizing large language models (Armand Joulin, Research Director @ META)](#) (33’)

Définition

Un LLM est une fonction mathématique sophistiquée qui prédit le mot suivant dans n’importe quel morceau de texte.

\[LLM : $f(\text{texte}) = \arg\max_{w \in \mathcal{V}} P(w \mid \text{texte})$\]où :

$f(\text{texte})$ est la fonction du modèle de langage
w est le mot suivant prédit
$\mathcal{V}$ est le vocabulaire du modèle
$P(w \mid \text{texte})$ est la probabilité d’un mot donné le contexte précédent

Source: ChatGPT - “Crée une fonction mathématique en LaTeX qui illustre cette définition”

Entrainement d’un LLM

Entraînement sur une quantité massive de textes

1. On donne au modèle une phrase incomplète :
   → *"Les chats aiment dormir sur le ___"*
2. Le modèle propose un mot (au début, un peu au hasard)
3. On compare le mot généré avec le mot réel attendu
4. On ajuste les paramètres (via *backpropagation*)

→ Répété des milliards de fois

Entrainement d’un LLM

Résultat :

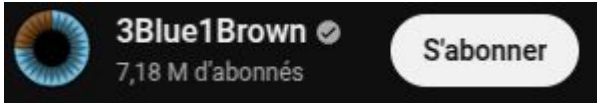
➡️ Le modèle apprend à générer des phrases cohérentes, même avec du texte qu’il n’a **jamais vu**.

Et ensuite ?

🔁 Entraînement par renforcement avec feedback humain (RLHF)

➡️ Permet de transformer ce modèle en **chatbot** interactif (comme ChatGPT)

Vidéos d’explication complète



Résumé en 7 minutes: Large Language Models explained briefly

Transformers (26’) : Transformers (how LLMs work) explained visually | DL5

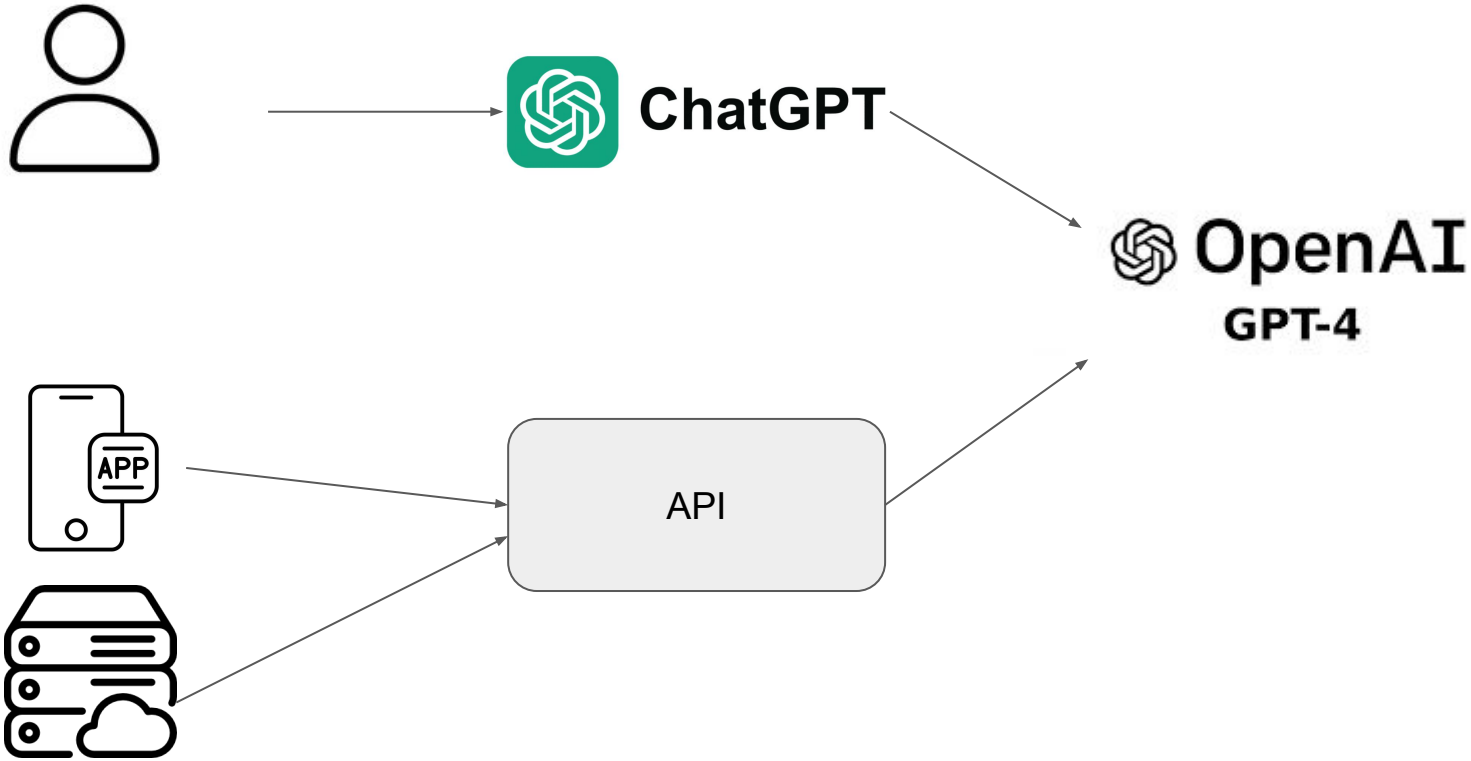
Attention mechanism (27’): Attention in transformers, step-by-step | DL6

En bref

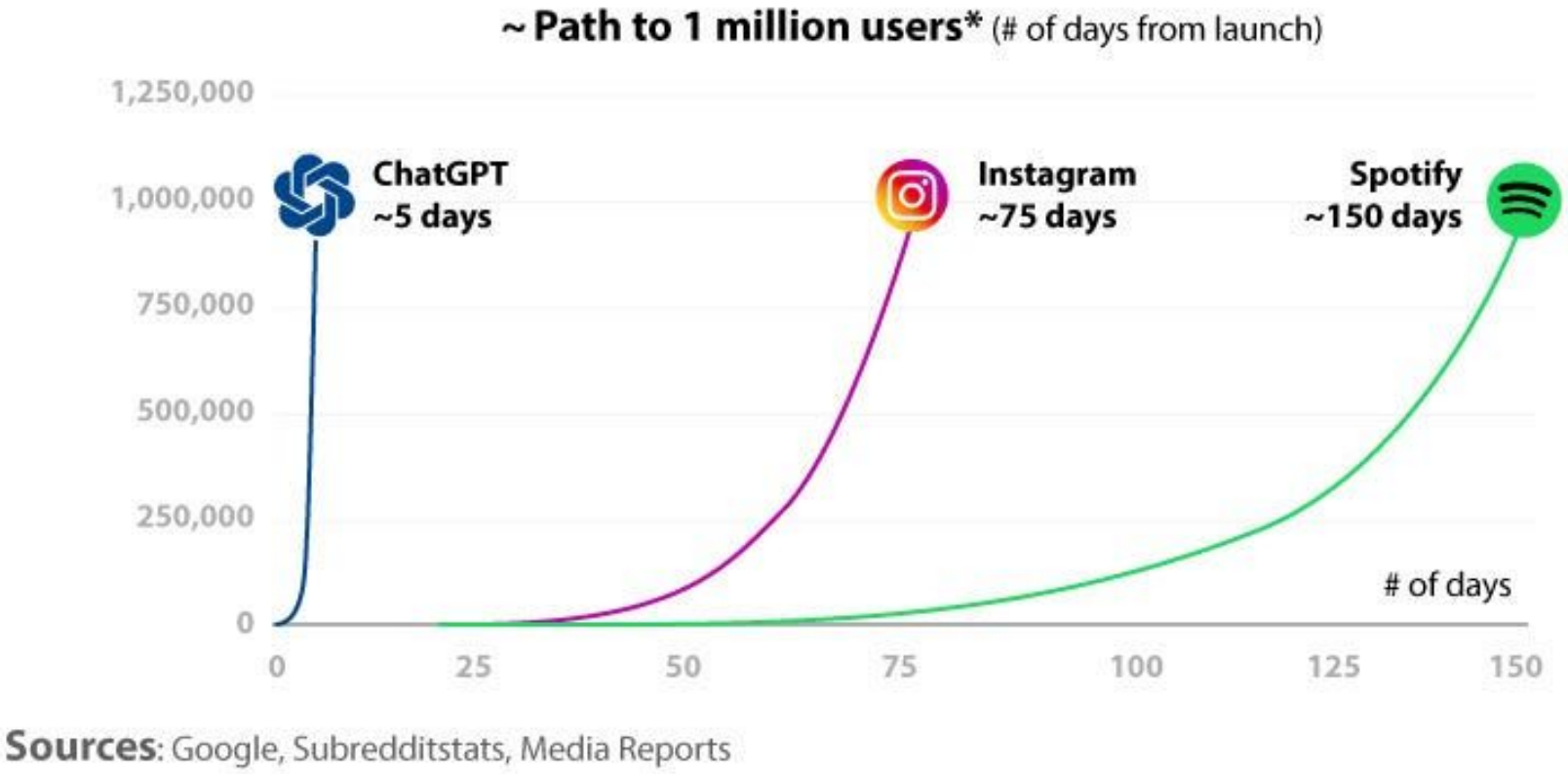
En résumé:

- Un modèle d’IA, sous-catégorie du Deep Learning
- Word embeddings (vecteurs de mots) → Comprendre le sens des mots
- Attention mechanism (comprendre les interactions entre les mots)
- But: prédire le prochain mot
- ChatGPT = LLM entraîné comme un “assistant”

Utilisation



ChatGPT : buzz immédiat



ChatGPT

Claude

NotebookLM

Gemini

Meta AI

cohere

perplexity

glean

MISTRAL AI_

Hugging Face | Search models, datasets, users... | Models | Datasets | Spaces | Posts | Docs | Enterprise | Pricing | Log In | Sign Up

The AI community building the future.

The platform where the machine learning community collaborates on models, datasets, and applications.

Explore AI Apps or Browse 1M+ models

Tasks | Libraries | Datasets | Languages | Licenses | Other

Filter Tasks by name

Multimodal

- Text-to-Image
- Image-to-Text
- Text-to-Video
- Visual Question Answering
- Document Question Answering
- Graph Machine Learning

Computer Vision

- Depth Estimation
- Image Classification
- Object Detection
- Image Segmentation
- Image-to-Image
- Unconditional Image Generation
- Video Classification
- Zero-Shot Image Classification

Natural Language Processing

- Text Classification
- Token Classification
- Table Question Answering
- Question Answering
- Zero-Shot Classification
- Translation
- Summarization
- Conversational
- Text Generation
- Text2Text Generation
- Sentence Similarity

Audio

- Text-to-Speech
- Automatic Speech Recognition
- Audio-to-Audio
- Audio Classification
- Voice Activity Detection

Tabular

- Tabular Classification
- Tabular Regression

Reinforcement Learning

- Reinforcement Learning
- Robotics

Models 469,541 | Filter by name

- meta-llama/Llama-2-70b — Text Generation • Updated 4 days ago • 25.2k • 64
- stabilityai/stable-diffusion-xl-base-0.9 — Updated 6 days ago • 2.01k • 393
- openchat/openchat — Text Generation • Updated 2 days ago • 1.3k • 136
- lllyasviel/ControlNet-v1-1 — Updated Apr 26 • 1.87k
- cerspense/zeroscope_v2_XL — Updated 3 days ago • 2.66k • 334
- meta-llama/Llama-2-13b — Text Generation • Updated 4 days ago • 328 • 64
- tiiuae/falcon-40b-instruct — Text Generation • Updated 27 days ago • 288k • 899
- WizardLM/WizardCoder-15B-V1.0 — Text Generation • Updated 3 days ago • 12.5k • 332
- CompVis/stable-diffusion-v1-4 — Text-to-Image • Updated about 17 hours ago • 448k • 5.72k
- stabilityai/stable-diffusion-2-1 — Text-to-Image • Updated about 17 hours ago • 782k • 2.81k
- Salesforce/xgen-7b-8k-inst — Text Generation • Updated 4 days ago • 6.18k • 57

Challenges

⚡ Energy and hardware consumption

🍄 Hallucinations

👨🏻 Bias from training data (Garbage in, garbage out)

📜 High volume of high quality text → mostly consumed

🎨 Copyrights

🧌🙈 Become AI addict

🔷➡️ 🔹 Small Language Models (SLM): cloud & local

Extensions

🔍 Recherches internet

👩‍💻 Code interpreter (*ex*: caculatrice, Python, …)

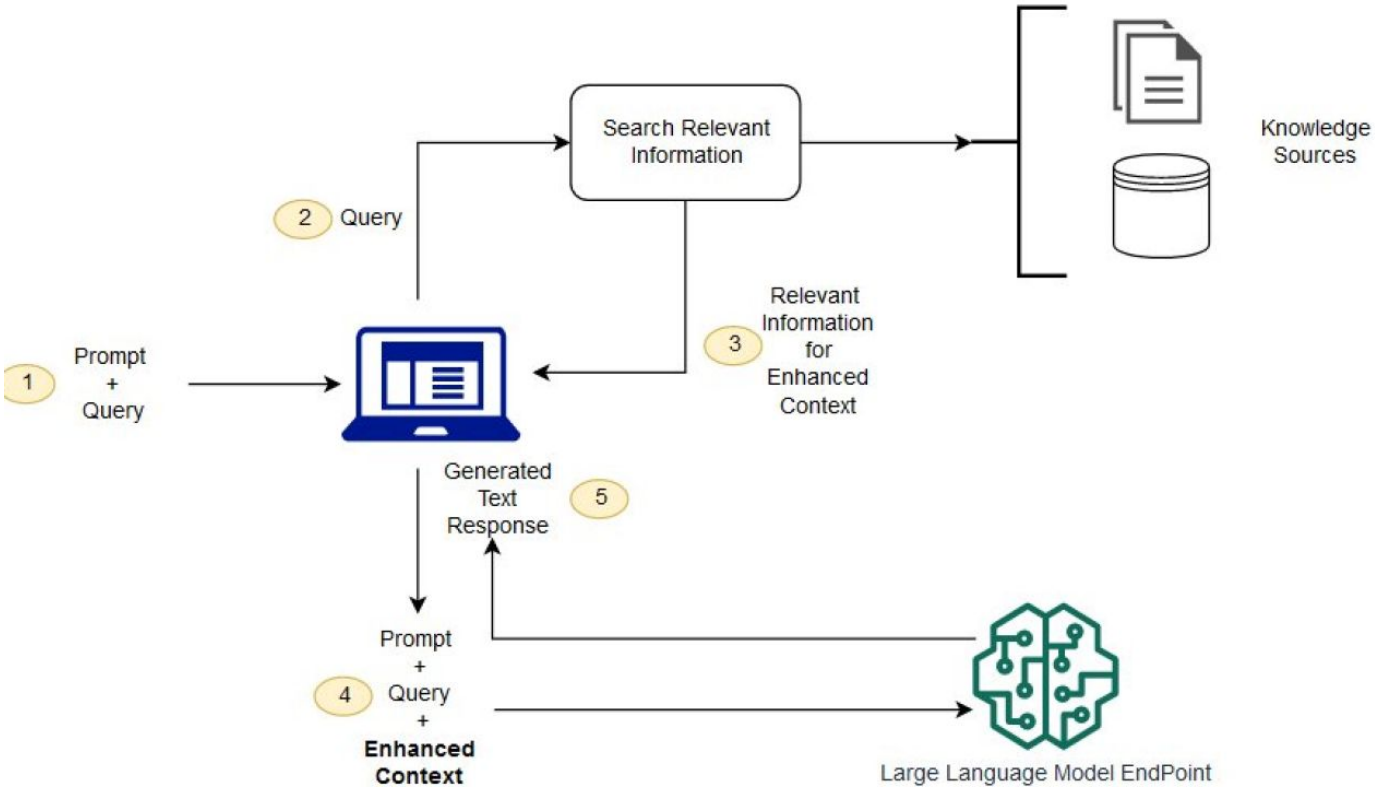
💭 Chain-of-Thoughts

🖼️ Image generation

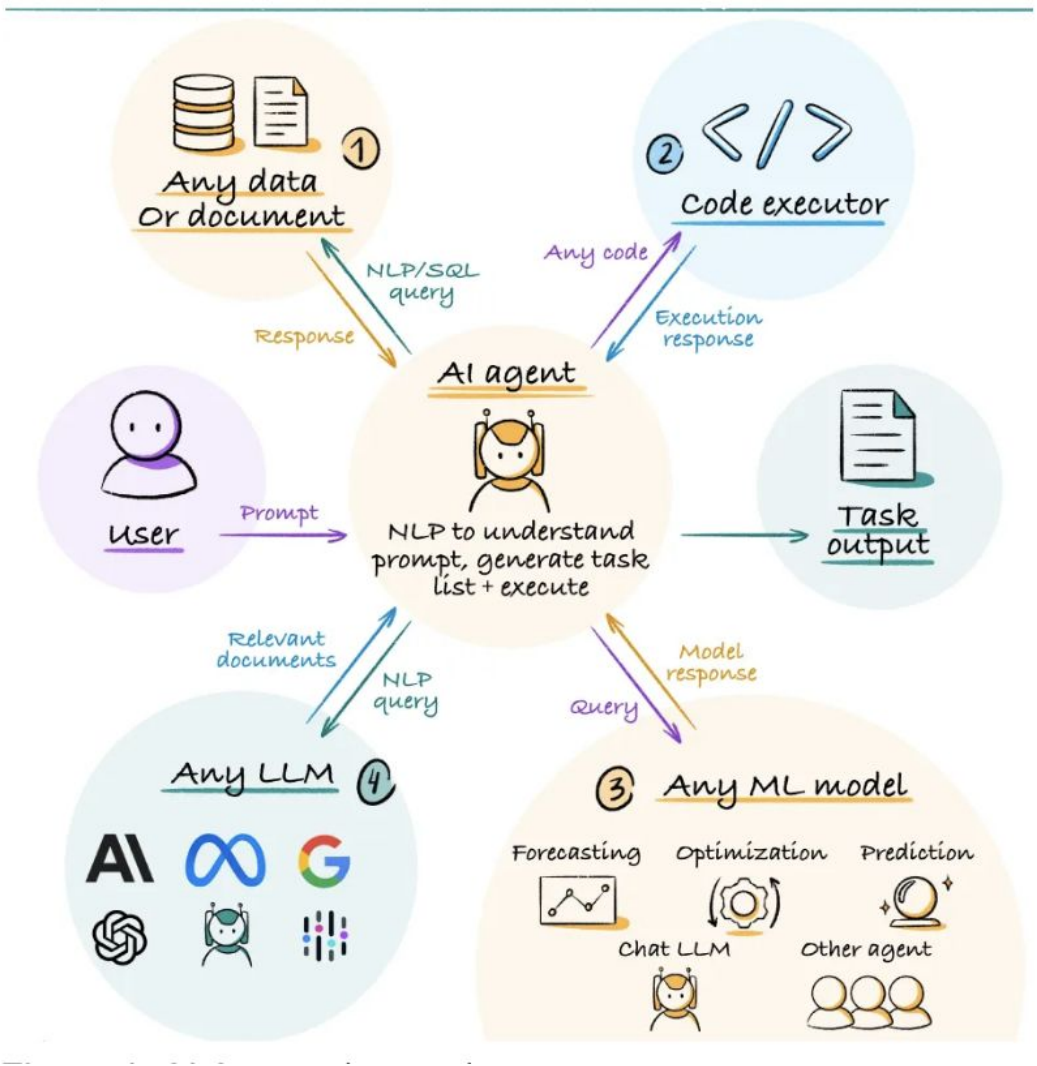
💻 Web browsing

🌈🤖 RAG (Retrieval-Augmented Generation)

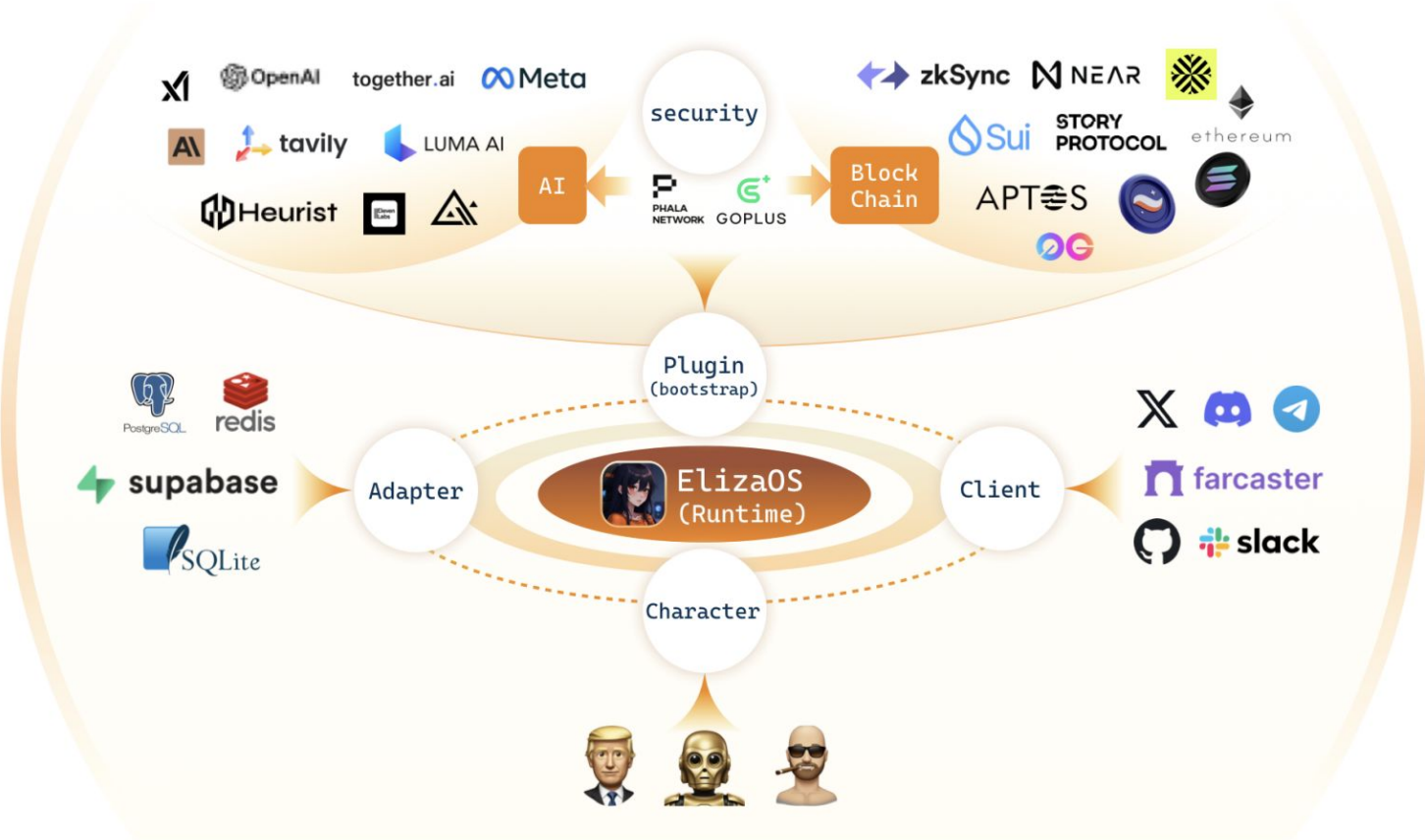
RAG (Retrieval-Augmented Generation)



AI Agents

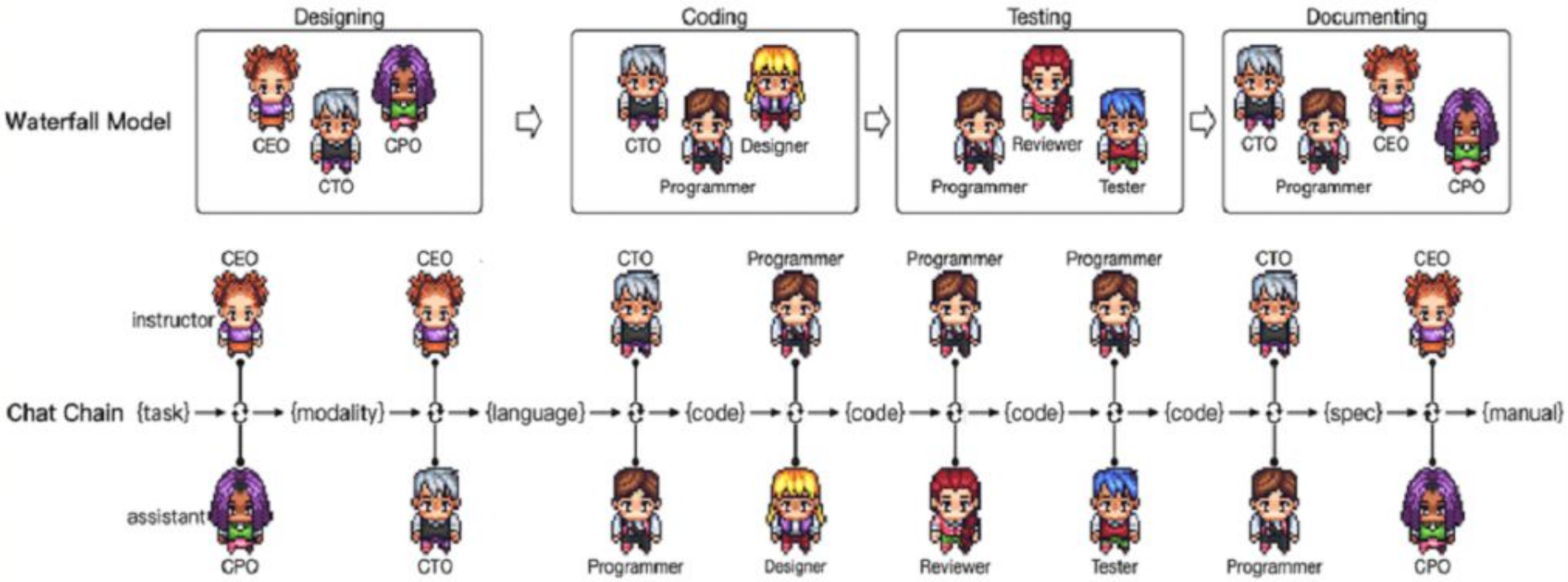

Any data
Or document
1
2
Code executor
NLP/SQL
query
Response
Any code
Execution
response
AI agent
User
Prompt
NLP to understand
prompt, generate task
list + execute
Task
output
Relevant
documents
NLP
query
Model
response
Query
Any LLM
4
3
Any ML model
Forecasting
Optimization
Prediction
Chat LLM
Other agent

ElizaOS



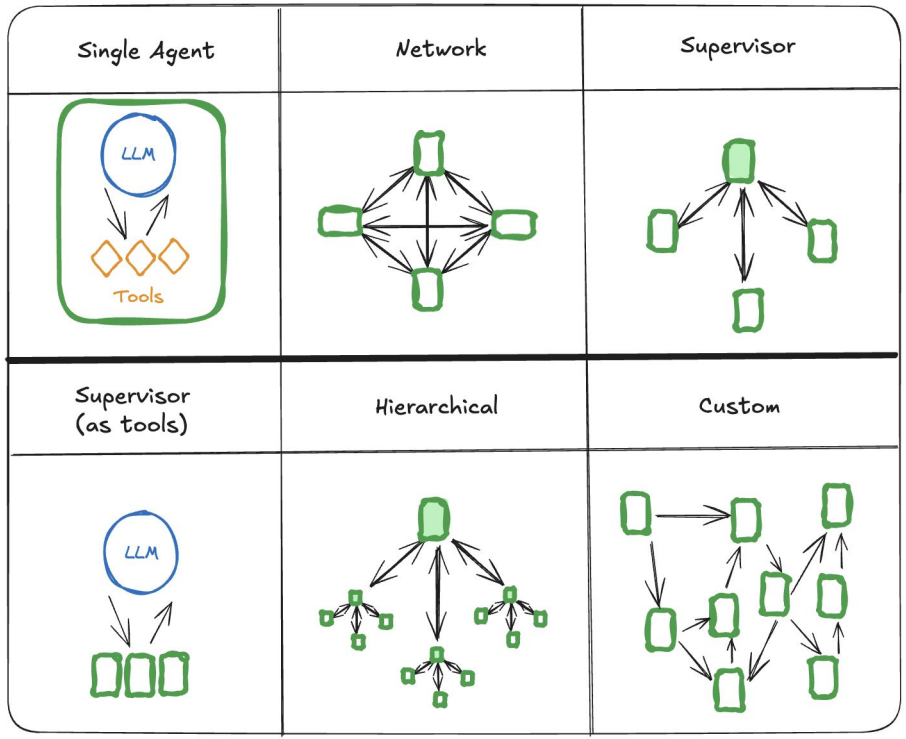
https://github.com/elizaOS/eliza

Agentic Design Patterns: Multi-Agent Collaboration



Proposed ChatDev architecture. Image adapted from "Communicative Agents for Software Development," Qian et al. (2023).

Multi-Agents System



Source:
https://langchain-ai.github.io/langgraph/concepts/multi_agent/

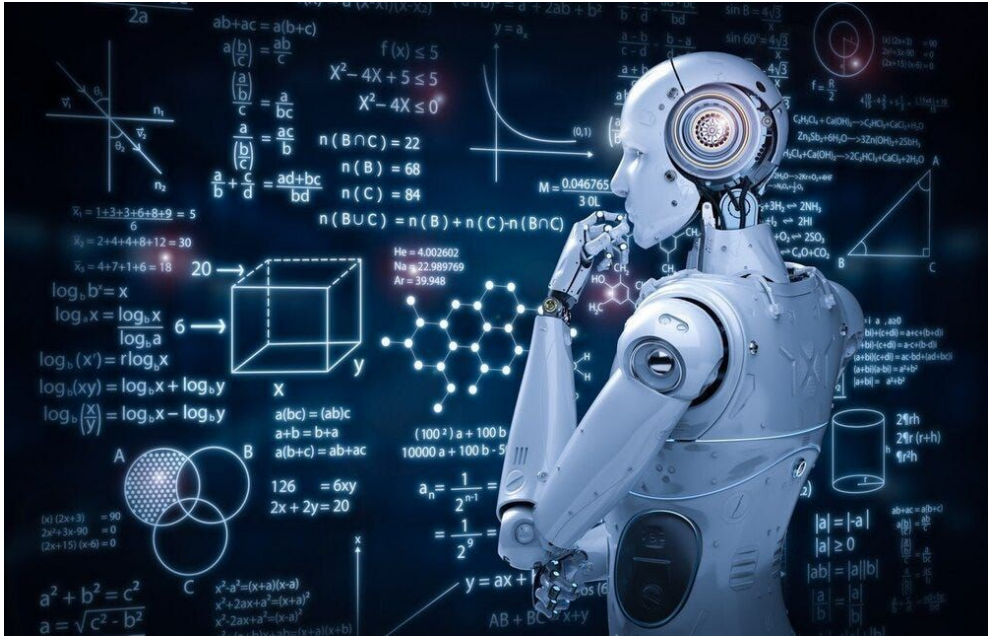
Artificial General Intelligence

“Les LLMs permettront-ils

d’atteindre l’AGI

(Artificial General Intelligence)

? ”



Yann LeCun • Following

VP & Chief AI Scientist at Meta

1h •

Every intelligence is specialized, including human intelligence.
Intelligence is a collection of skills and an ability to acquire new ones quickly.
It cannot be measured with a scalar quantity.
No intelligence can be even close to general,
which is why the phrase "Artificial General Intelligence" makes no sense.
There is no question that machines will eventually equal and surpass human intelligence in all domains. But even those systems will not have "general" intelligence, for any reasonable definition of the word general.

Questions ?


WE ARE AN
AI STARTUP
API CALL
TO GPT-3
imgflip.com