



HAL
open science

Evolution des LLMs Open-Source

Yannis Bendi-Ouis, Xavier Hinaut

► **To cite this version:**

Yannis Bendi-Ouis, Xavier Hinaut. Evolution des LLMs Open-Source. Unithé ou Café à l'Inria, Feb 2024, Bordeaux, France. hal-04850997

HAL Id: hal-04850997

<https://hal.science/hal-04850997v1>

Submitted on 20 Dec 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

Ínria

Evolution des
LLMs Open-Source



Ínria

Evolution des
LLMs Open-Source
Weights



Inria

Yannis Bendi-Ouis

Doctorant dans l'équipe Mnemosyne
Encadré par Xavier Hinaut




Email :

yannis.bendi-ouis@inria.fr

Site web :

www.naowak.fr

Introduction

- > ChatGPT 
- > LLM : Large Language Model
- > Beaucoup de paramètres => beaucoup de calcul ⚠
- > Très peu accessible
- > Certaines entreprises peuvent avoir un monopole et garder leur modèle opaque
- > La recherche publique peine à rattraper
- > Enfin... jusqu'à maintenant !



Modèle Open Source



Modèle Privé



Association à but non lucratif



Entreprise à but lucratif



Papier scientifique



Bibliothèque programmation

01

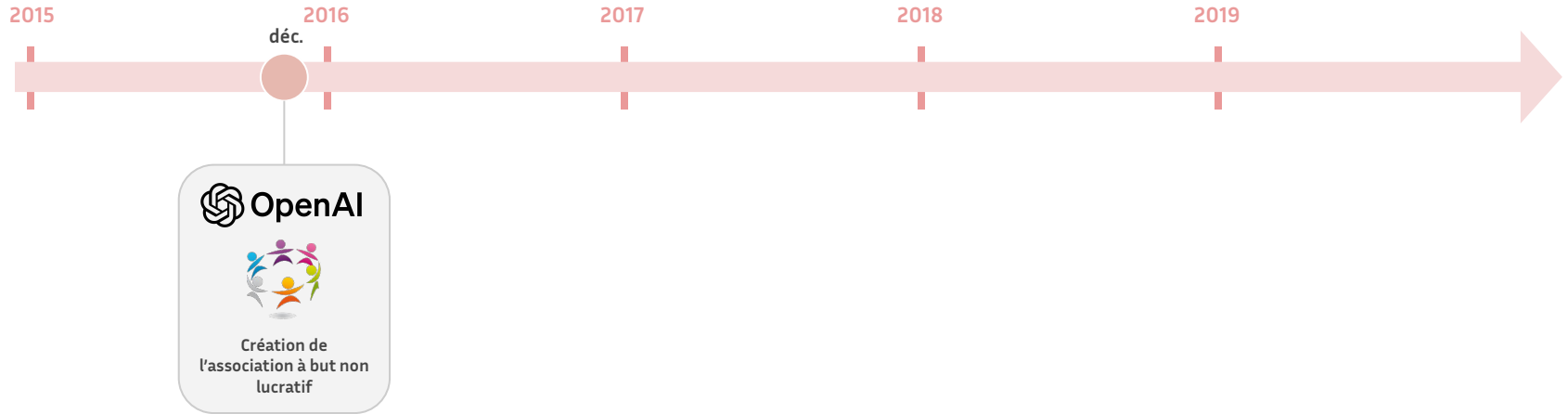
Timeline 2015-2019

Une recherche privée très publique

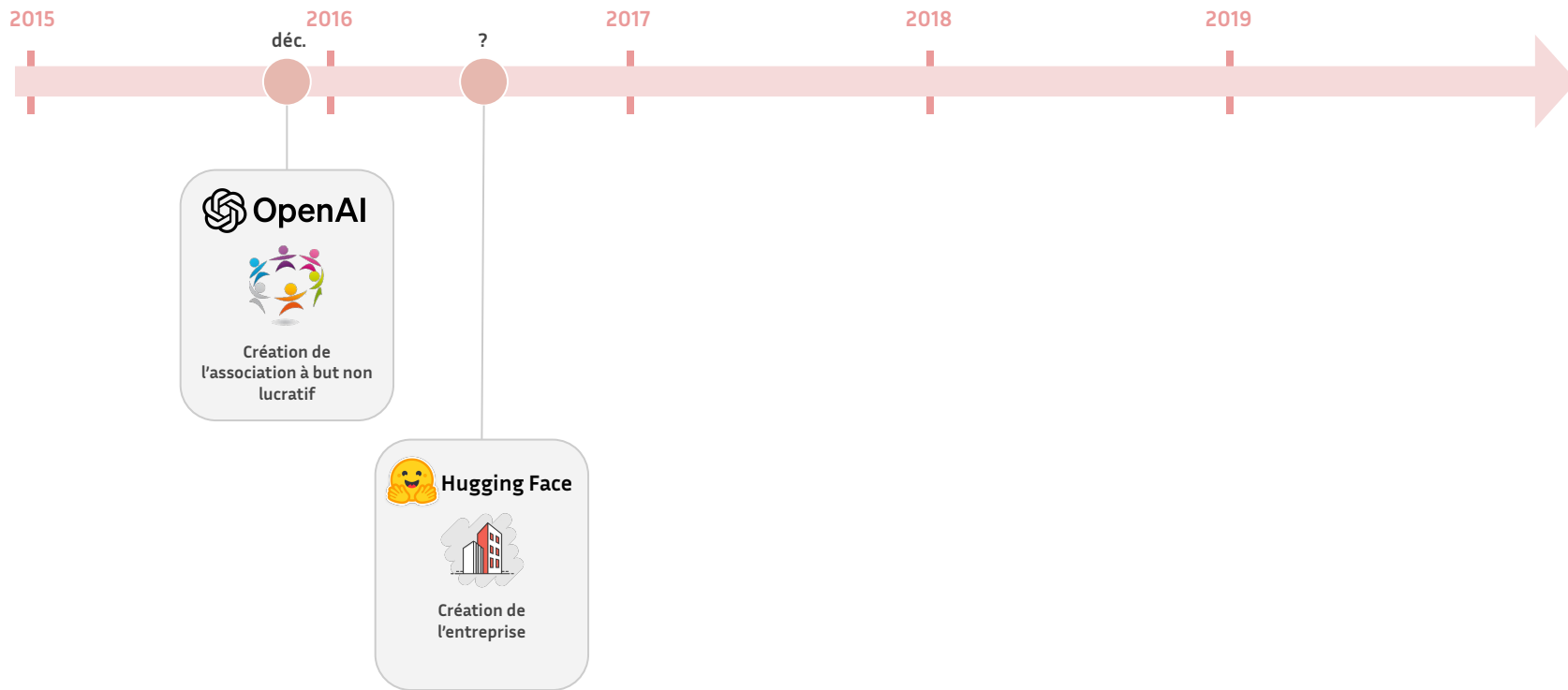
Timeline 2015 - 2019



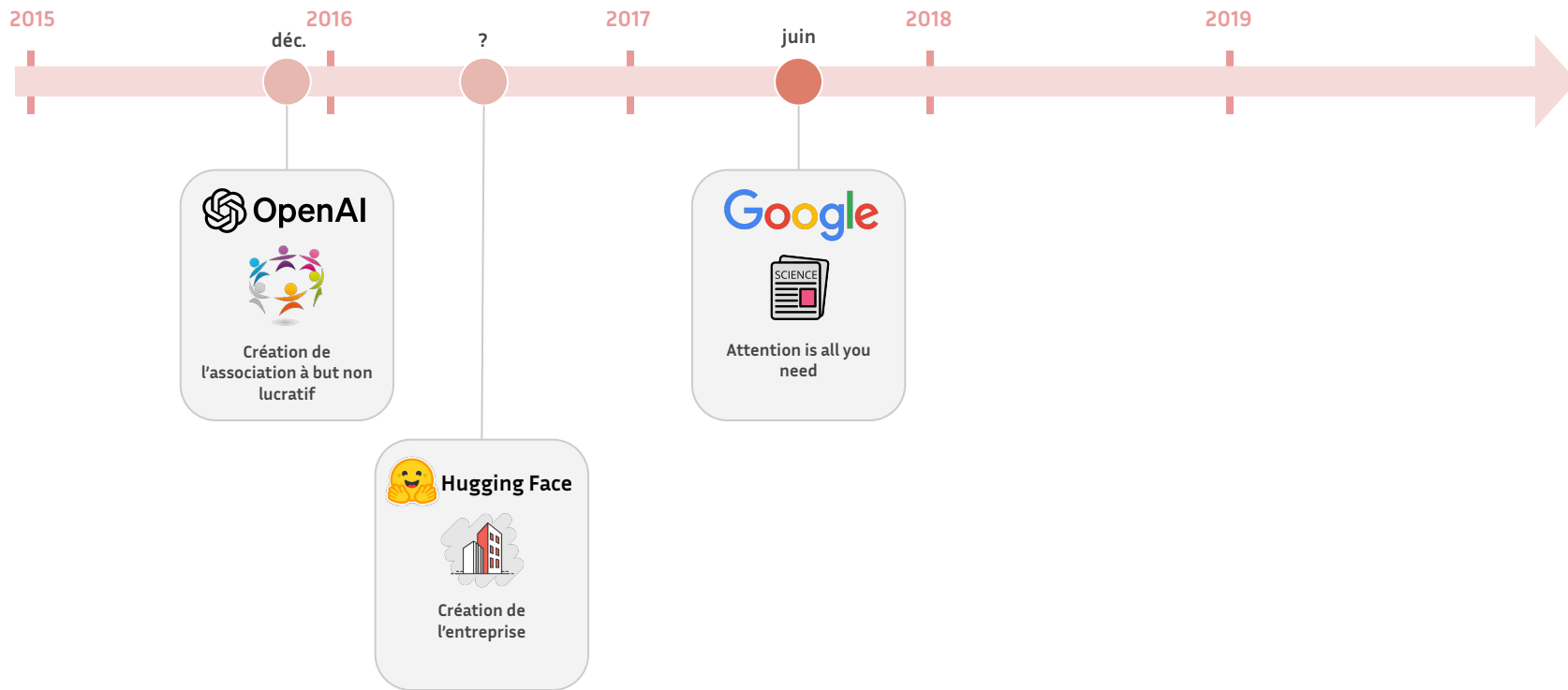
Timeline 2015 - 2019



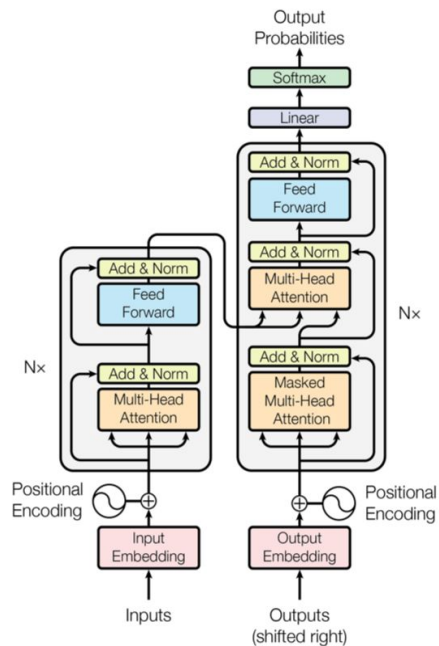
Timeline 2015 - 2019



Timeline 2015 - 2019

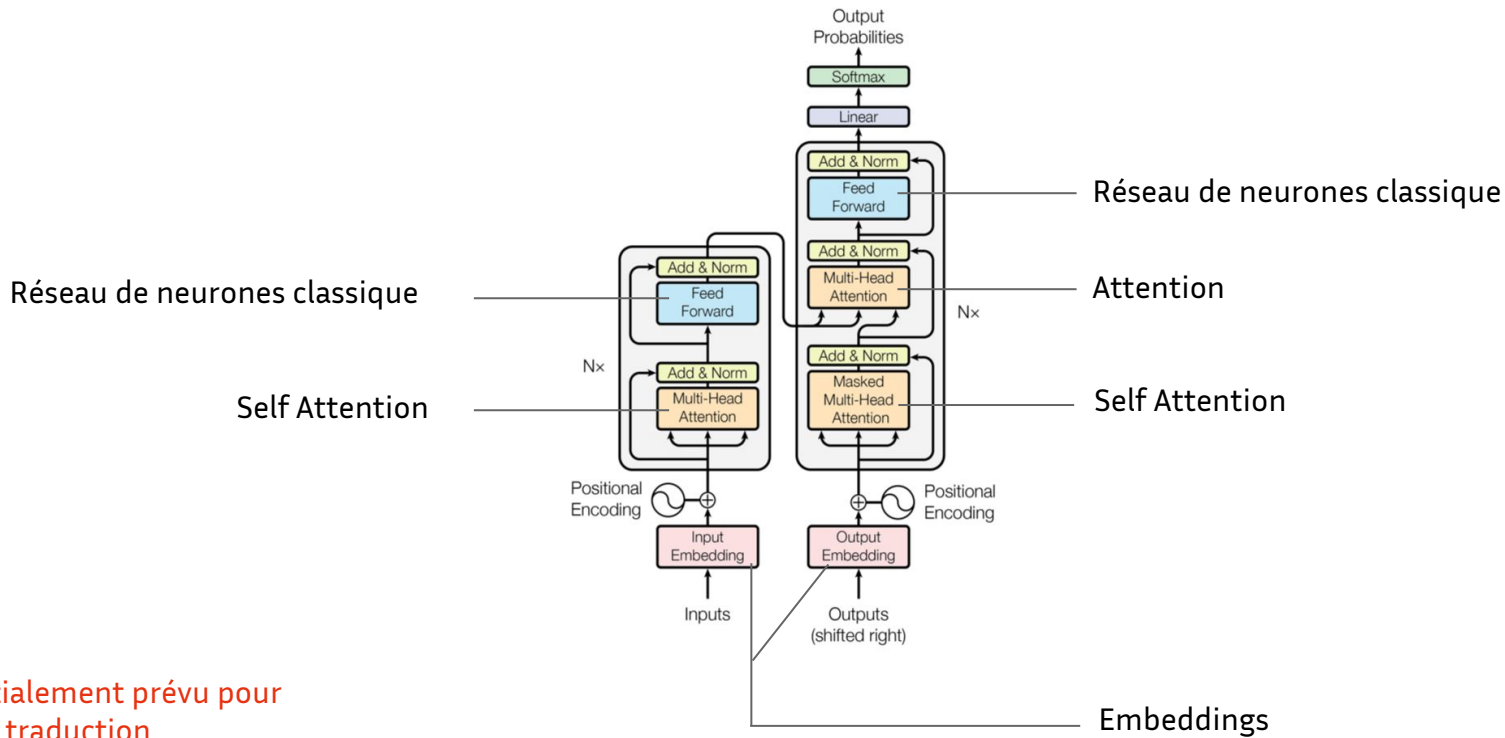


Attention is all you need (2017)



* Initialement prévu pour de la traduction

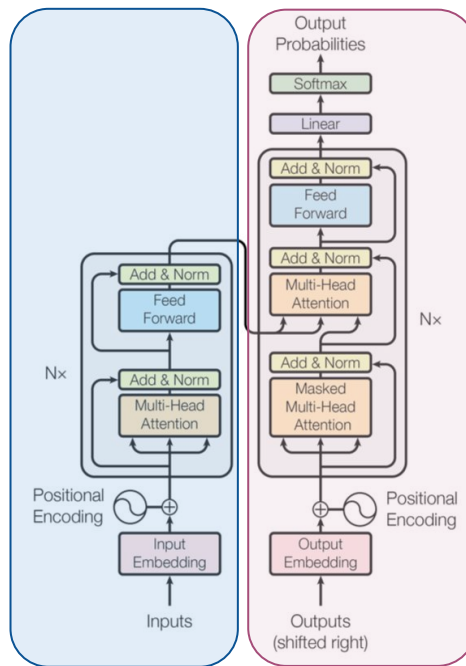
Attention is all you need (2017)



* Initialement prévu pour de la traduction

Attention is all you need (2017)

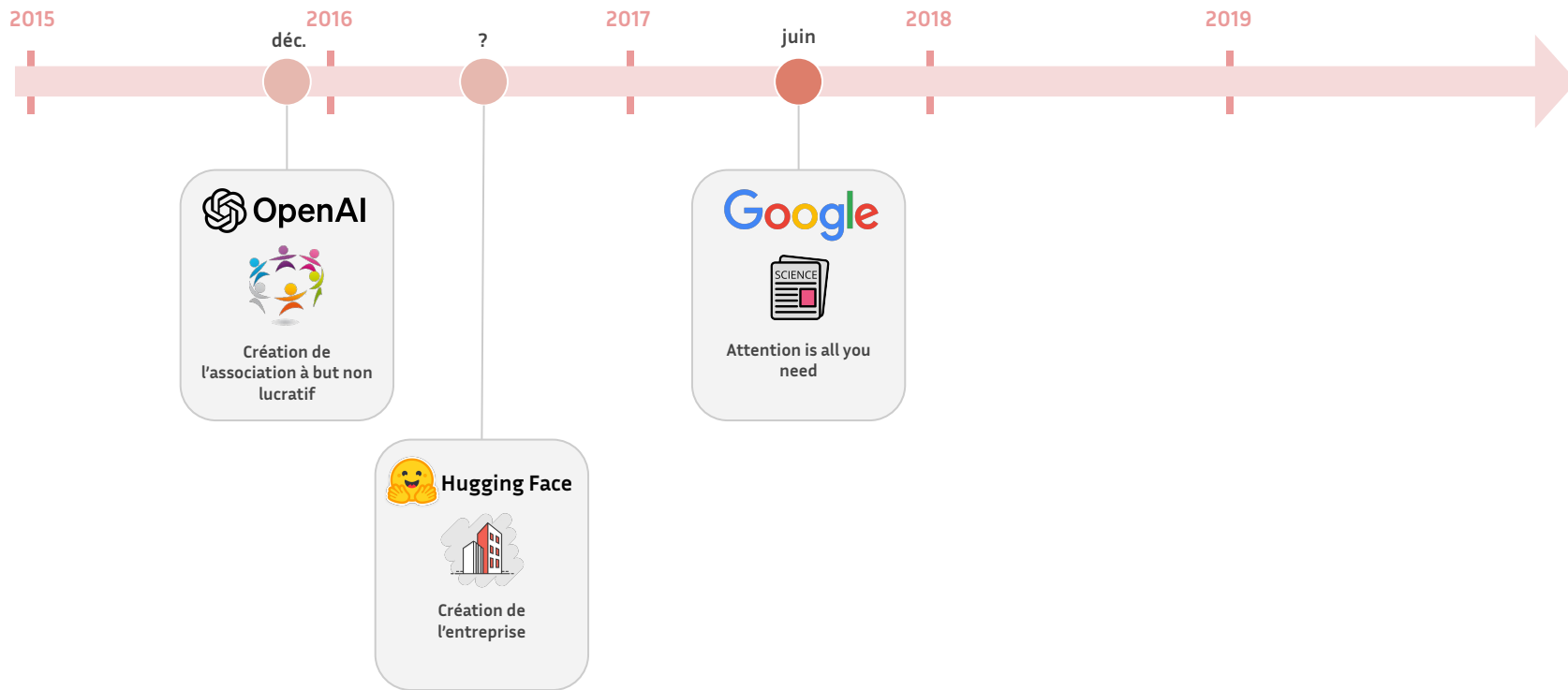
Encodeur



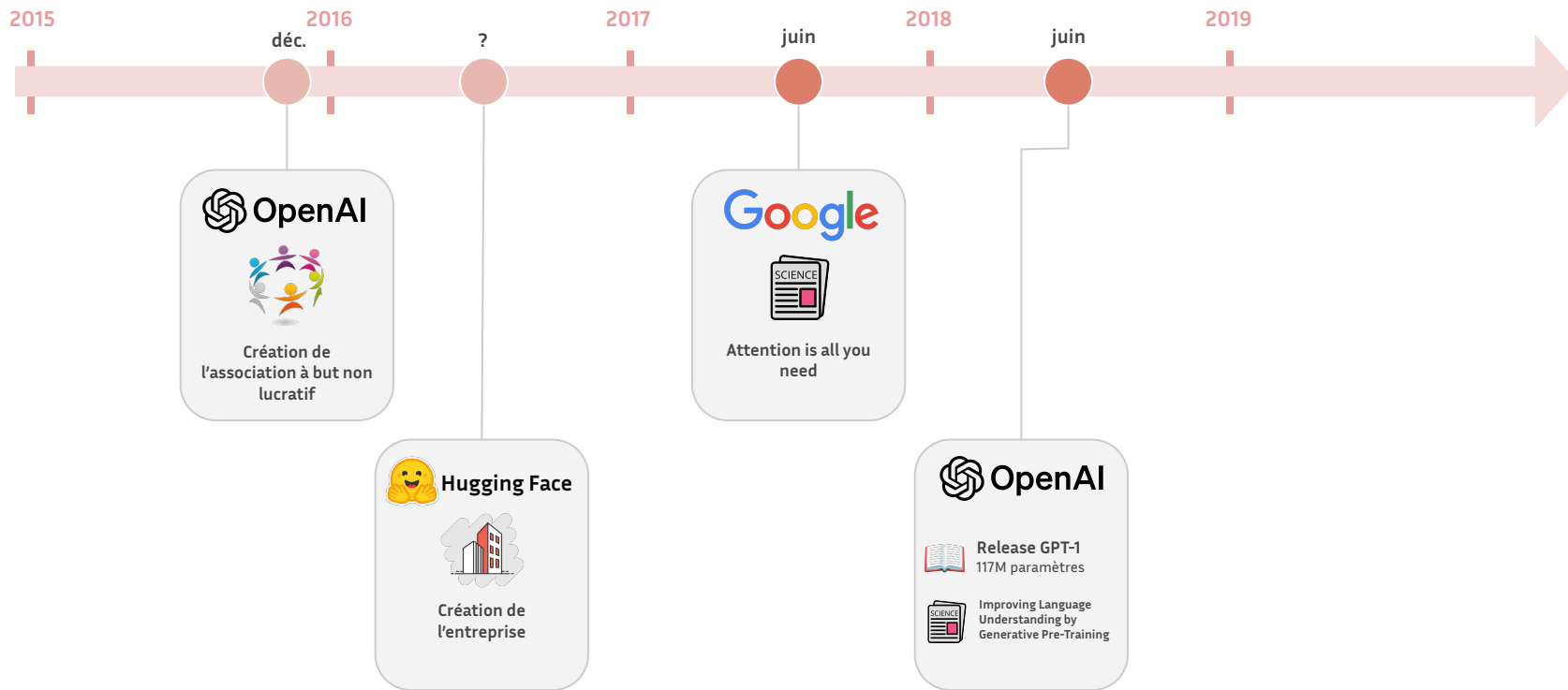
Décodeur

* Initialement prévu pour de la traduction

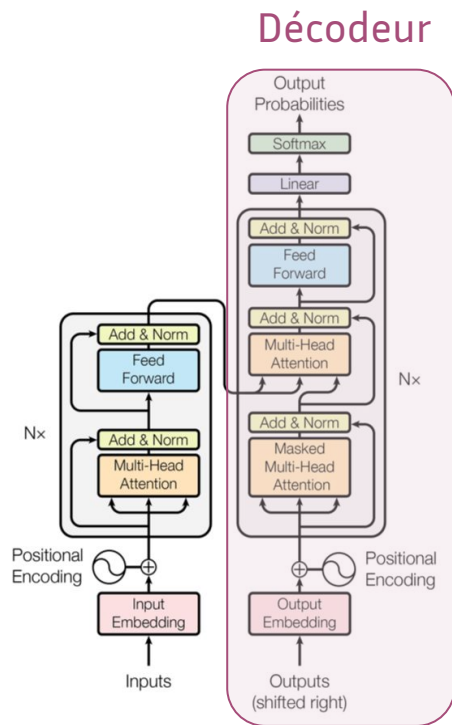
Timeline 2015 - 2019



Timeline 2015 - 2019



GPT-1



12 couches - 768 dims
117M de paramètres
Dataset: BookCorpus - (~3 GB)

What follows is an explanation about Transformers in NLP.

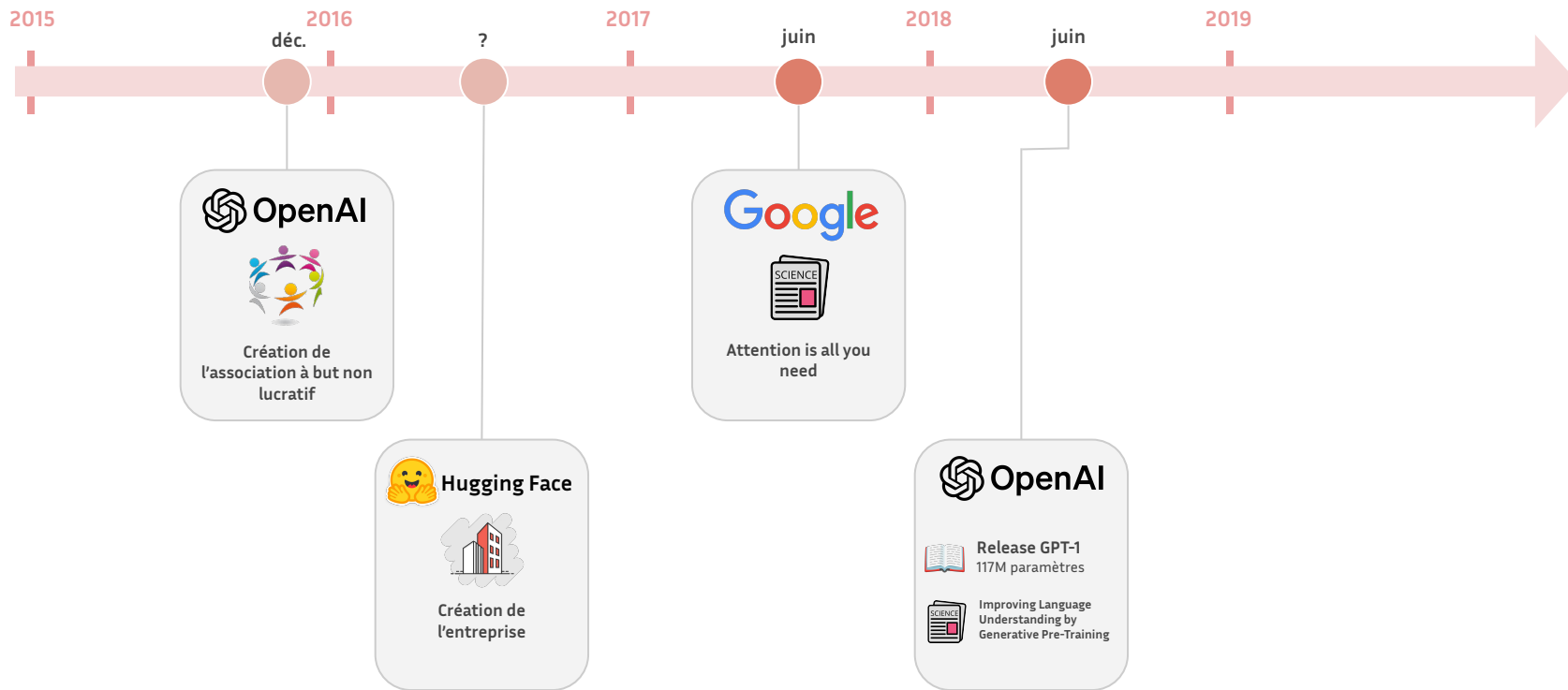
but the scientific community has a strict anti - mormon policy regarding transformers. "

" no, " smith objected. " that's not the point. what the hell was done by the trans - continental railroad to the point that it could be the cause of the cross - referencing the date. "

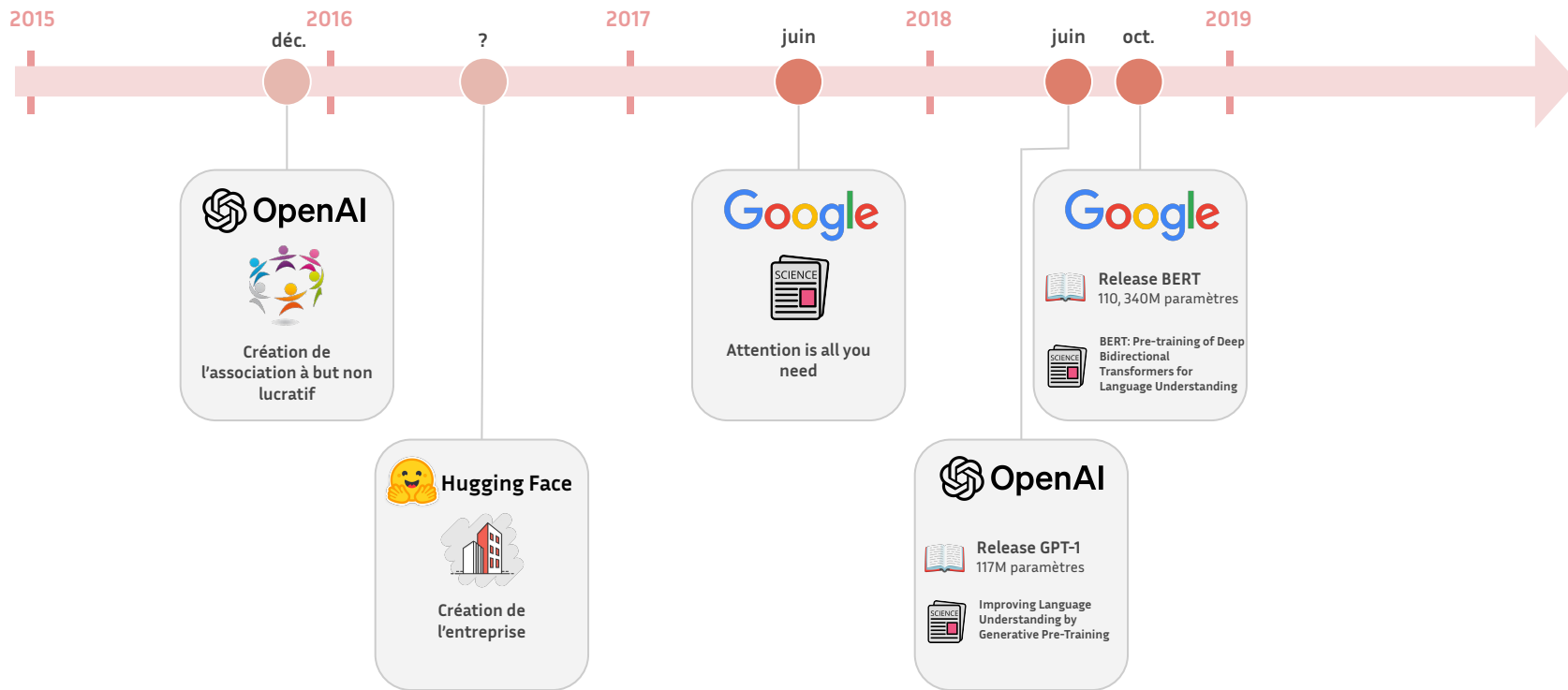
" well, you said you knew about the cross - referencing of the cross? " " i'm the guy who told you, " smith insisted.

" ah. yes... it took me a year to track you down... and you were right. it wasn't just any cross - referencing. there was a very specific crossroad in the legend that goes back

Timeline 2015 - 2019

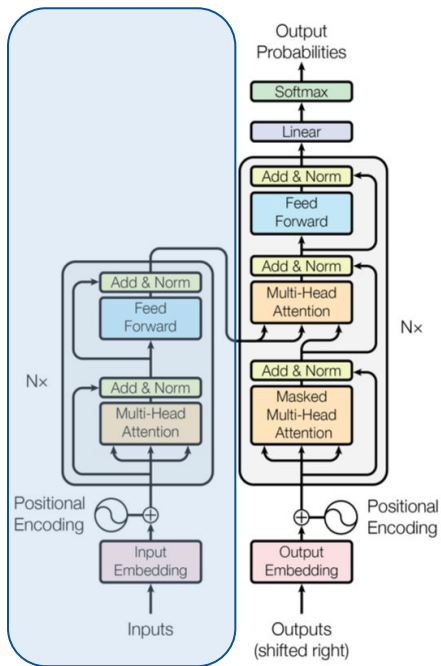


Timeline 2015 - 2019



BERT

Encodeur



12 couches - 768 dims

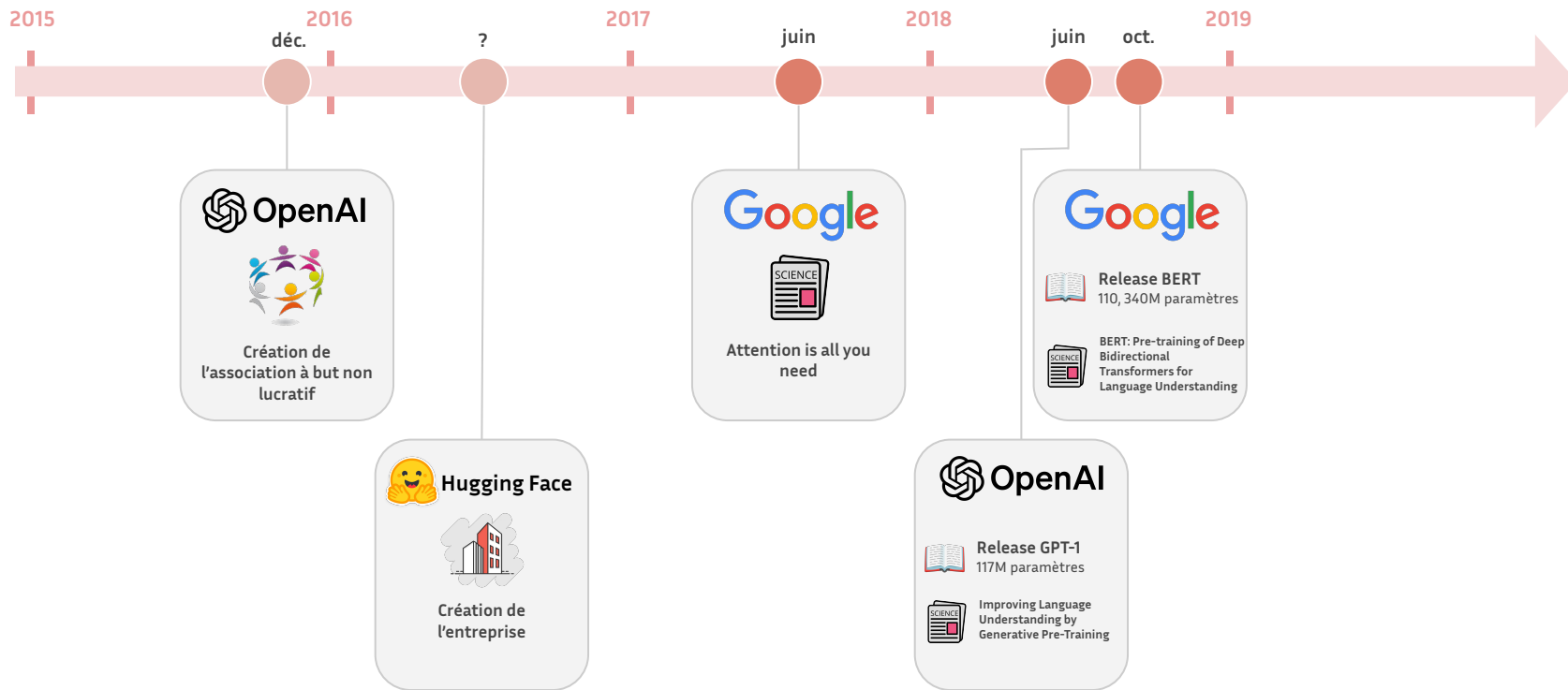
110M, 340M de paramètres

BookCorpus (~3 GB)

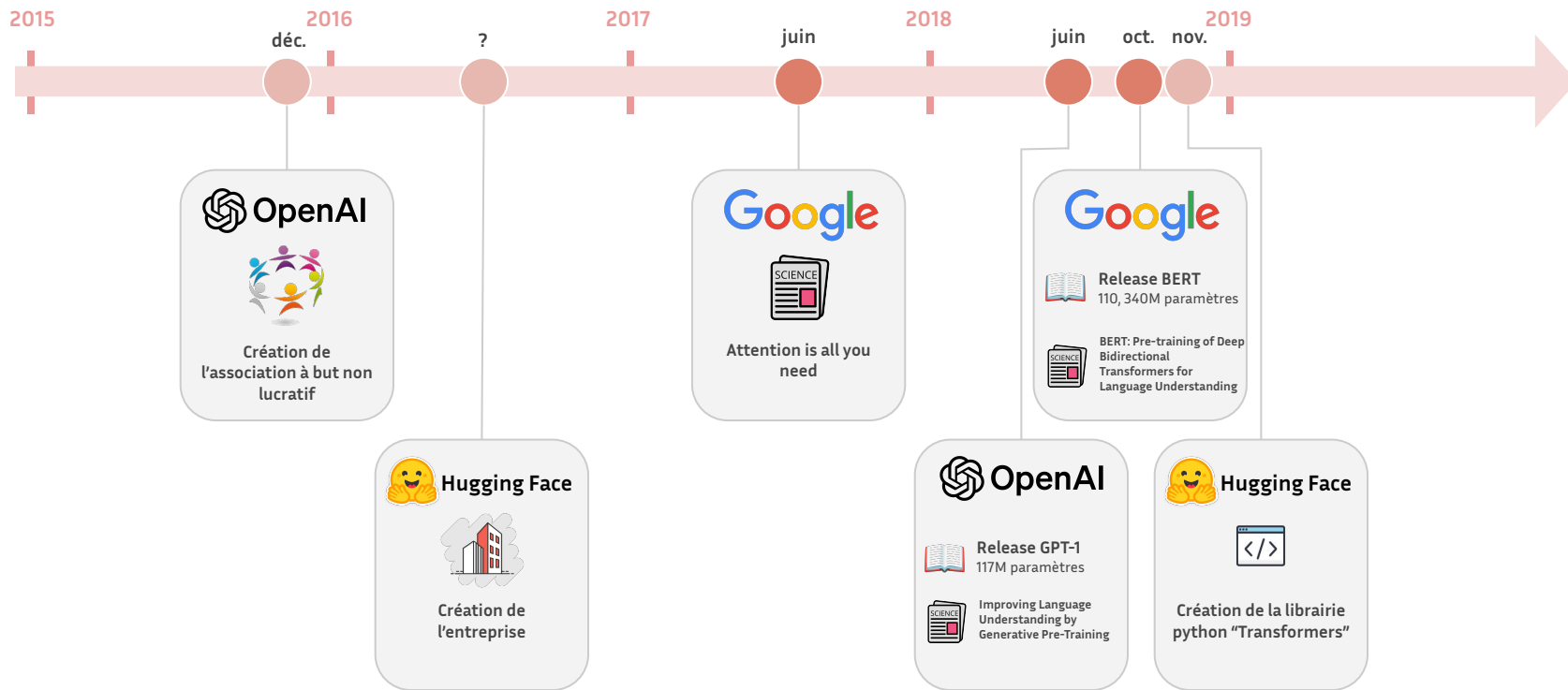
English Wikipedia (~9 GB)

The screenshot shows the BERT application interface. On the left, there is a search bar with the text "Rechercher un mot..." and a dropdown menu "parmi les titres". Below the search bar, there are two sections: "Nuage de mots du voisinage" (Word cloud) and "50 plus proches voisins" (50 closest neighbors). The word cloud contains words like "emploi", "ministre", "travail", "entreprises", "crise", "oui", "choix", "français", "euros", "milliards", "niveau", "président", "chômage", "relance", "évidemment", "question", "salariés", "maire", "croissance", "000", "pays", "économique", "réforme", "bruno", "dette", "temps", "emplois". The "50 plus proches voisins" section lists 10 interview snippets. On the right, there is a document snippet titled "Déclaration de M. Jean Castex, Premier ministre, sur l'annonce d'un plan d'investissement dans les compétences et de réduction des métiers en tension, Châtelleraut le 27 septembre 2021." The document text includes "Intervenants : Jean Castex", "Thématiques : Société, Education - Recherche, Formation, Economie, Emploi - Travail, Politique de l'emploi", "Mots-clés : Politique de l'emploi, Politique de la formation, AFPA, Secteur d'activité, Plan gouvernemental", "Date : 27 août 2021", and a paragraph starting with "Vous le savez, notre pays vit actuellement une situation économique que je qualifierais de paradoxale." Below the document snippet, there is a checkbox "Colorier les points sélectionnés (augmente le temps de calcul)".

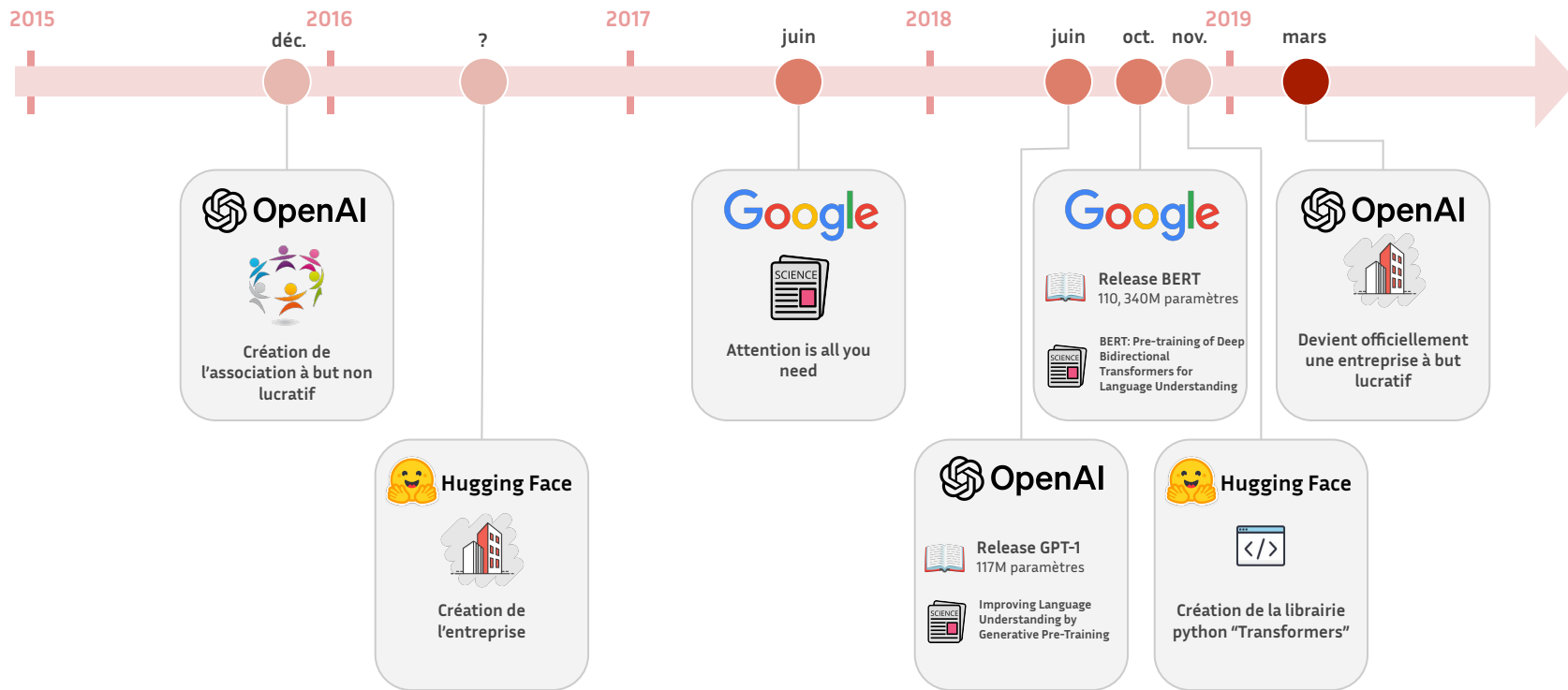
Timeline 2015 - 2019



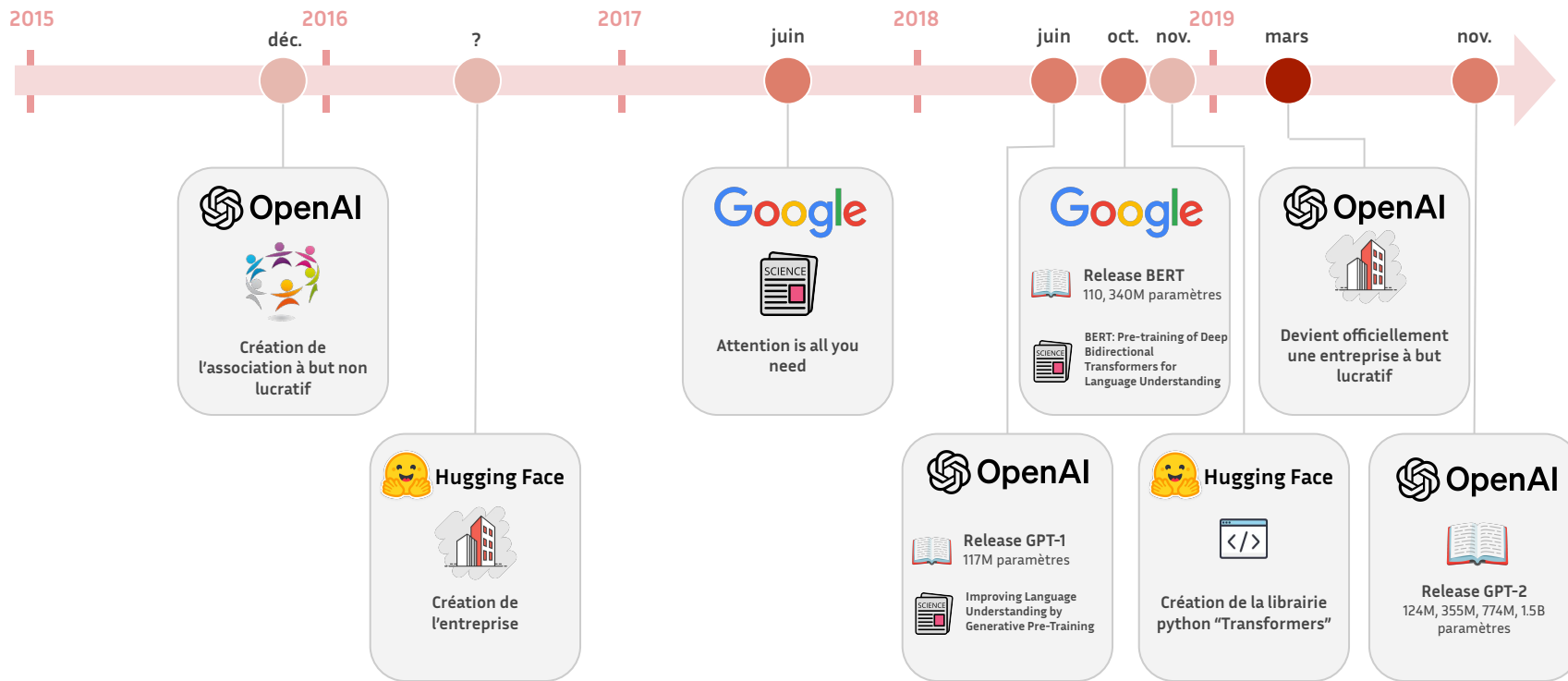
Timeline 2015 - 2019



Timeline 2015 - 2019

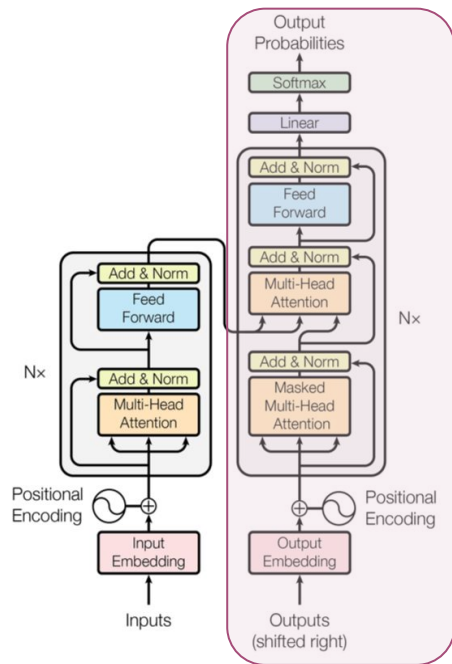


Timeline 2015 - 2019



GPT-2

Décodeur



12 couches - 768 dims
124M, 355M, 774M, 1.5B de paramètres
BookCorpus (~3 GB)
WebText (~40 GB)

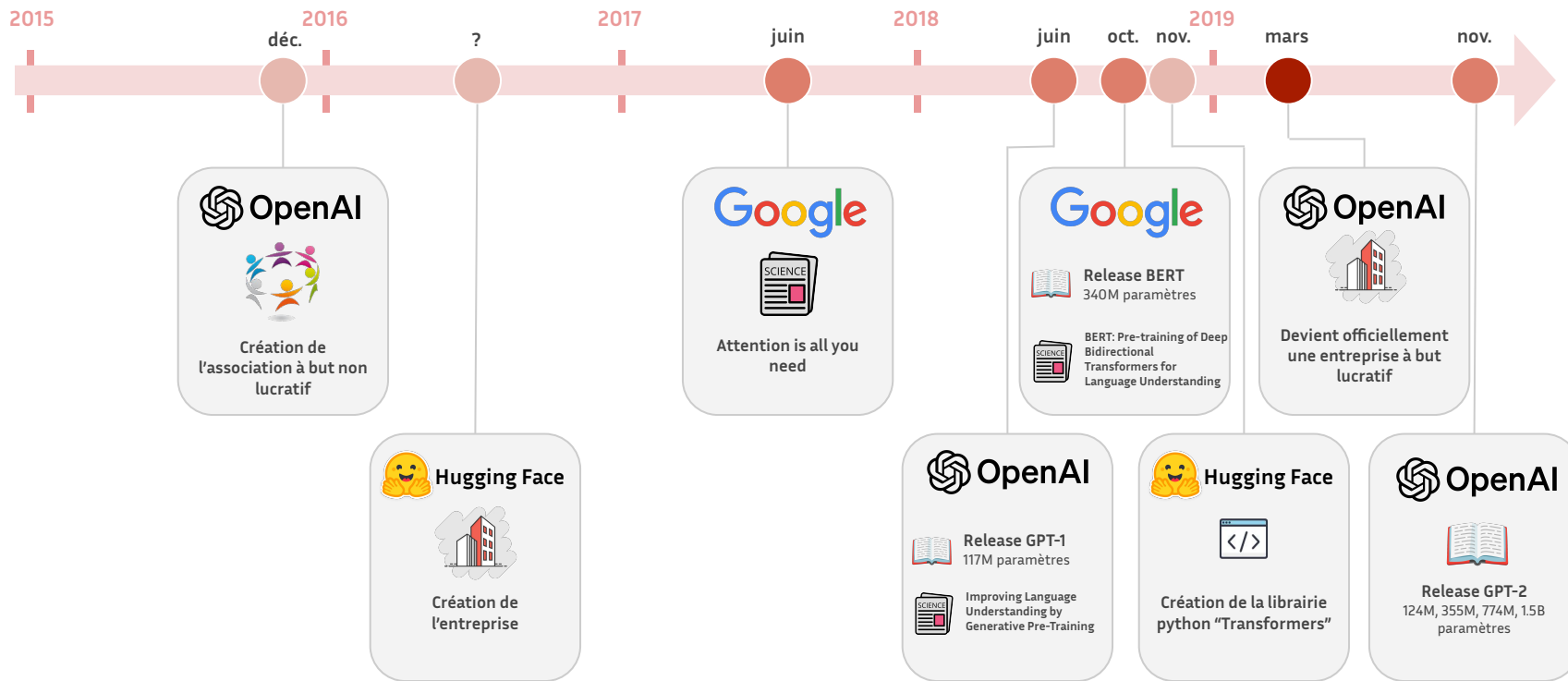
What follows is an explanation about Transformers in NLP.

So what is it that makes a robot move forward?

The way we normally think about it is by observing the movement of a particular object. What is the object? It is called a robot, i.e it is a physical object that in our world can behave in certain specific ways based on its behavior.

In this case, the objects are robots, which are like automatons but have a central controller inside that gives them certain properties such as direction, speed, and position. Robots can be robots that are made of metal, wood, rubber, or other materials which move in a particular way. For instance, some robots make sounds, others can be programmed to move a

Timeline 2015 - 2019



02

Timeline 2020-2022

CloseAI, ou quand le privé ferme ses portes

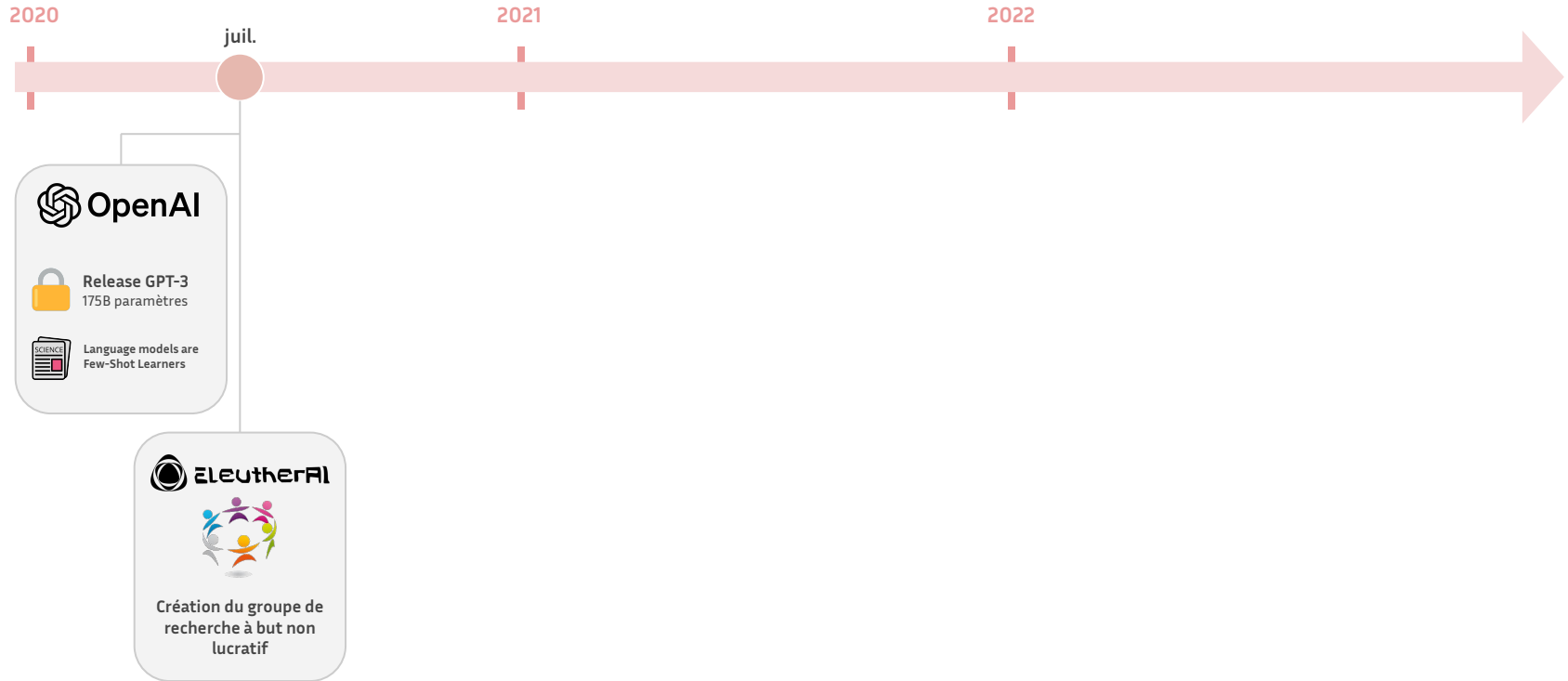
Timeline 2020 - 2022



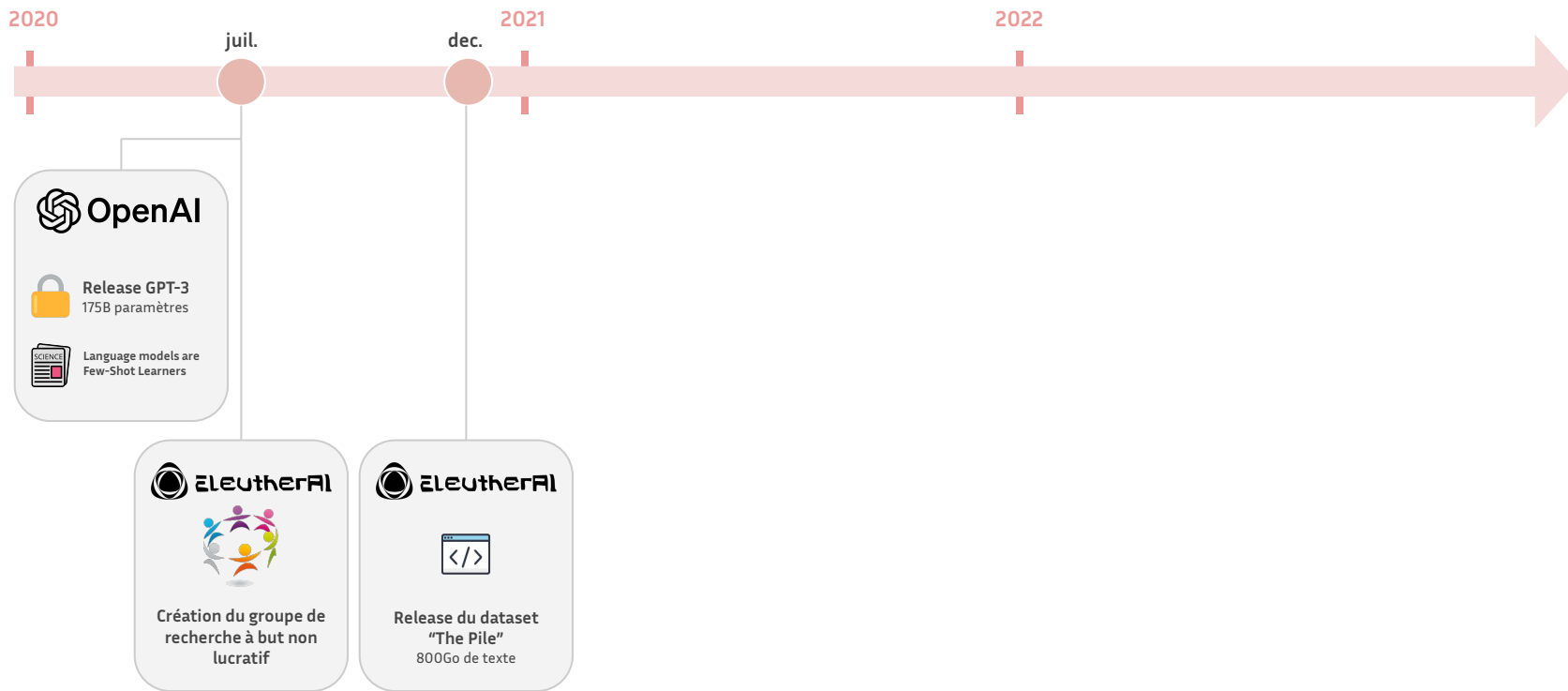
Timeline 2020 - 2022



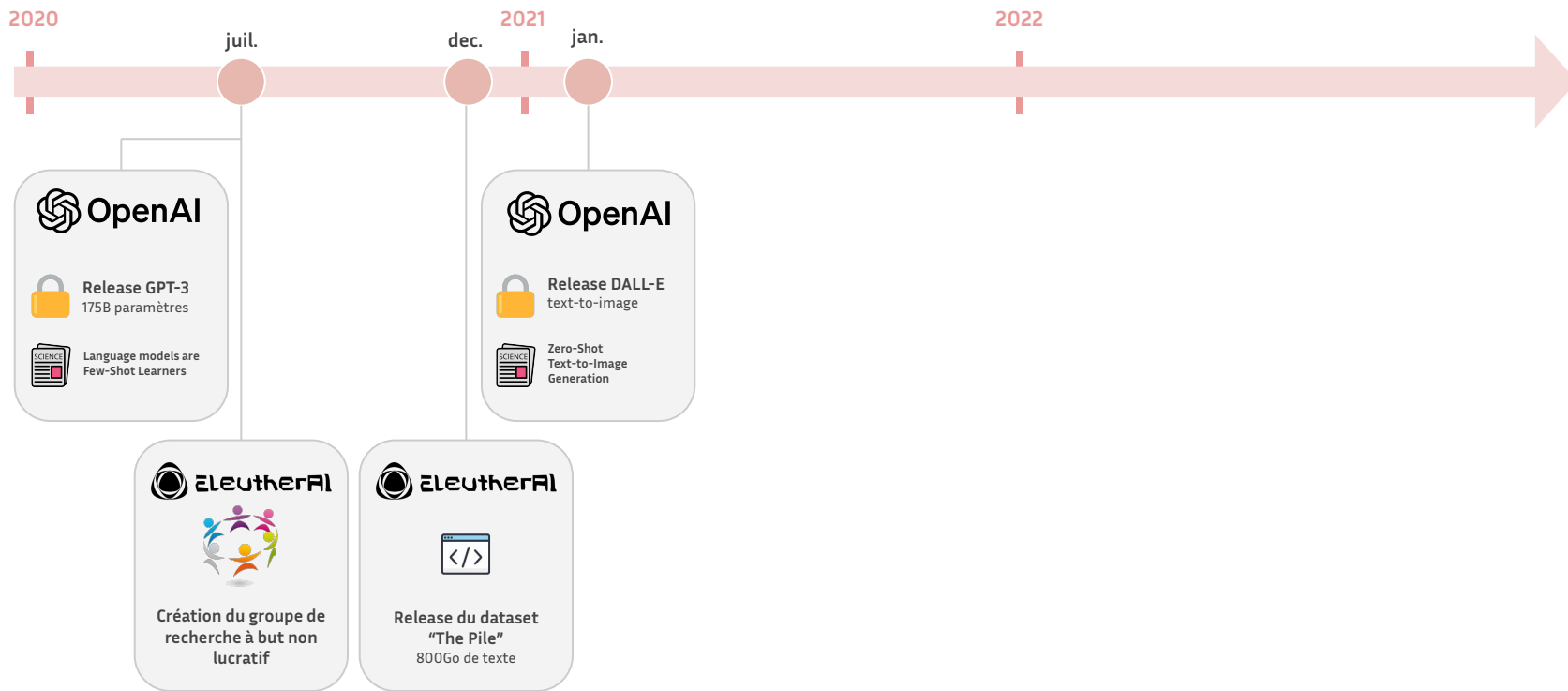
Timeline 2020 - 2022



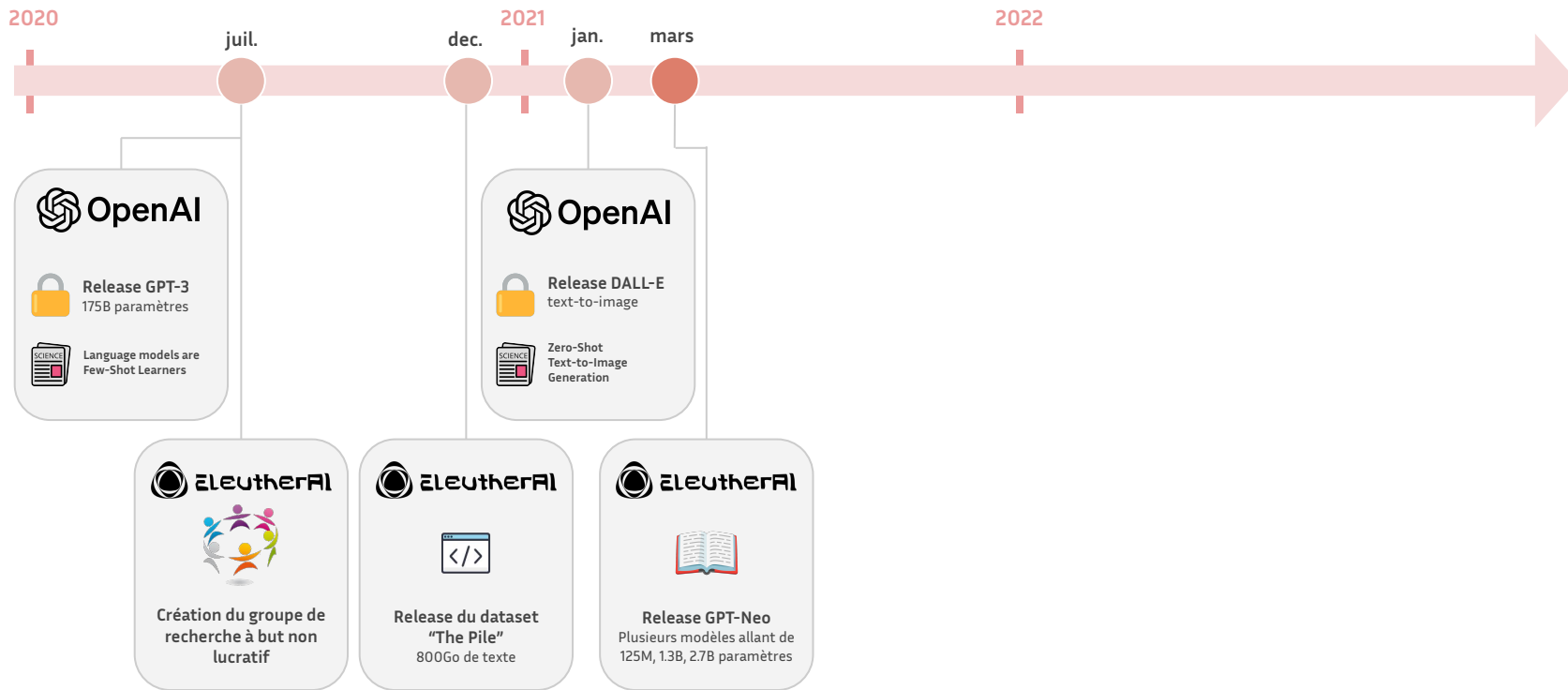
Timeline 2020 - 2022



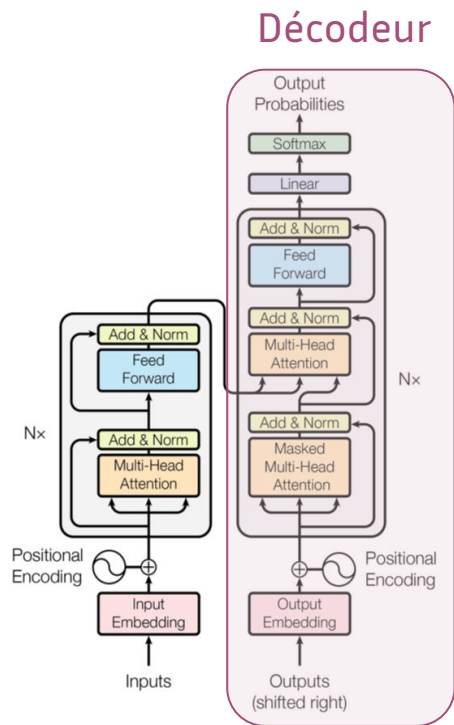
Timeline 2020 - 2022



Timeline 2020 - 2022



GPT-Neo



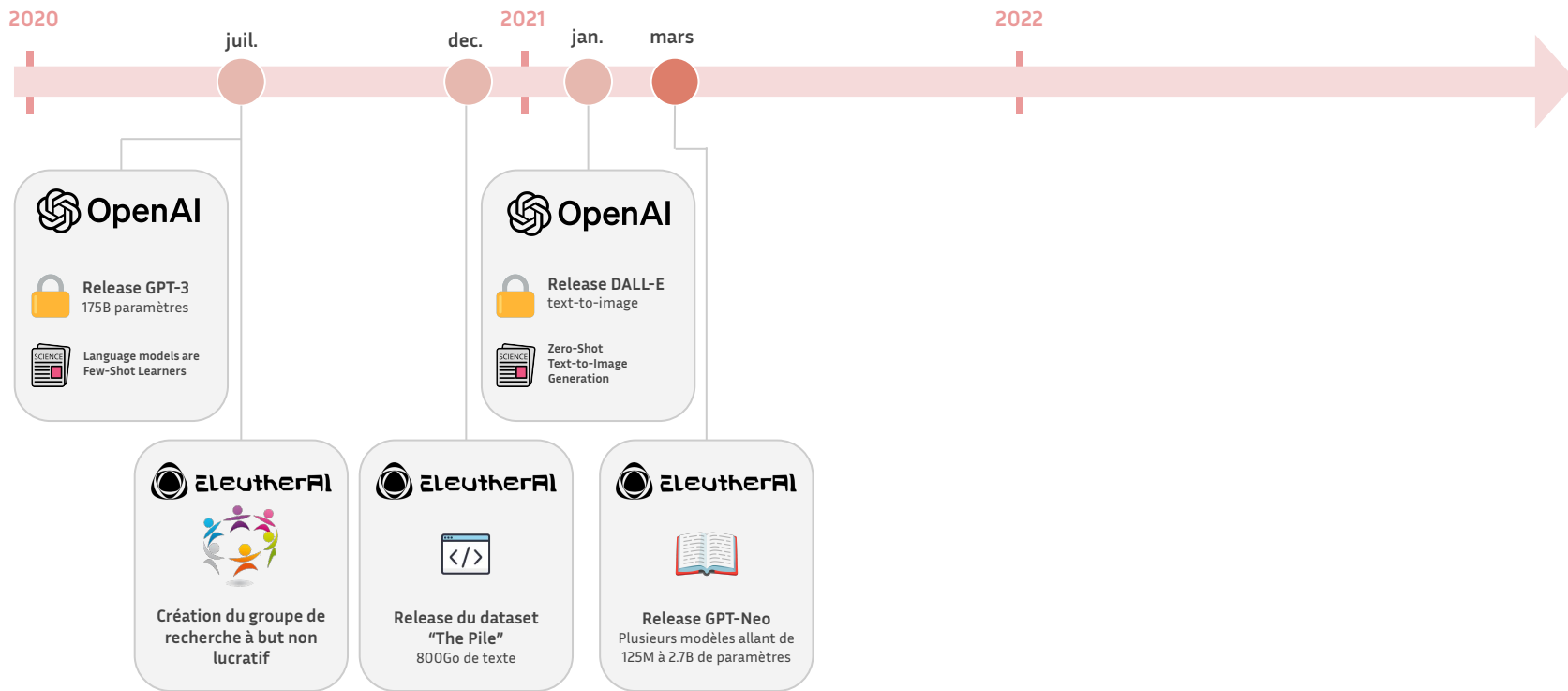
32 couches - 2560 dims
125M, 1.3B, 2.7B de paramètres
The Pile (~800 GB)

What follows is an explanation about Transformers in NLP.

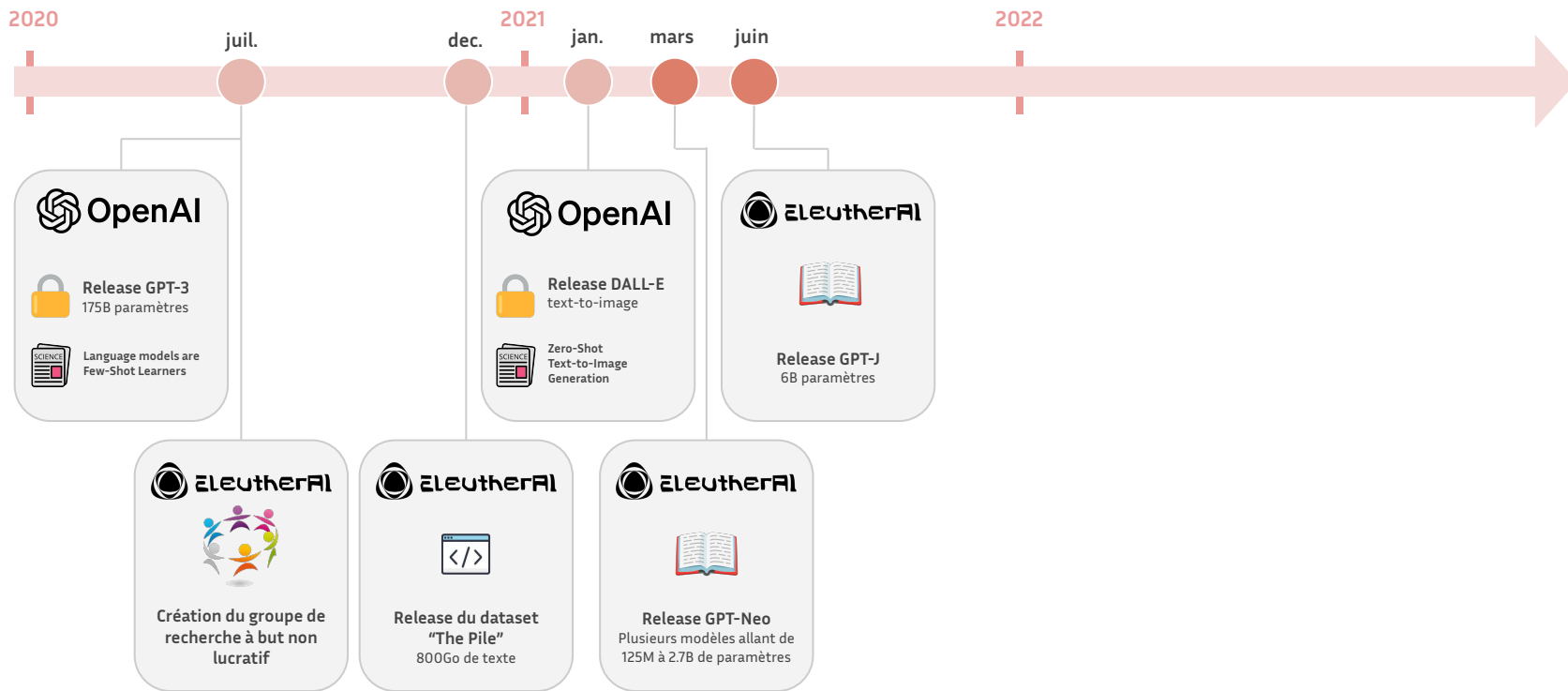
As most NLP algorithms that I have seen so far are rule-based they tend to be difficult to apply to an unknown text. The algorithm that I have developed, however, can be used in a NLP application where the user needs to perform some analysis and then have the results presented in new text. The algorithm is based on the following assumption: There is no context, no information, no context.

To make it a little more clear, I will share my results, what they are saying and why I think they are meaningful. I hope that the following can be used as a reference in such an application. This is a brief, almost naive, explanation of my algorithm. It is clear that it can easily be extended to fit more complex scenarios.

Timeline 2020 - 2022

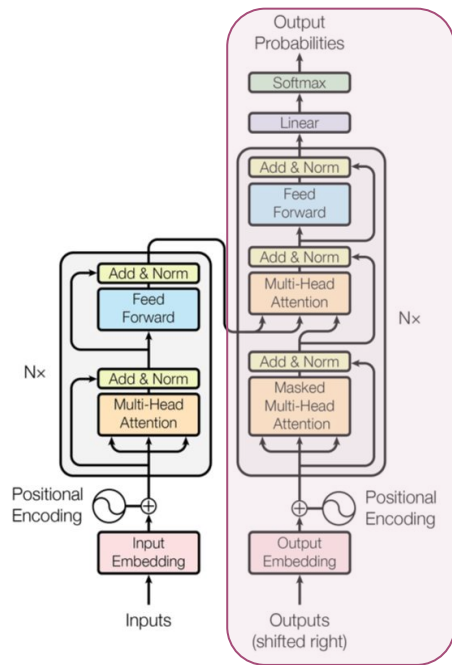


Timeline 2020 - 2022



GPT-J

Décodeur



28 couches - 4096 dims

6B de paramètres

The Pile (~800 GB)

← Prompt

← Texte généré par le modèle

Qu'est ce que la réforme des retraites prévoit?...

Nous avons travaillé à cette question et la réforme des retraites prévoit une réforme de notre système de répartition. Nous prendrons le problème de l'équilibre de notre système de retraite à bras le corps. Nous regarderons notre système de retraite pour en dégager les éléments nécessaires à sa stabilité. Après beaucoup de travaux académiques nous pourrions mettre en place ce rééquilibrage si cela paraît nécessaire. Mais je crois que quand nous parlons retraites, quand nous prenons pour référence les travaux académiques, nous partons du principe qu'un système de retraite est un système de répartition, c'est notre système de répartition qui est au coeur de la société, la rémunération, la pension de retraite et nous devons le maintenir dans la durée.

GPTJ-Overton
Par Yannis Bendi-Ouis
www.naowak.fr



Qu'est ce que la commune de paris de 1871 ?

La Commune de Paris, un mouvement d'inspiration anarchiste et qui durera 72 jours (du 18 mars 1871 au 28 mai 1871), est un épisode de la Révolution française qui a bouleversé la capitale, Paris. Elle est en partie considérée comme le dernier épisode de l'anarchisme français et comme l'un des épisodes les plus sombres de la Commune, avec la répression menée par la Préfecture de Paris, qui ne s'achèvera qu'avec l'armée versaillaise.

Une fois encore, il est difficile de qualifier la Commune de Paris de mouvement anarchiste tant elle est imprégnée de l'idéologie de la bourgeoisie révolutionnaire de la fin du XIXe siècle. Mais dans sa répression, elle a été, avec la Terreur de 1793, le premier exemple de la prise de pouvoir par des minorités armées.

Les débuts de la Commune de Paris
Cet épisode de la Révolution française a lieu dans un contexte particulier : le siège de Paris par les Prussiens, puis par les Allemands. En effet, depuis le traité de Francfort (1871) qui a mis fin à la guerre franco-prussienne, la France a perdu son rang de première puissance militaire et mondiale au profit de l'Allemagne.

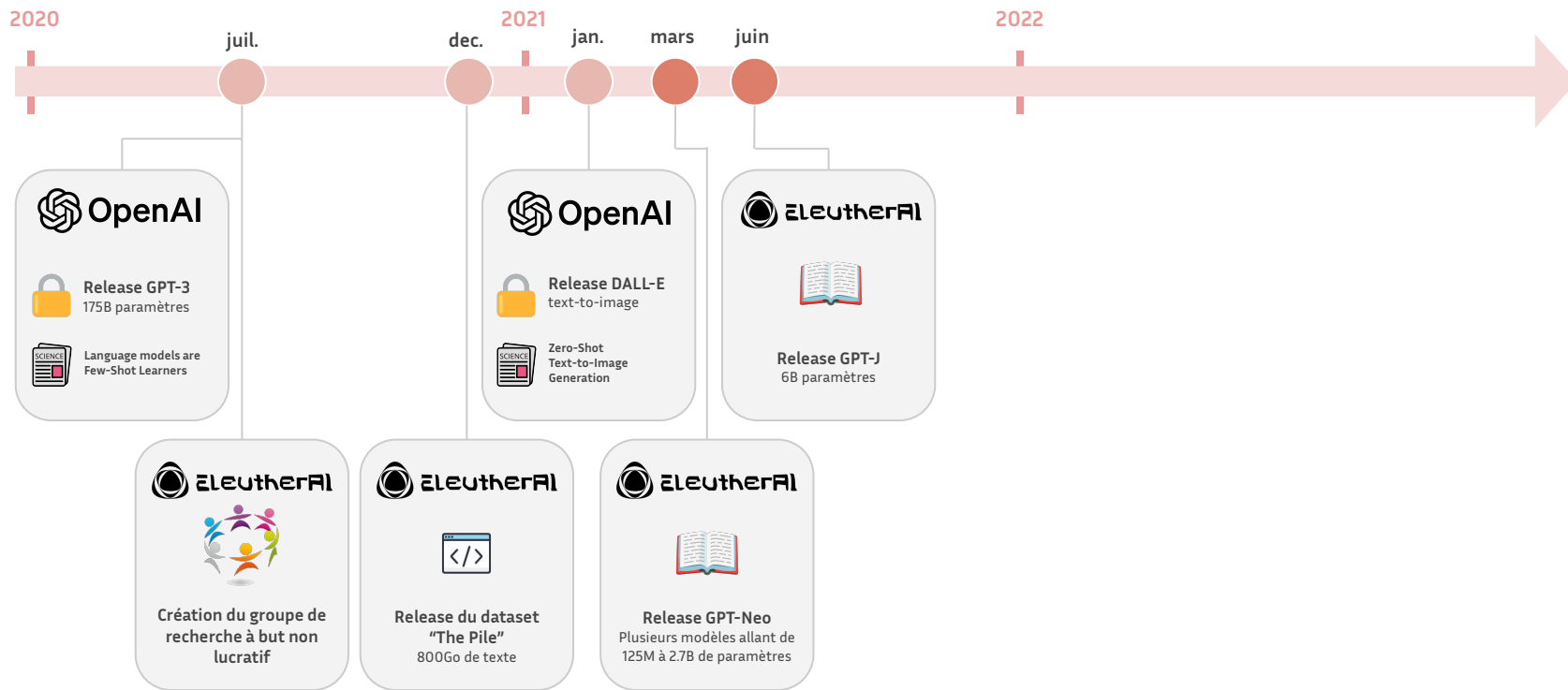
[...]

Les chatbots t'intéressent ?
Contacte moi !

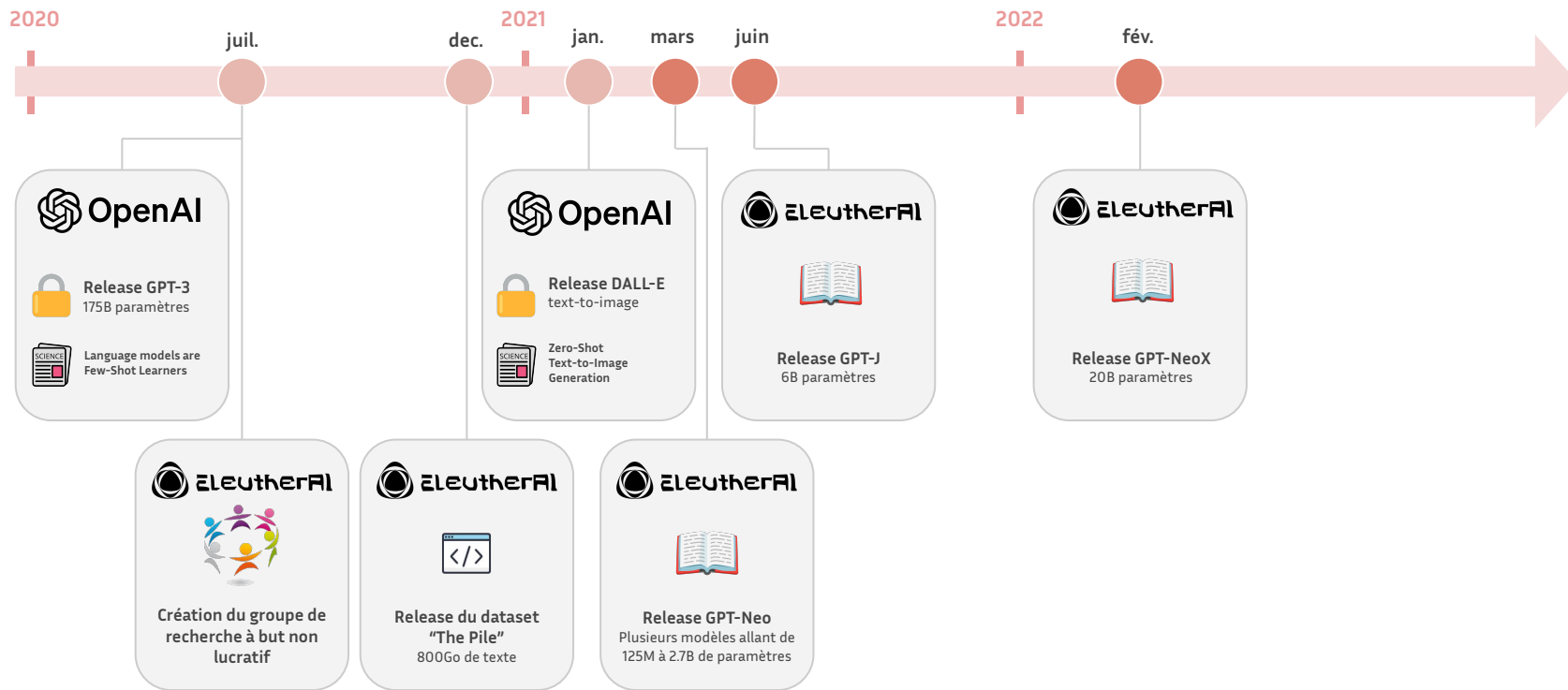


GPTJ-Overton
Par Yannis Bendi-Ouis
www.naowak.fr

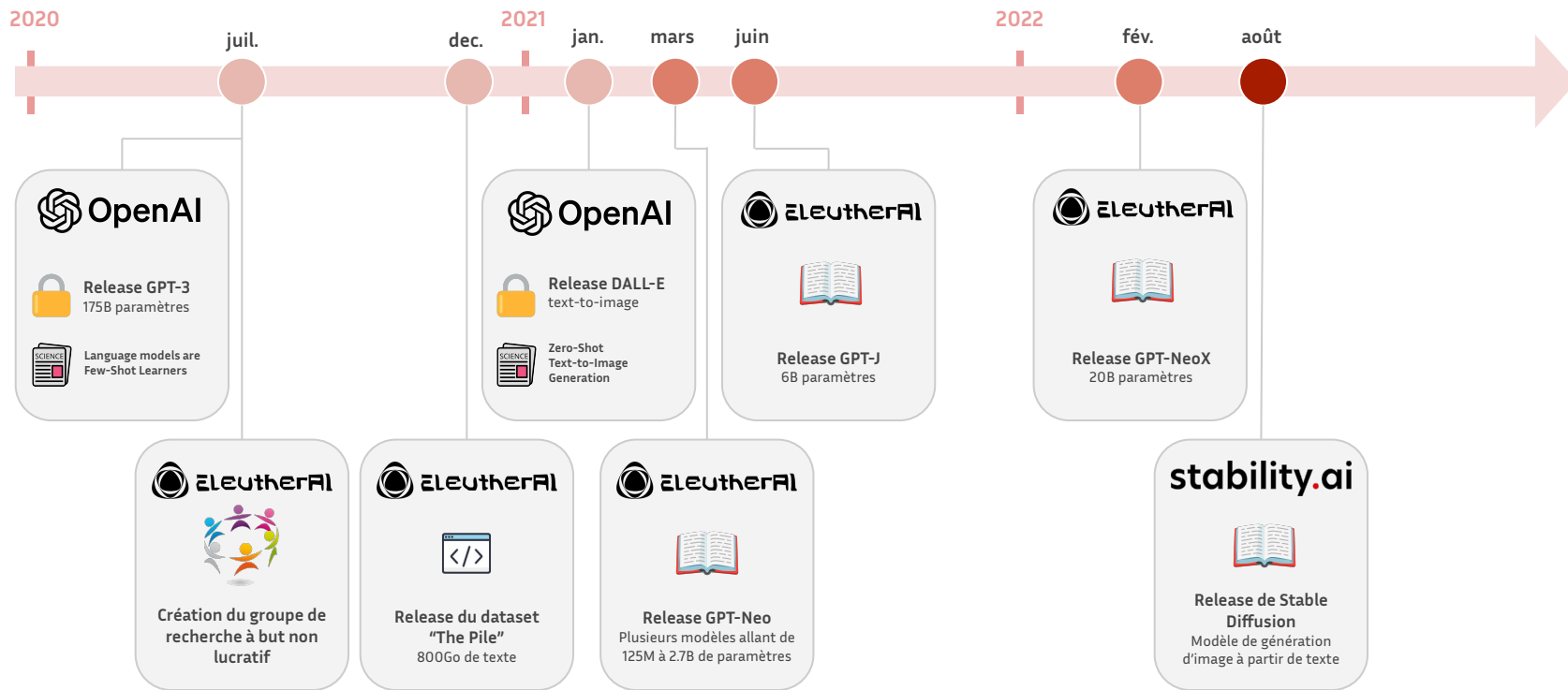
Timeline 2020 - 2022



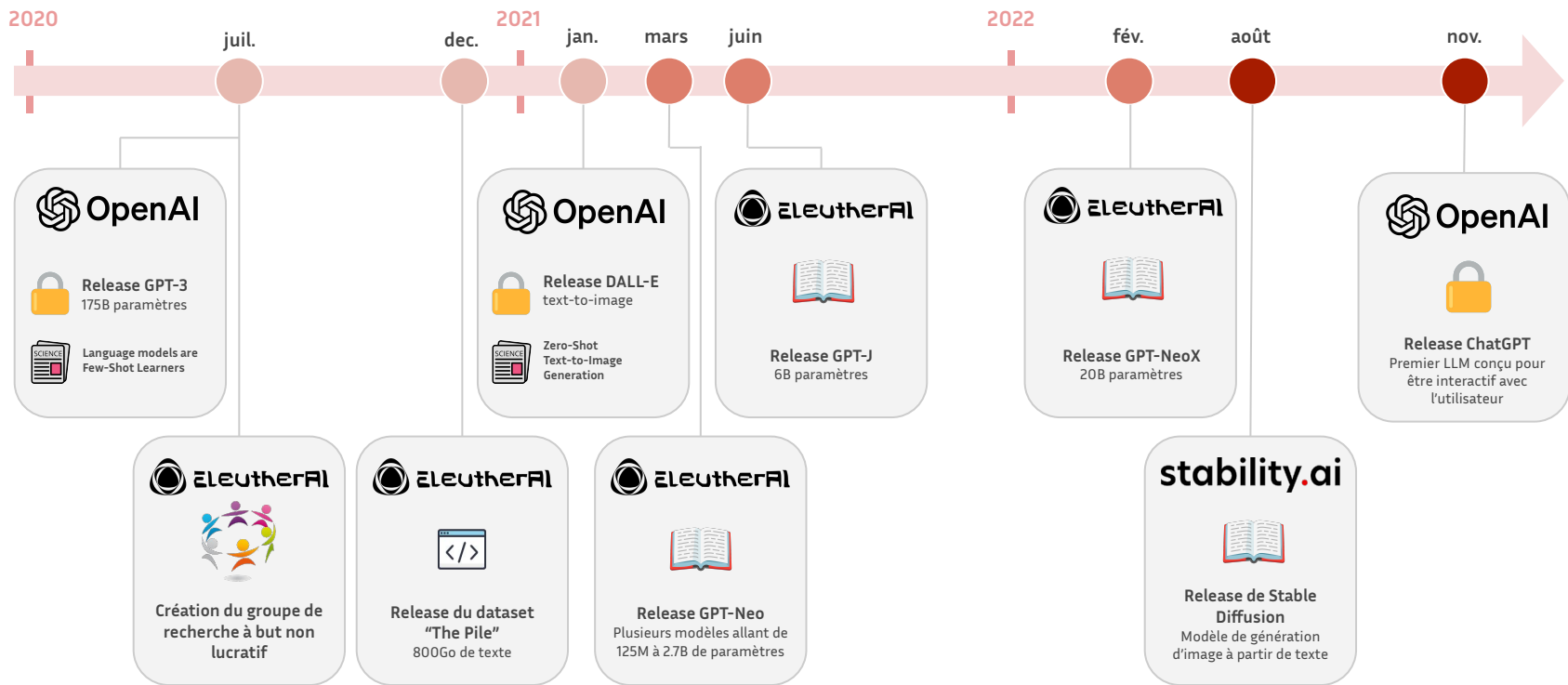
Timeline 2020 - 2022



Timeline 2020 - 2022



Timeline 2020 - 2022

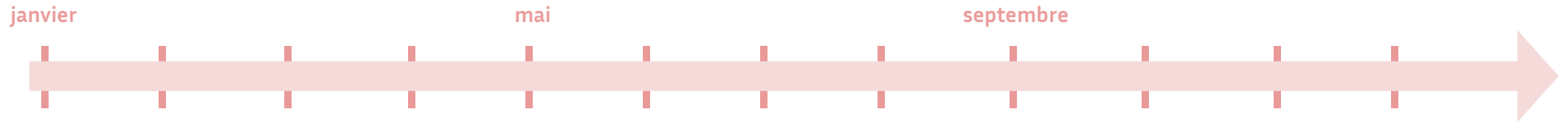


03

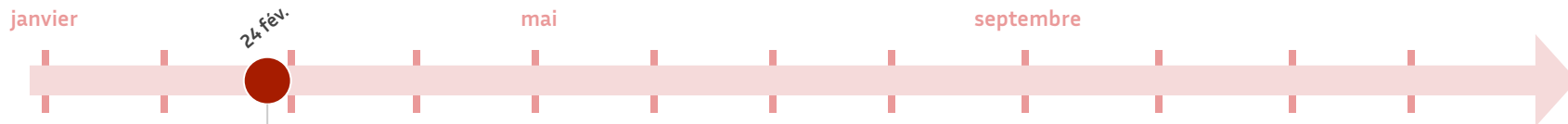
Timeline 2023


Des LLMs pour tous ?


Timeline 2023




Timeline 2023

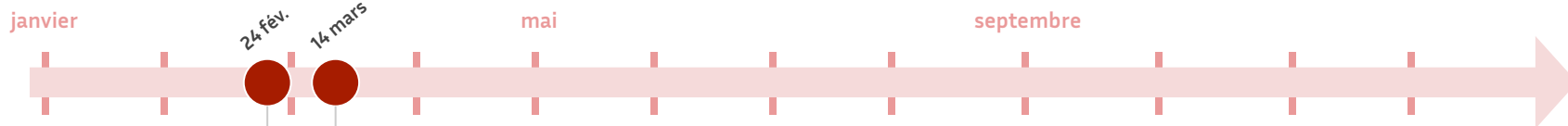



 **Meta**


 **Release LLaMa-1**
7, 13, 33, 70B
paramètres


 **LLaMa: Open and Efficient
Foundation Language
Models**


Timeline 2023




 **Meta**

 **Release LLaMa-1**
7, 13, 33, 70B
paramètres

 **LLaMa: Open and Efficient
Foundation Language
Models**

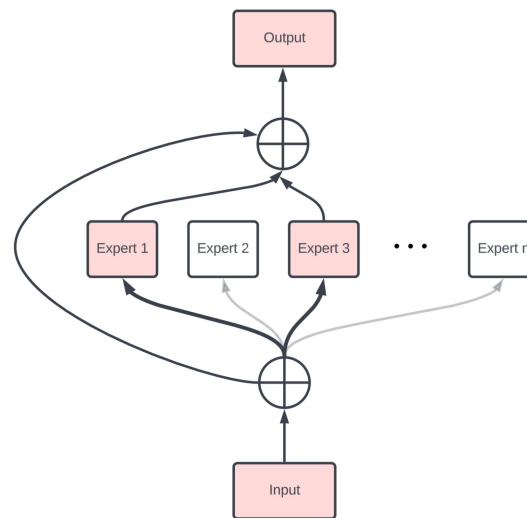
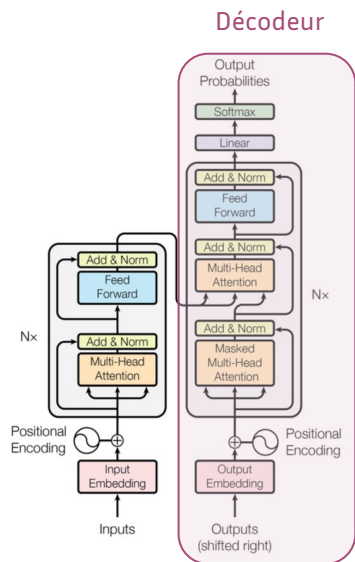
 **OpenAI**



Release GPT-4
16x111B ou 8x220B MoE -
1800B paramètres

GPT-4

? couches - ? dims
~1800B de paramètres
Mixture of Expert 8x220B ou 16x111B



Dégradation des modèles d'OpenAI ?

De moins en moins bon ?



Dégradation des modèles d'OpenAI ?

De moins en moins bon ?



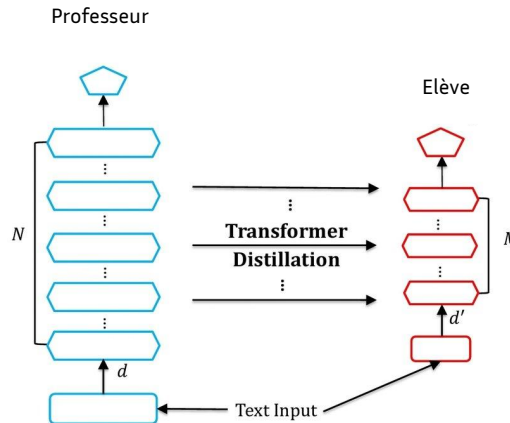
> Finetuning pour supprimer certains comportements

Dégradation des modèles d'OpenAI ?

De moins en moins bon ?



- > Finetuning pour supprimer certains comportement
- > Distillation dans de plus petit modèle

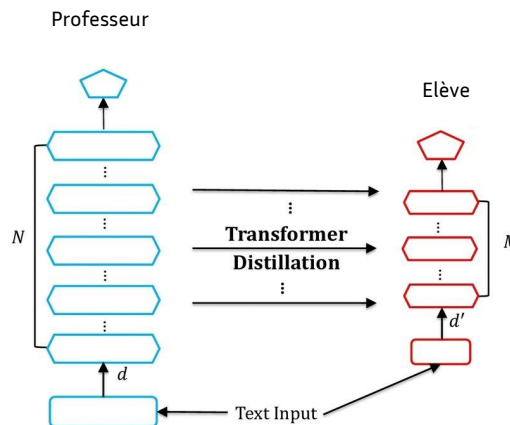


Dégradation des modèles d'OpenAI ?

De moins en moins bon ?

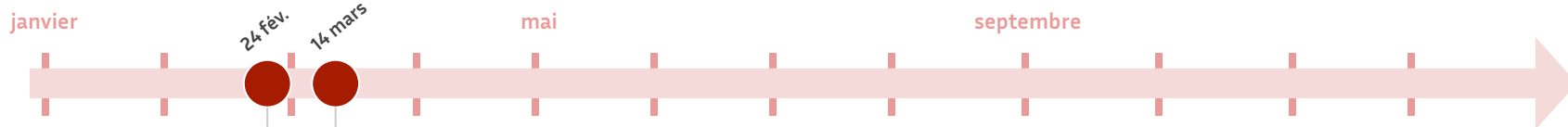



- > Finetuning pour supprimer certains comportement
- > Distillation dans de plus petit modèle





System description		
System	Model	#P
T5	t5-large	770M
CodeT5	codet5-large	770M
GPT-3	text-davinci-003	175B
ChatGPT	gpt-3.5-turbo	20B
StarCoder	starcoder	15.5B
CodeT5+	codet5p-16b	16B
CodeGen	codegen-350m	350M
Diffusion-LM	Custom	50M
GENIE	Custom	93M
CODEFUSION	Custom	75M


Timeline 2023




 **Meta**

 **Release LLaMa-1**
7, 13, 33, 70B
paramètres

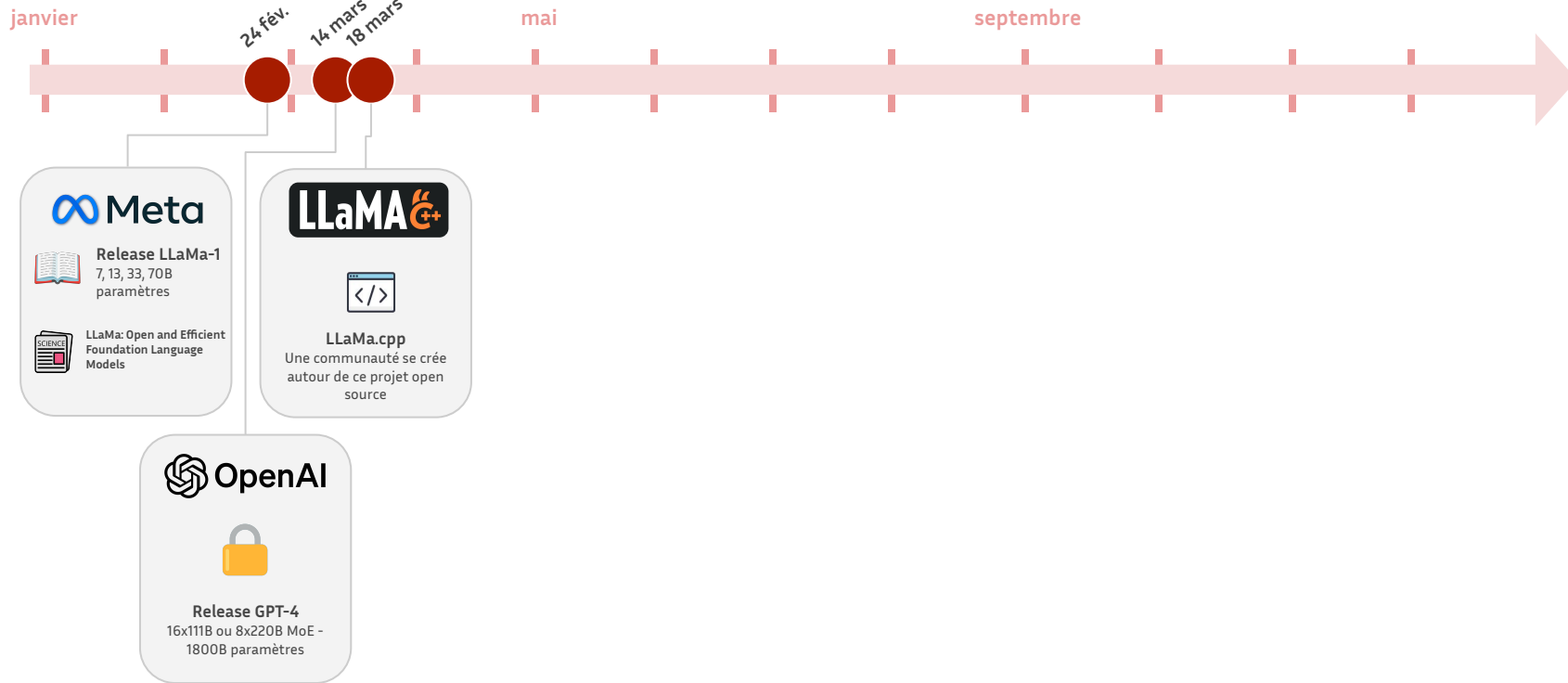
 **LLaMa: Open and Efficient
Foundation Language
Models**

 **OpenAI**

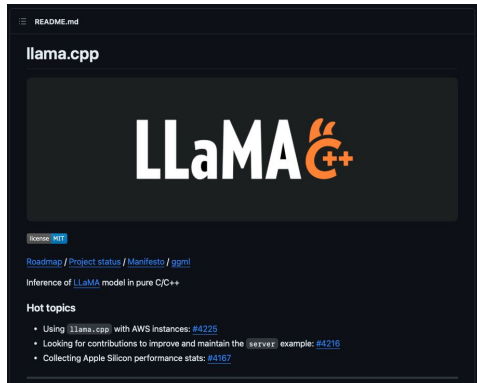


Release GPT-4
16x111B ou 8x220B MoE -
1800B paramètres

Timeline 2023



LLaMa.cpp



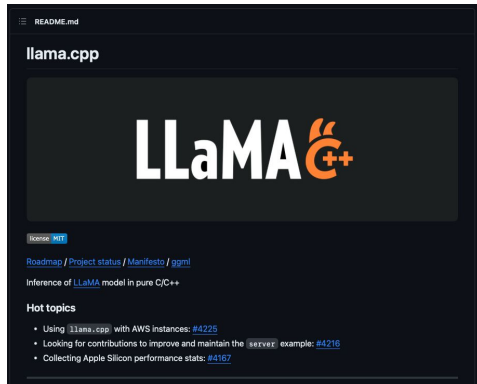
LocalLlama

r/LocalLLaMA

Rejoindre



LLaMa.cpp



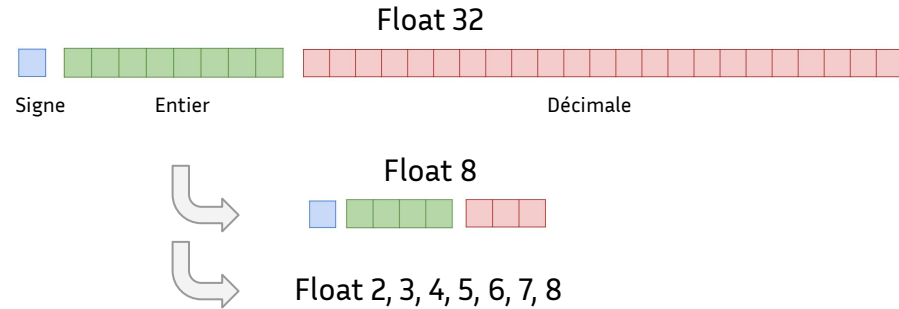
LocalLlama

r/LocalLLaMA

Rejointe



Quantification



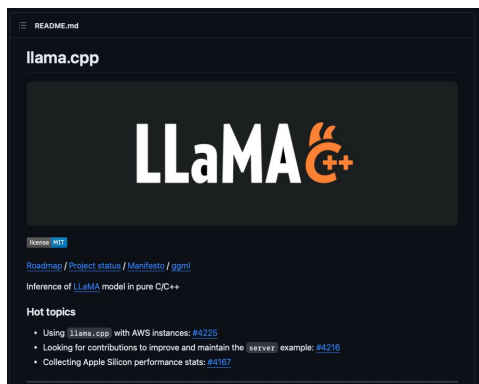
> Taille d'un modèle quantifié

7B ~ 30 Go → 4 bits → ~ 4 Go

> RAM / VRAM nécessaire

7B ~ 30 Go → 4 bits → ~ 7 Go

LLaMa.cpp



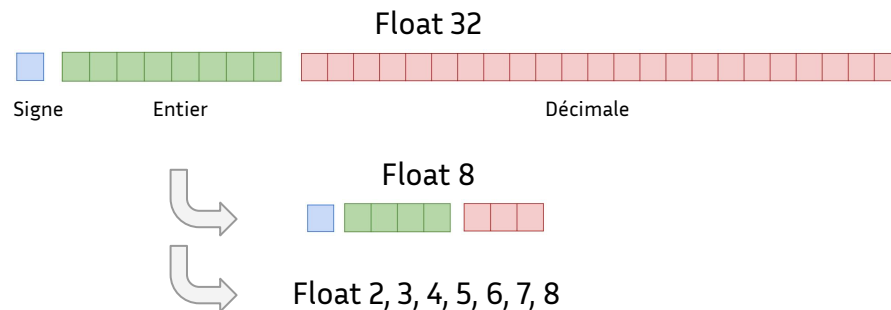
LocalLlama

r/LocalLLaMA

Rejointe



Quantification



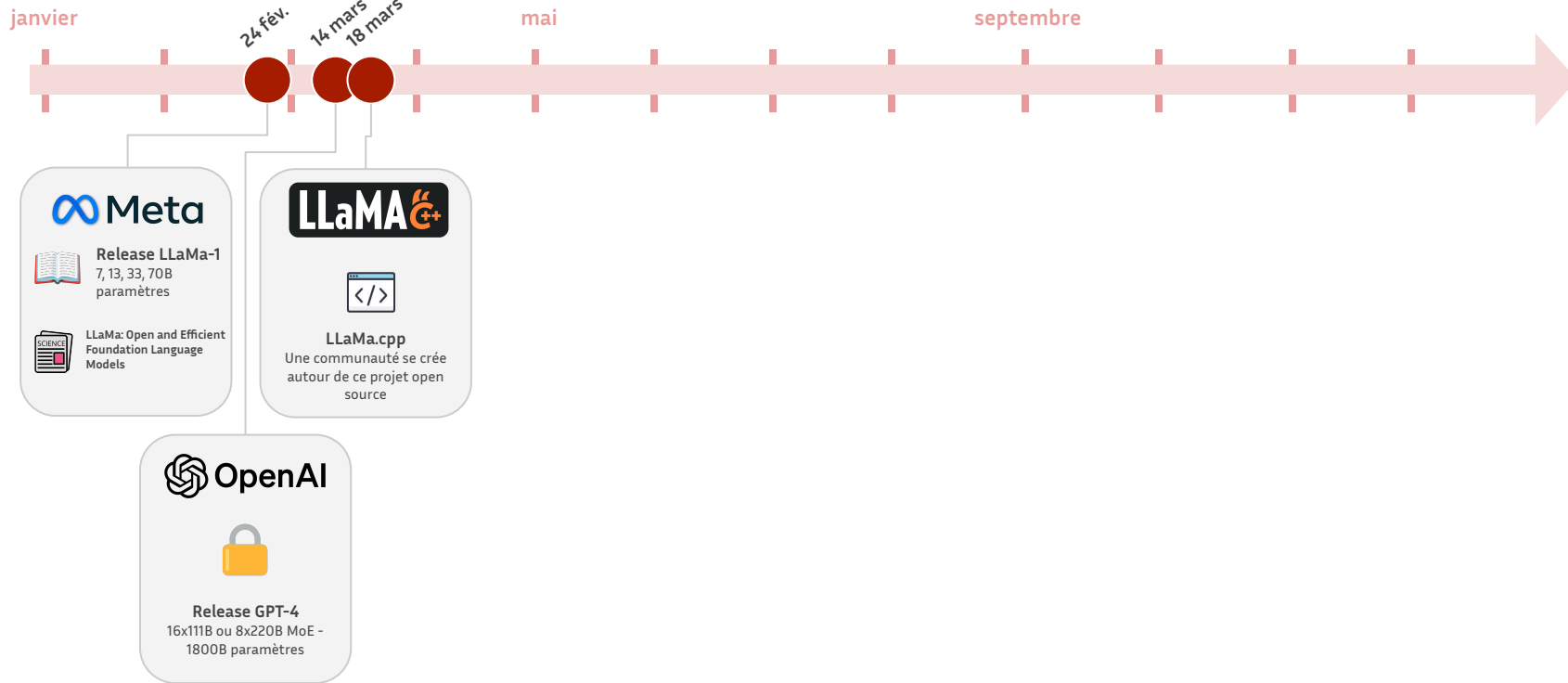
> Taille d'un modèle quantifié

7B ~ 30 Go → 4 bits → ~ 4 Go

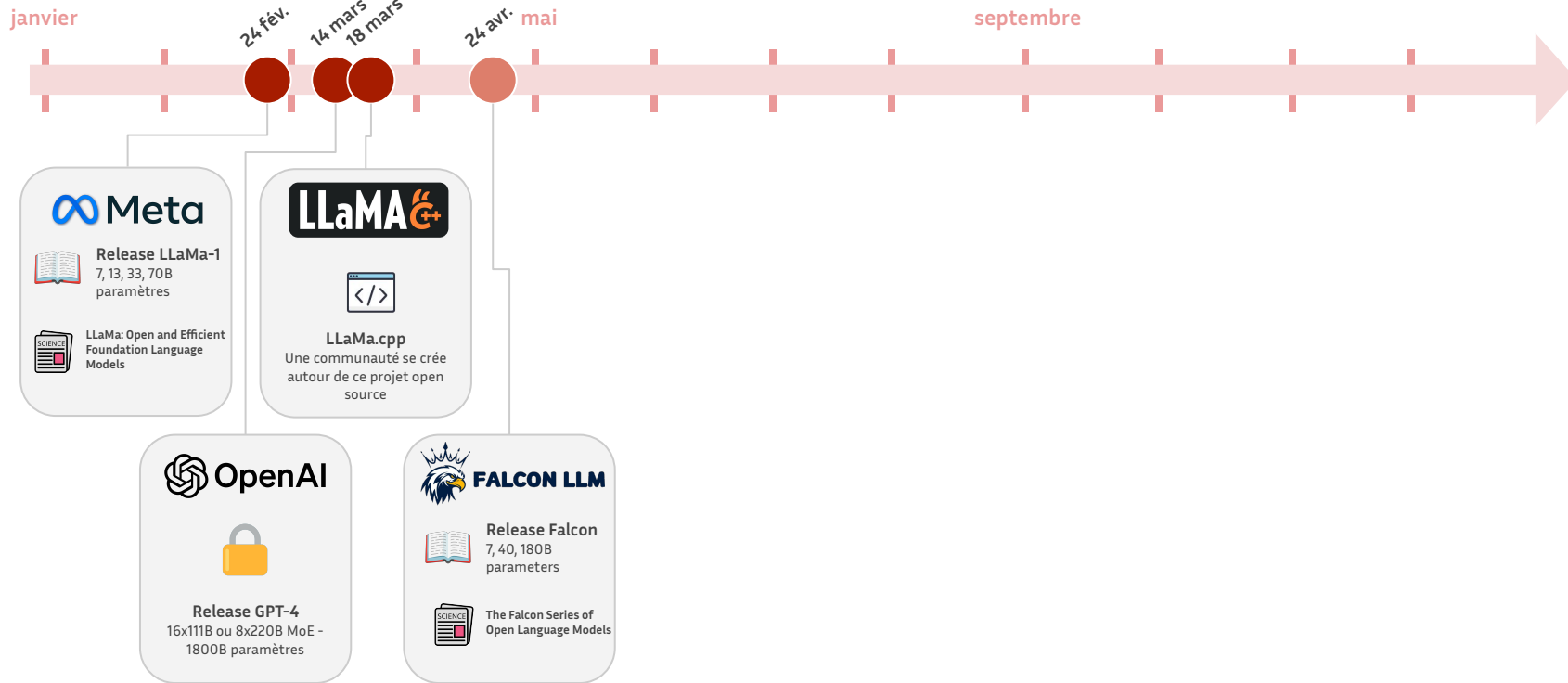
> RAM / VRAM nécessaire

7B ~ 30 Go → 4 bits → ~ 7 Go

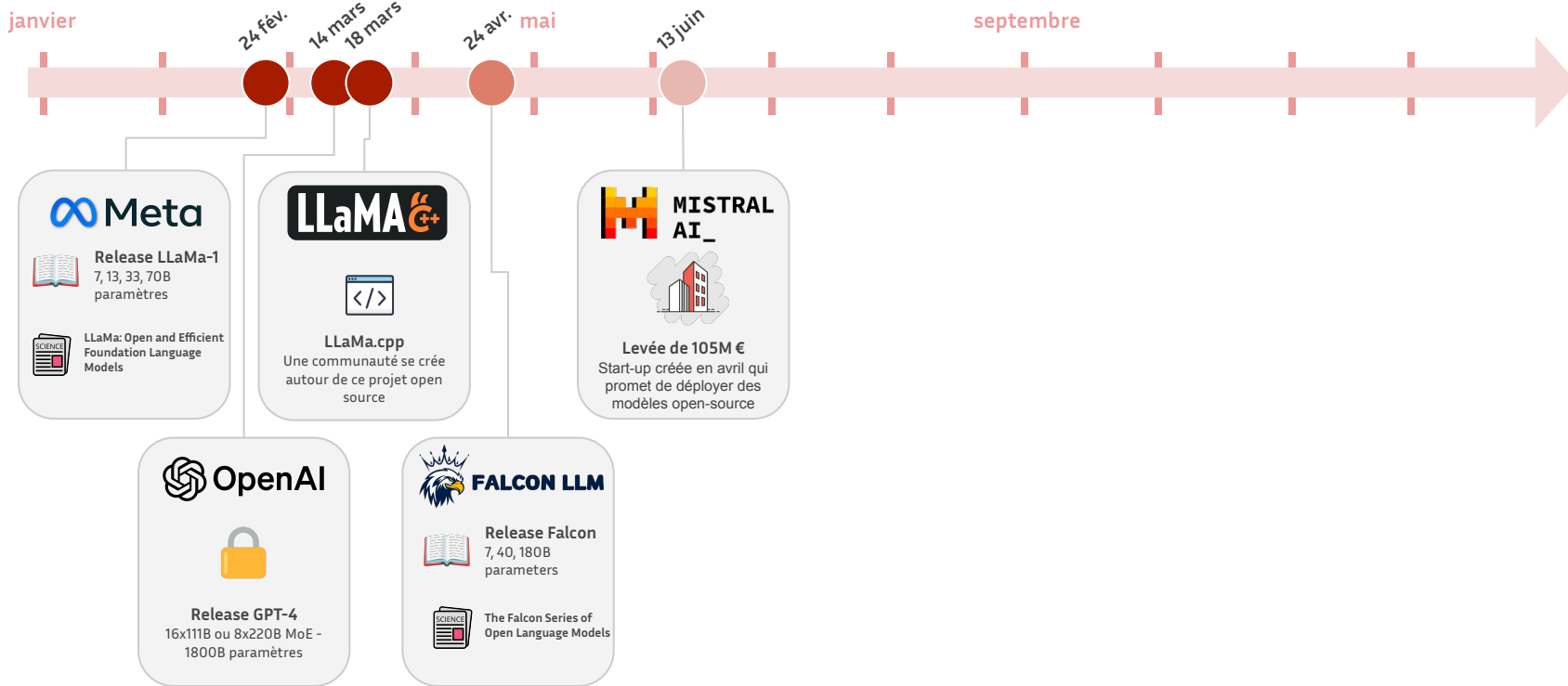
Timeline 2023



Timeline 2023



Timeline 2023



Timeline 2023

janvier

24 fév.


14 mars
18 mars


24 avr. mai


13 juin

18 juil.

septembre


Meta

 **Release LLaMa-1**
7, 13, 33, 70B
paramètres

 **LLaMa: Open and Efficient
Foundation Language
Models**


LLaMA





LLaMa.cpp
Une communauté se crée
autour de ce projet open
source


**MISTRAL
AI_**





Levée de 105M €
Start-up créée en avril qui
promet de déployer des
modèles open-source



OpenAI





Release GPT-4
16x111B ou 8x220B MoE -
1800B paramètres



FALCON LLM

 **Release Falcon**
7, 40, 180B
parameters

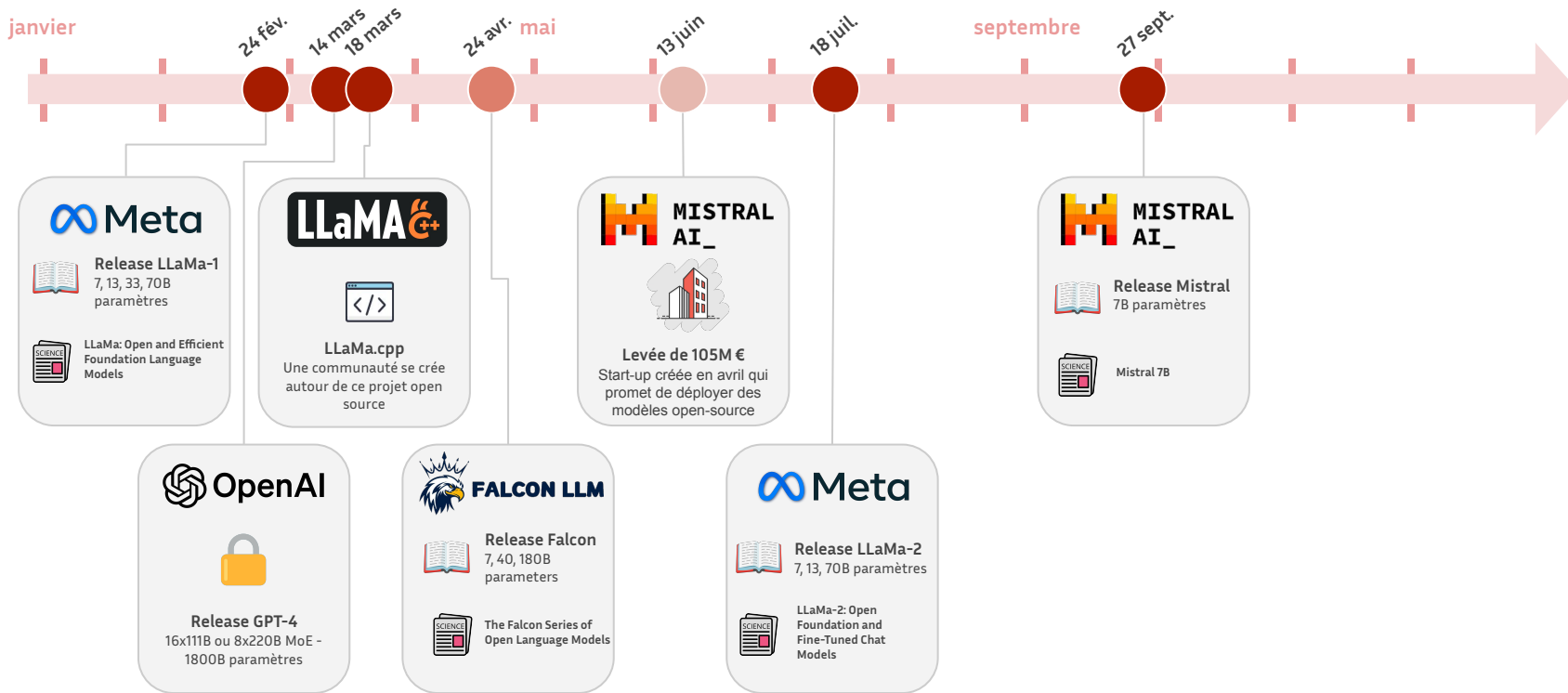
 **The Falcon Series of
Open Language Models**


Meta

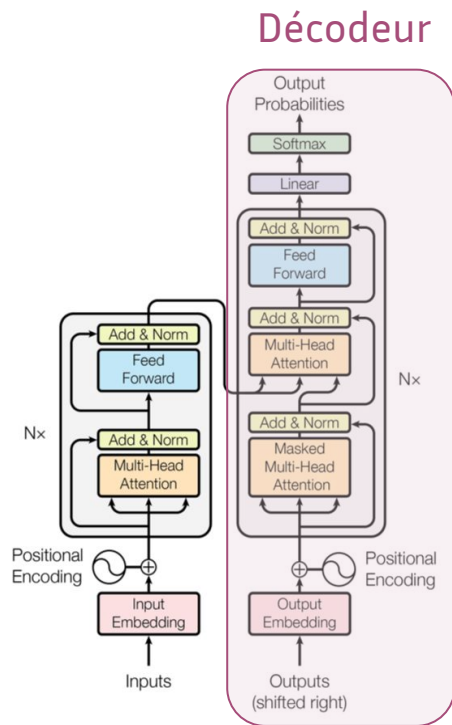
 **Release LLaMa-2**
7, 13, 70B paramètres

 **LLaMa-2: Open
Foundation and
Fine-Tuned Chat
Models**

Timeline 2023



Mistral



32 couches - 4096 dims

7B de paramètres

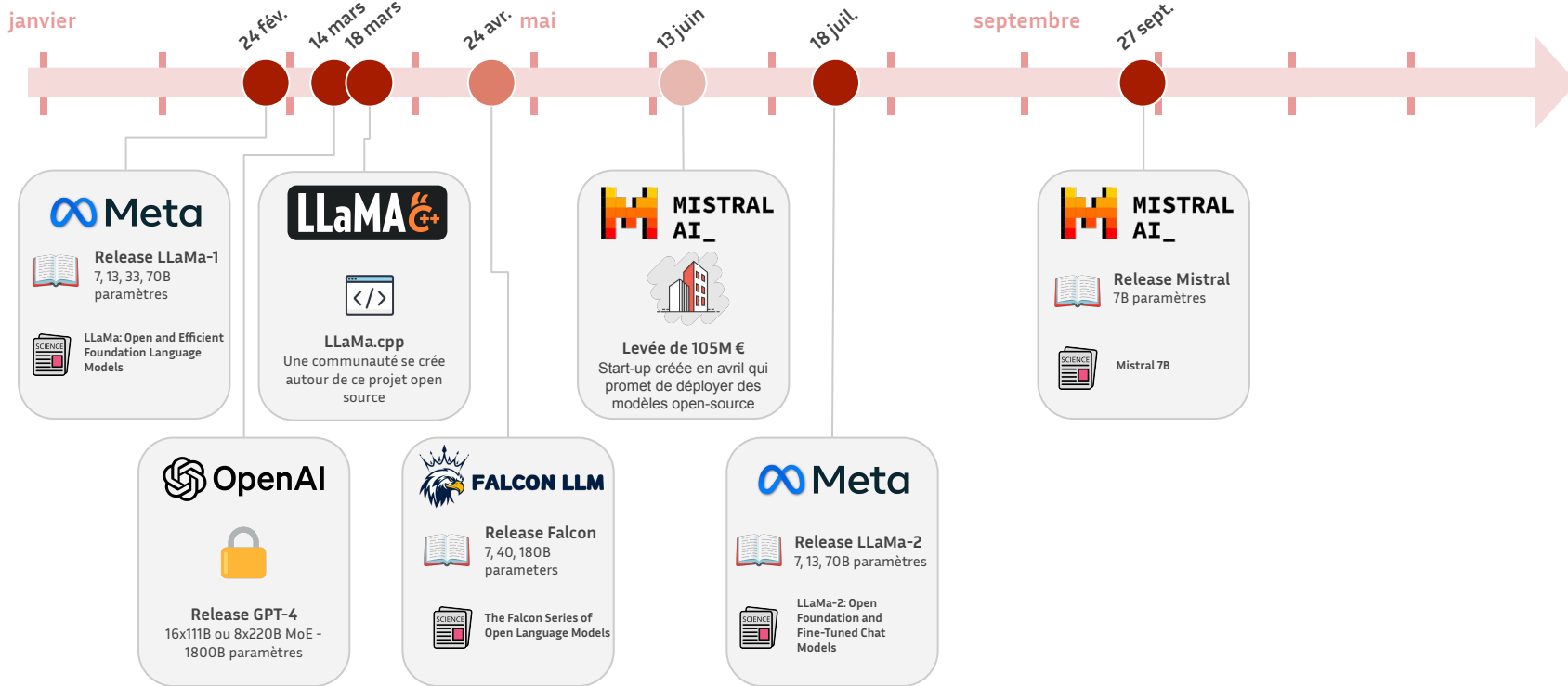
Dataset : ?

What follows is an explanation about Transformers in NLP.

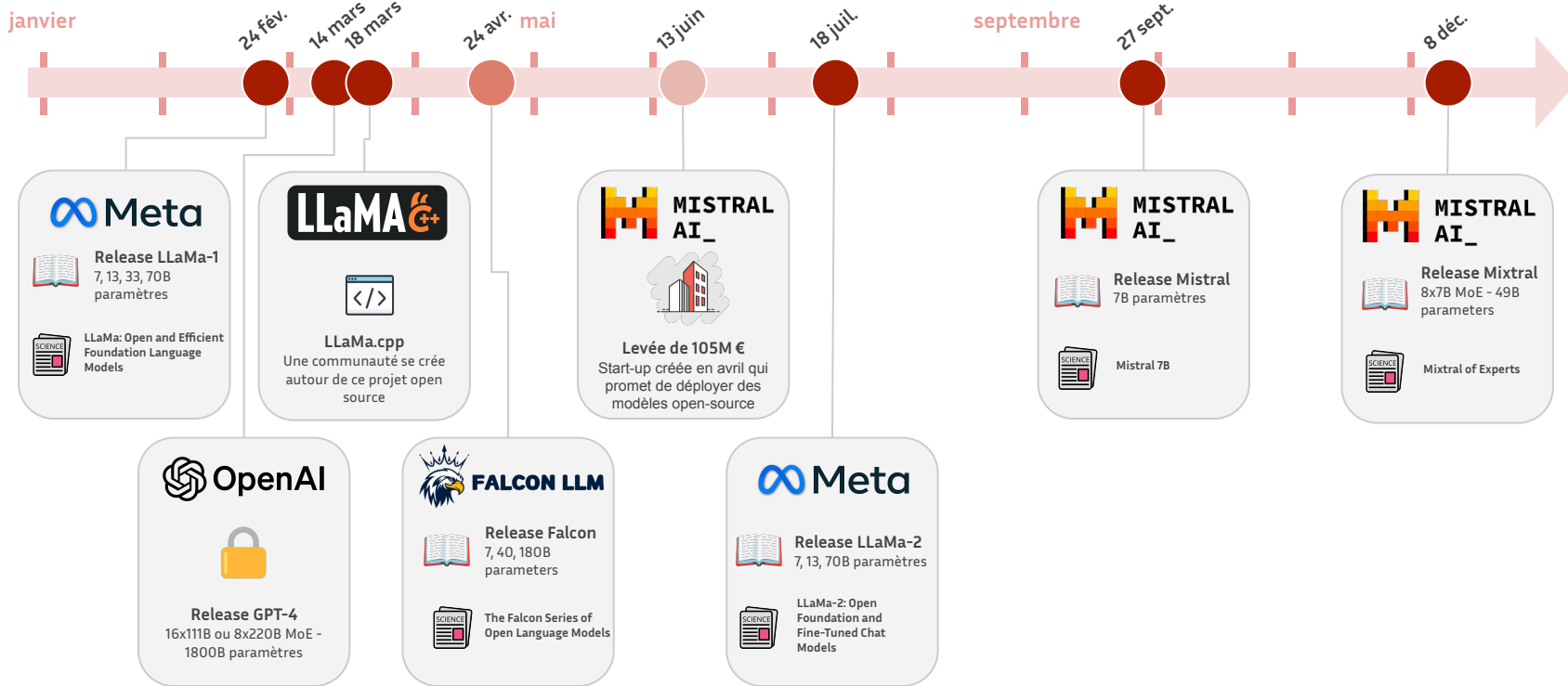
Transformers are a type of neural network model architecture that has revolutionized the field of Natural Language Processing (NLP) and achieved state-of-the-art results on various language understanding tasks, such as machine translation, question answering, text summarization, and speech recognition.

The Transformer model was introduced by Vaswani et al. in their 2017 paper "Attention Is All You Need." It departs from traditional Recurrent Neural Networks (RNNs) or Long Short-Term Memory networks (LSTMs), which process sequences one token at a time, by using self-attention mechanisms to compute relationships between all pairs of tokens within a sequence in parallel. This allows the model to capture long-range dependencies more effectively and makes it particularly well-suited for handling longer texts.

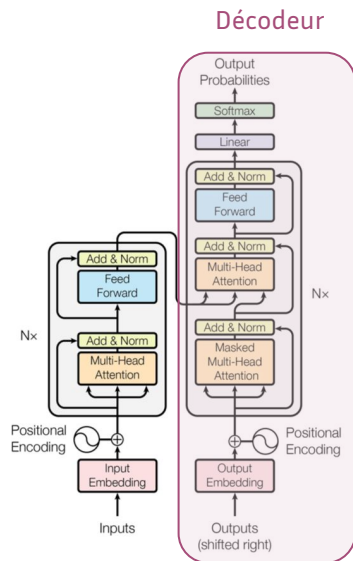
Timeline 2023



Timeline 2023



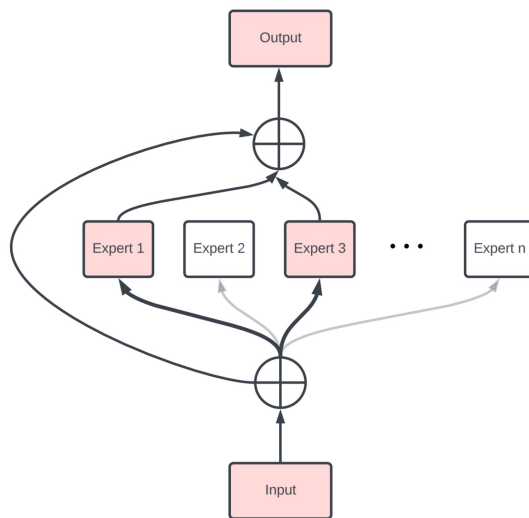
Mixtral



32 couches - 4096 dims

~49B de paramètres

Mixture of Expert 8x7B

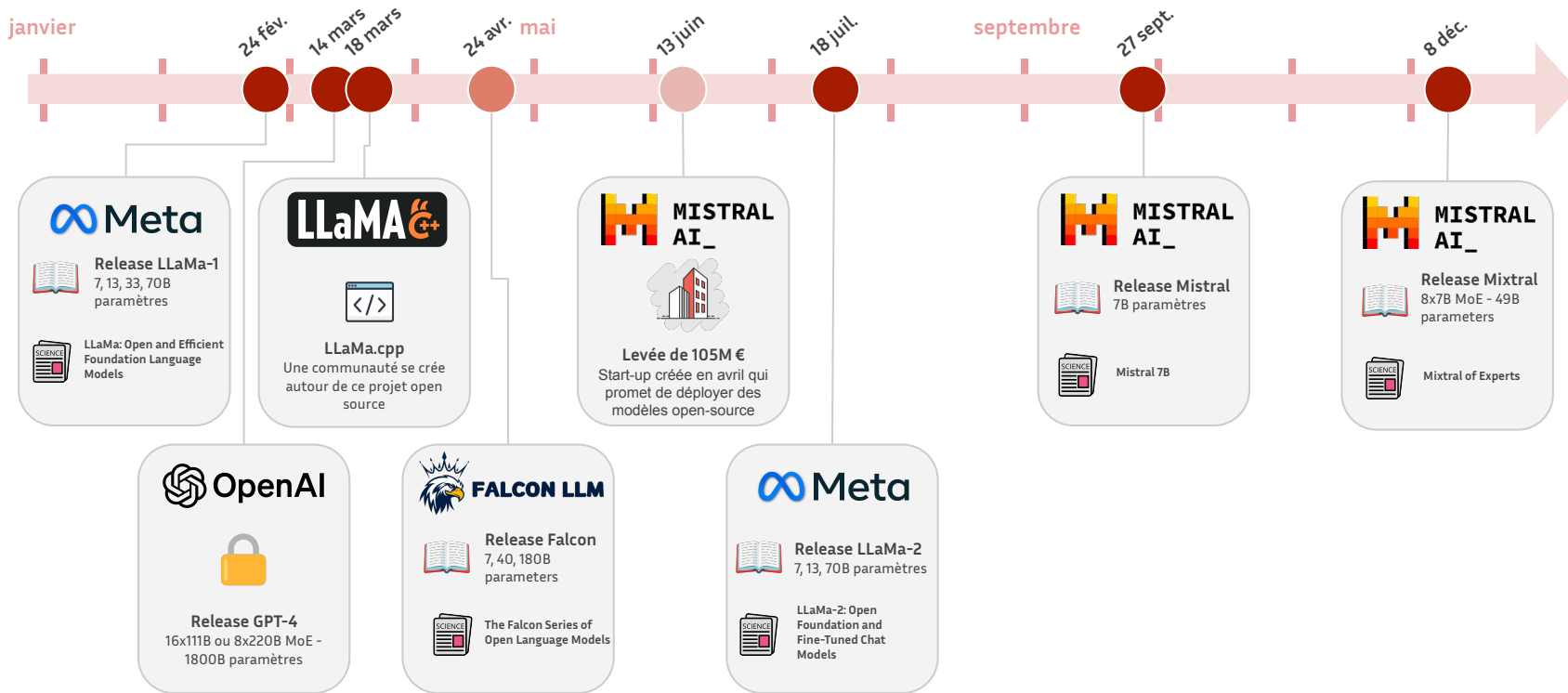


What follows is an explanation about Transformers in NLP.

Sure, I'm happy to help explain Transformers in the context of Natural Language Processing (NLP)!

Transformers are a type of model architecture used in machine learning, and they have been particularly successful in the field of NLP. The key innovation of Transformer models is their use of self-attention mechanisms, which allow the model to consider the entire input sequence when making predictions, rather than processing it sequentially like traditional recurrent neural networks (RNNs). This makes Transformer models more parallelizable and therefore faster to train on large datasets.

Timeline 2023



04

Evaluation des modèles

De l'open source dans le top 10 ?

Chatbot Arena

Expand to see 20+ Arena players

Model A

Dis moi une blague à propos de l'Inria.

Deux chercheurs de l'Inria discutent dans un café :

- Tu savais que l'Inria a développé un nouveau langage de programmation ?
- Ah oui ? C'est quoi comme langage ?
- C'est un langage tellement puissant qu'il peut programmer le temps !
- Incroyable ! Et comment ça marche ?
- Je ne sais pas, il n'est pas encore fini...

Bonus:

- Quelle est la différence entre un chercheur de l'Inria et un Jedi ?
- Un Jedi peut utiliser la Force, mais un chercheur de l'Inria peut utiliser Google.

Model B

Dis moi une blague à propos de l'Inria.

Pour vous raconter une blague à propos de l'Inria, il faut d'abord comprendre de quoi nous parlons. L'Inria est l'Institut National de Recherche en Informatique et en Automatique, une institution française de recherche en informatique et en automatique. Voici une blague à leur sujet :

Un chercheur de l'Inria se rendit un jour dans un village reculé pour étudier les habitudes des villageois en matière d'informatique. Au cours de son séjour, il remarqua que tous les habitants utilisaient des abeilles comme support d'information.

Curieux, il demanda à l'un des villageois pourquoi ils utilisaient des abeilles pour stocker des données. Le villageois lui répondit : "C'est simple, ici, nous appelons ça 'stock-age apicole' !"

A is better B is better Tie Both are bad

Enter your prompt and press ENTER Send

Chatbot Arena

The screenshot displays the Chatbot Arena interface with two chatbots, Model A and Model B, competing to generate a better response to the prompt: "Dis moi une blague à propos de l'Inria." Model A's response is a list of bullet points about a programming language. Model B's response is a short story about a researcher at Inria. At the bottom, there are four buttons for voting: "A is better", "B is better", "Tie", and "Both are bad". A text input field at the bottom left contains the prompt, and a "Send" button is on the right.

Expand to see 20+ Arena players

Model A

Dis moi une blague à propos de l'Inria.

Deux chercheurs de l'Inria discutent dans un café :

- Tu savais que l'Inria a développé un nouveau langage de programmation ?
- Ah oui ? C'est quoi comme langage ?
- C'est un langage tellement puissant qu'il peut programmer le temps !
- Incroyable ! Et comment ça marche ?
- Je ne sais pas, il n'est pas encore fini...

Bonus:

- Quelle est la différence entre un chercheur de l'Inria et un Jedi ?
- Un Jedi peut utiliser la Force, mais un chercheur de l'Inria peut utiliser Google.

Model B

Dis moi une blague à propos de l'Inria.

Pour vous raconter une blague à propos de l'Inria, il faut d'abord comprendre de quoi nous parlons. L'Inria est l'Institut National de Recherche en Informatique et en Automatique, une institution française de recherche en informatique et en automatique. Voici une blague à leur sujet :

Un chercheur de l'Inria se rendit un jour dans un village reculé pour étudier les habitudes des villageois en matière d'informatique. Au cours de son séjour, il remarqua que tous les habitants utilisaient des abeilles comme support d'information.

Curieux, il demanda à l'un des villageois pourquoi ils utilisaient des abeilles pour stocker des données. Le villageois lui répondit : "C'est simple, ici, nous appelons ça 'stock-age apicole' !"

A is better B is better Tie Both are bad

Enter your prompt and press ENTER Send

bard-jan-24-gemini-pro (2ème)

vicuna-33b (18ème)

Leaderboard - Chatbot Arena

Rank ▲	🤖 Model ▲	☆ Arena Elo ▼	Organization ▲	License ▲
1	GPT-4-Turbo	1249	OpenAI	Proprietary
2	Bard (Gemini Pro)	1215	Google	Proprietary
3	GPT-4-0314	1189	OpenAI	Proprietary
4	GPT-4-0613	1161	OpenAI	Proprietary
5	Mistral Medium	1150	Mistral	Proprietary
6	Claude-1	1150	Anthropic	Proprietary
7	Claude-2.0	1131	Anthropic	Proprietary
8	Mixtral-8x7b-Instruct-v0.1	1123	Mistral	Apache 2.0
9	Gemini Pro (Dev API)	1122	Google	Proprietary
10	Claude-2.1	1119	Anthropic	Proprietary
11	GPT-3.5-Turbo-0613	1117	OpenAI	Proprietary
12	Gemini Pro	1114	Google	Proprietary
13	Yi-34B-Chat	1112	01 AI	Yi License
14	Claude-Instant-1	1110	Anthropic	Proprietary
15	WizardLM-70B-v1.0	1106	Microsoft	Llama 2 Community
16	Tulu-2-DPO-70B	1105	AllenAI/UW	AI2 ImpACT Low-risk

Leaderboard - Chatbot Arena

Rank ▲	🤖 Model ▲	☆ Arena Elo ▼	Organization ▲	License ▲
1	GPT-4-Turbo	1249	OpenAI	Proprietary
2	Bard (Gemini Pro)	1215	Google	Proprietary
3	GPT-4-0314	1189	OpenAI	Proprietary
4	GPT-4-0613	1161	OpenAI	Proprietary
5	Mistral Medium	1150	Mistral	Proprietary
6	Claude-1	1150	Anthropic	Proprietary
7	Claude-2.0	1131	Anthropic	Proprietary
8	Mixtral-8x7b-Instruct-v0.1	1123	Mistral	Apache 2.0
9	Gemini Pro (Dev API)	1122	Google	Proprietary
10	Claude-2.1	1119	Anthropic	Proprietary
11	GPT-3.5-Turbo-0613	1117	OpenAI	Proprietary
12	Gemini Pro	1114	Google	Proprietary
13	Yi-34B-Chat	1112	01 AI	Yi License
14	Claude-Instant-1	1110	Anthropic	Proprietary
15	WizardLM-70B-v1.0	1106	Microsoft	Llama 2 Community
16	Tulu-2-DPO-70B	1105	AllenAI/UW	AI2 ImpACT Low-risk

Leaderboard - Chatbot Arena

Rank ▲	🤖 Model ▲	☆ Arena Elo ▼	Organization ▲	License ▲
17	GPT-3.5-Turbo-0314	1104	OpenAI	Proprietary
18	Vicuna-33B	1094	LMSYS	Non-commercial
19	Starling-LM-7B-alpha	1089	UC Berkeley	CC-BY-NC-4.0
20	Llama-2-70b-chat	1081	Meta	Llama 2 Community
22	OpenChat-3.5	1077	OpenChat	Apache-2.0
21	NV-Llama2-70B-SteerLM-Chat	1077	Nvidia	Llama 2 Community
23	OpenHermes-2.5-Mistral-7b	1075	NousResearch	Apache-2.0
24	pplx-70b-online	1074	Perplexity AI	Proprietary
25	GPT-3.5-Turbo-1106	1072	OpenAI	Proprietary
26	SOLAR-10.7B-Instruct-v1.0	1065	Upstage AI	CC-BY-NC-4.0
27	Dolphin-2.2.1-Mistral-7B	1064	Cognitive Computations	Apache-2.0
28	WizardLM-13b-v1.2	1058	Microsoft	Llama 2 Community
29	Zephyr-7b-beta	1050	HuggingFace	MIT
30	MPT-30B-chat	1043	MosaicML	CC-BY-NC-SA-4.0
31	CodeLlama-34B-instruct	1042	Meta	Llama 2 Community
32	Vicuna-13B	1040	LMSYS	Llama 2 Community

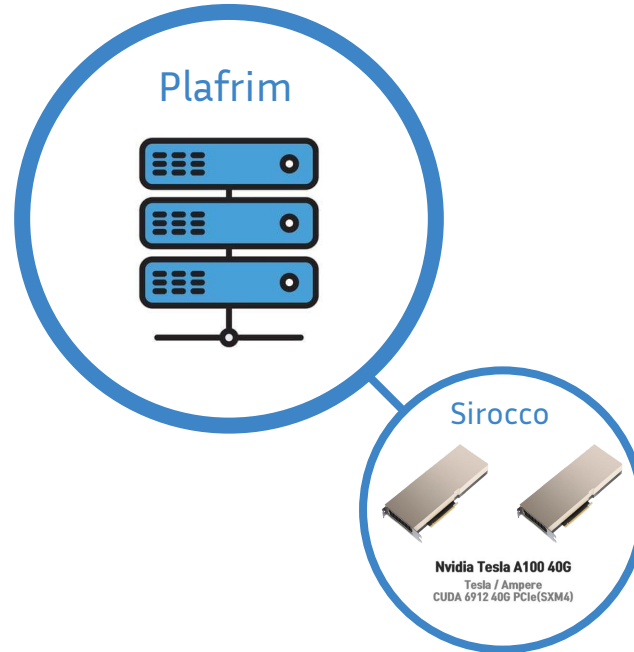
05

Mixtral sur Plafrim,
c'est possible ?

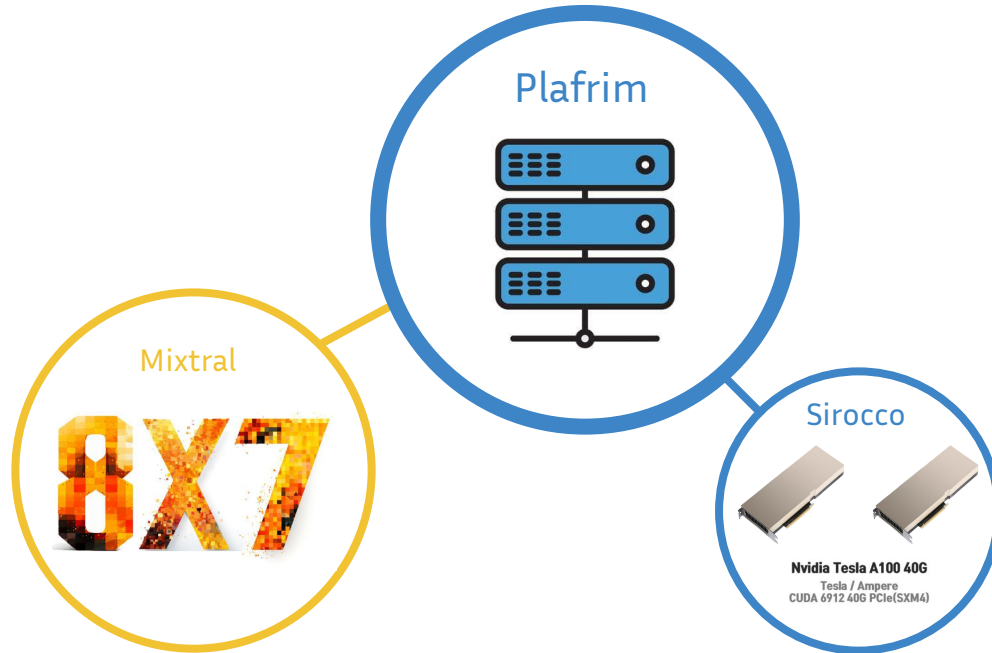
Mixtral sur Plafrim



Mixtral sur Plafrim



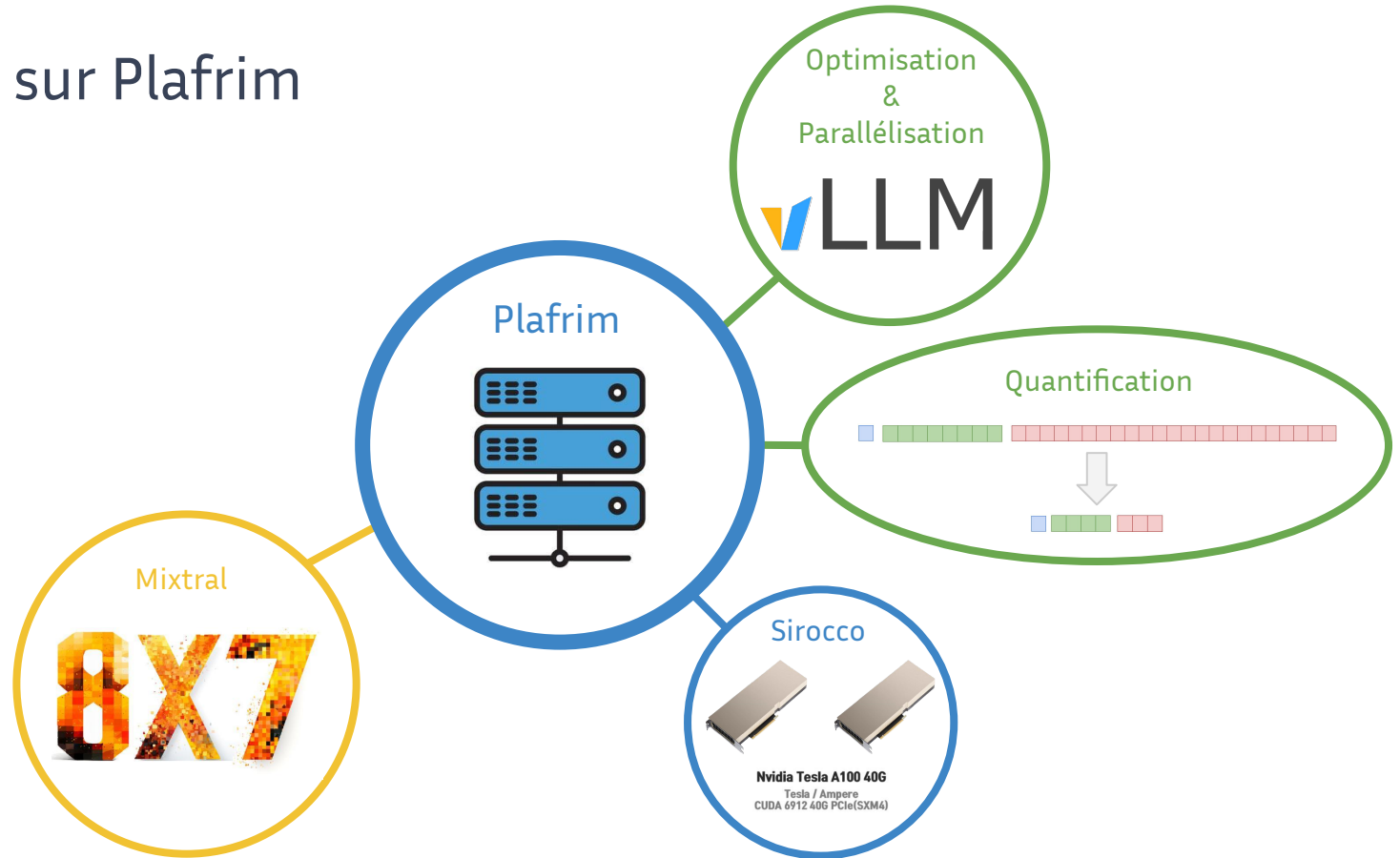
Mixtral sur Plafrim



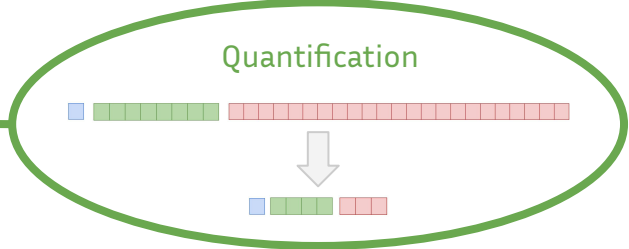
Mixtral sur Plafrim



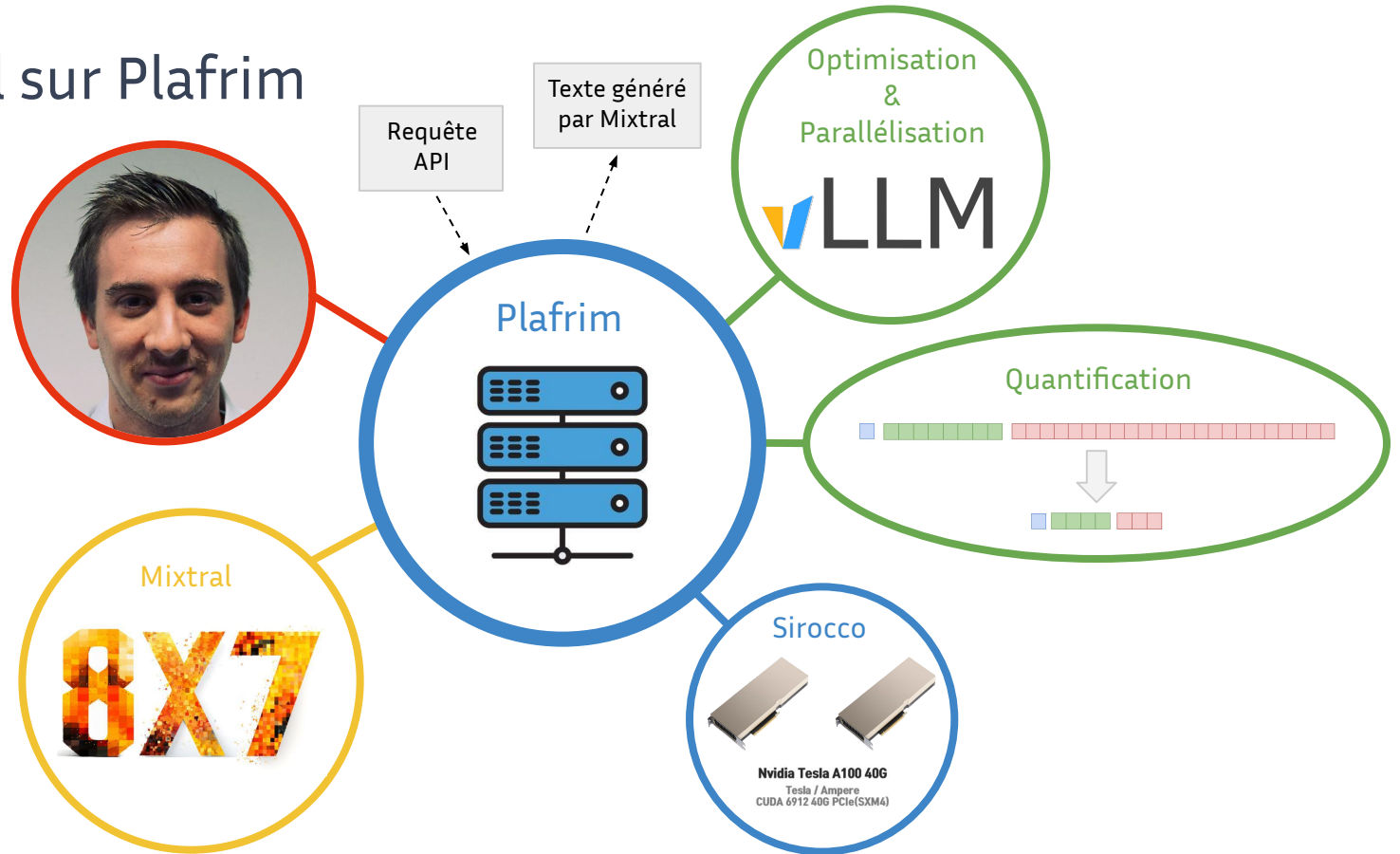
Mixtral sur Plafrim



Mixtral sur Plafrim



Mixtral sur Plafrim



05

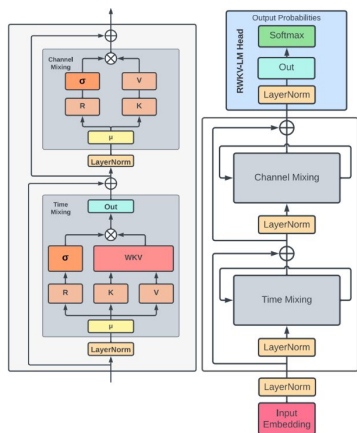
Des LLMs qui ne sont pas des Transformers

D'autres architectures

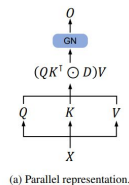
- > Calcul de l'attention des Transformers est très couteux (n^2)

D'autres architectures

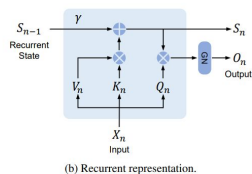
> Calcul de l'attention des Transformers est très couteux (n^2)



RWKV

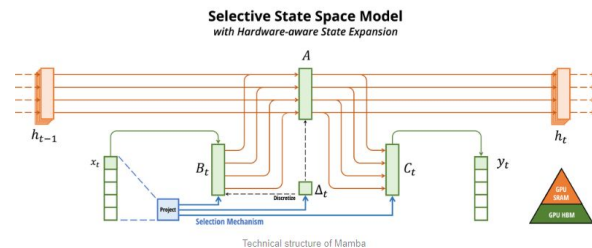


(a) Parallel representation.



(b) Recurrent representation.

RetNet



MAMBA

Merci !