Thésaurus vs Intelligence Artificielle : « Un thésaurus a-t-il encore du sens dans un monde d'IA ? »

Rencontres Hortidoc
18/11/2025 - Magalie Weber, Sophie Aubin & Maël Lesavourey

* Sommaire

« Un thésaurus a-t-il encore du sens dans un monde d'IA ? »

- Thésaurus et IA : alliés ou rivaux ?
- Quelques repères historiques
- IA générative, comment ça marche ?
- Vers le meilleur des deux mondes ?

Thésaurus et IA : alliés ou rivaux ?

De quoi parle-t-on ?

* Intelligence Artificielle (IA)

Ensemble **des théories et des techniques permettant à des machines d'accomplir des tâches** et de résoudre des problèmes normalement réservés aux humains et à certains animaux (**raisonnement, apprentissage**...)

(Yann Lecun, collège de France).

* Plusieurs formes d'Intelligence Artificielle

IA numérique vs IA symbolique

The big Controversy

Modéliser le cerveau :

« Penser s'apparente à un calcul massivement parallèle de **fonctions élémentaires.**

L'information est un **signal** avant d'être un code »[1]

Forger une opinion :

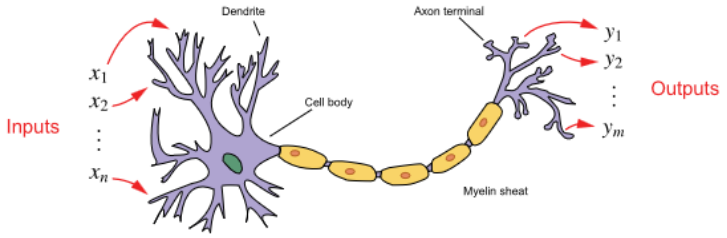
« Penser, c'est calculer des **symboles** qui ont à la fois une réalité matérielle et une valeur sémantique de représentation »[1]

L'information est une donnée symbolique de **haut niveau.**

Connexionniste vs **Symbolique**

Modéliser le cerveau
Modelling the brain

Forger une opinion
Making a mind




```
Tout [homme] est [mortel]
[Socrate] est un [homme]
Donc [Socrate] est [mortel]
```

[1] D Cardon, JP Cointet, A Mazieres (2018), La revanche des neurones https://dx.doi.org/10.3917/res.211.0173

source:Formation d'Introduction au Deep Learning FIDLE

* Approches inductives et déductives

The big Controversy

Approche **inductive**

Observations
Résultats
→ Programme (modèle)

Approche **déductive**

Observations
Programme
→ Résultats

Connexionniste vs **Symbolique**

Modéliser le cerveau
Modelling the brain

Forger une opinion
Making a mind

Observations ▶ Règles

Modèle

Expert

Règles ▶ Programme

source:Formation d'Introduction au Deep Learning FIDLE

* Thésaurus

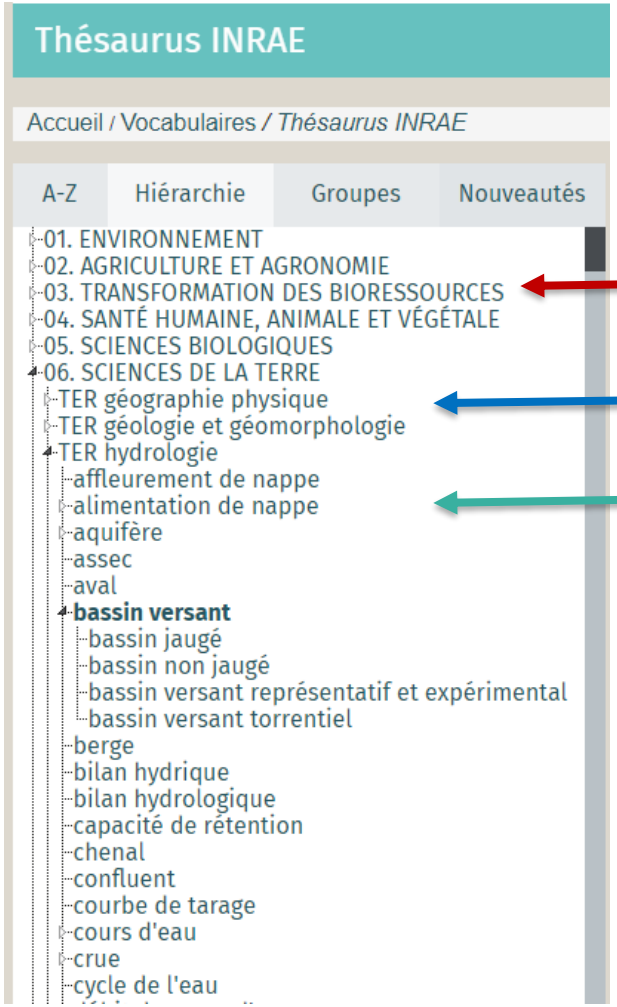
Ensemble organisé de termes contrôlés et normalisés qui expriment les **concepts** utiles à la description de contenus **propres à un domaine de connaissance**.

Langage documentaire défini par la norme ISO 25964 et le standard SKOS (Simple Knowledge Organization System) du W3C.

voir Vocabulaires Ouverts@INRAE https://vocabulaires-ouverts.inrae.fr/types-de-vocabulaires-et-usages

Thésaurus INRAE

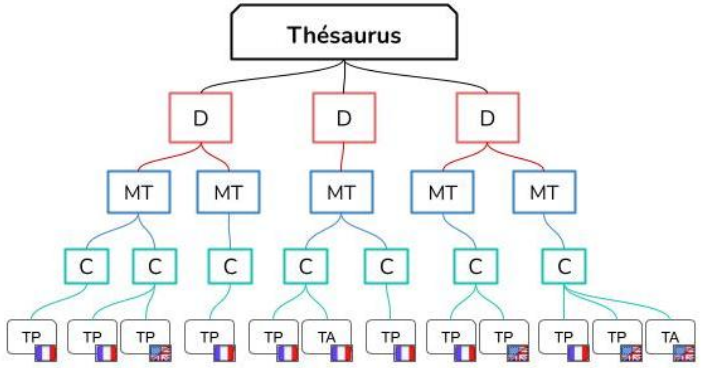
V2.5 (05-11-2025)



12 domaines

62 microthésaurus

16 420 concepts



23 300 termes FR
22 063 termes EN

\+

- des définitions textuelles
- des collections (groupes)
- des alignements avec d'autres vocabulaires

Aubin, S. et al. 2024. Révéler et valoriser vos données avec le thésaurus INRAE. *NOV'AE - Ingénierie et savoir-faire innovants*. régulier 03 NOV'AE (sept. 2024). DOI:https://doi.org/10.17180/novae-2024-NO-art03.

* les relations

Le thésaurus présente deux types de relations distinctes :

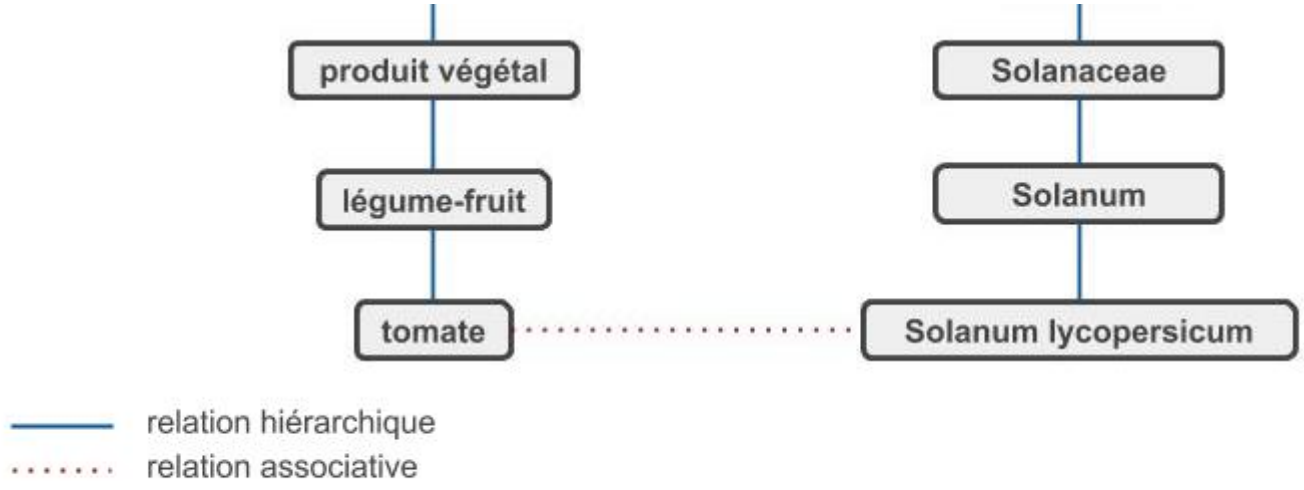
- Des **relations hiérarchiques** entre un concept générique et un spécifique qui se traduit par :

c.spécifique est un type de/ partie de *c.générique*

tomate est un type *de légume-fruit*

- Des **relations associatives** entre certains concepts pouvant être liés sémantiquement:

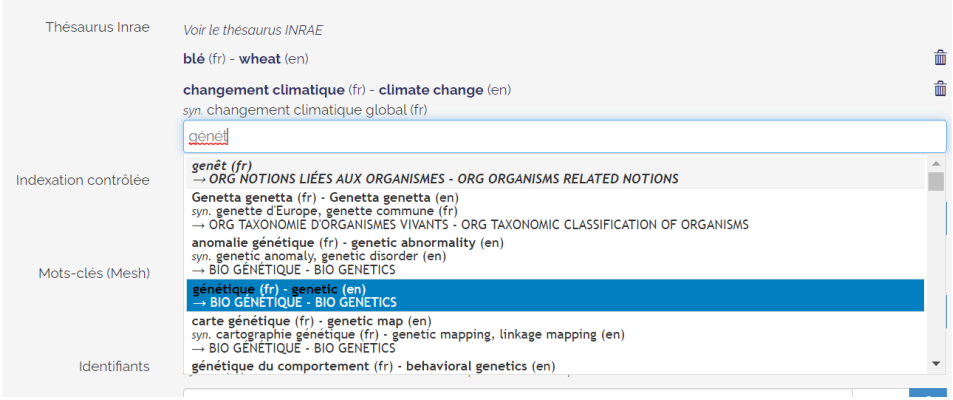
Solanum lycopersicum est lié à *tomate*



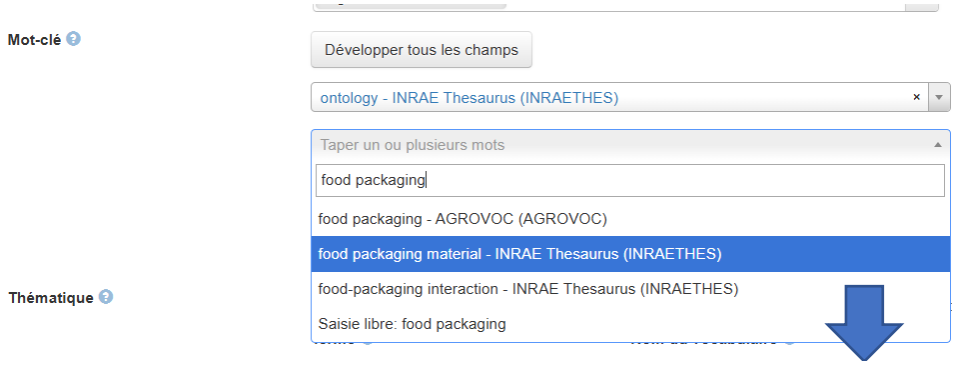
https://vocabulaires-ouverts.inrae.fr/a-propos-du-thesaurus-inrae/

* Intégration dans les SI INRAE

- HAL INRAE : intégré depuis mars 2021 [vidéo de démonstration (1’27’’)]



- Data INRAE : connecteur intégré depuis octobre 2024



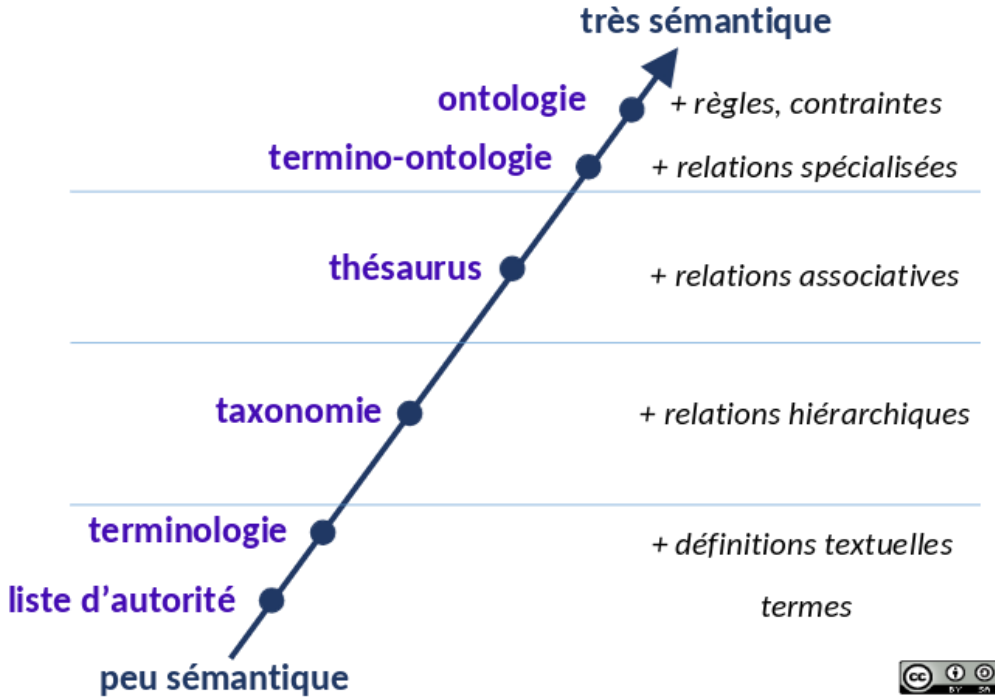
* Typologie des vocabulaires contrôlés

Raisonner
Décrire le réel pour valider, simuler, générer de nouvelles connaissances

Abstraire
Indexer dans une ou plusieurs langues, documenter de manière non ambigue

Classer
Organiser les objets numériques selon leurs propriétés ou leur thématique

Normaliser
S'affranchir de la variabilité de la langue, proposer une recherche à facettes



Vocabulaires Ouverts@INRAE https://vocabulaires-ouverts.inrae.fr/types-de-vocabulaires-et-usages

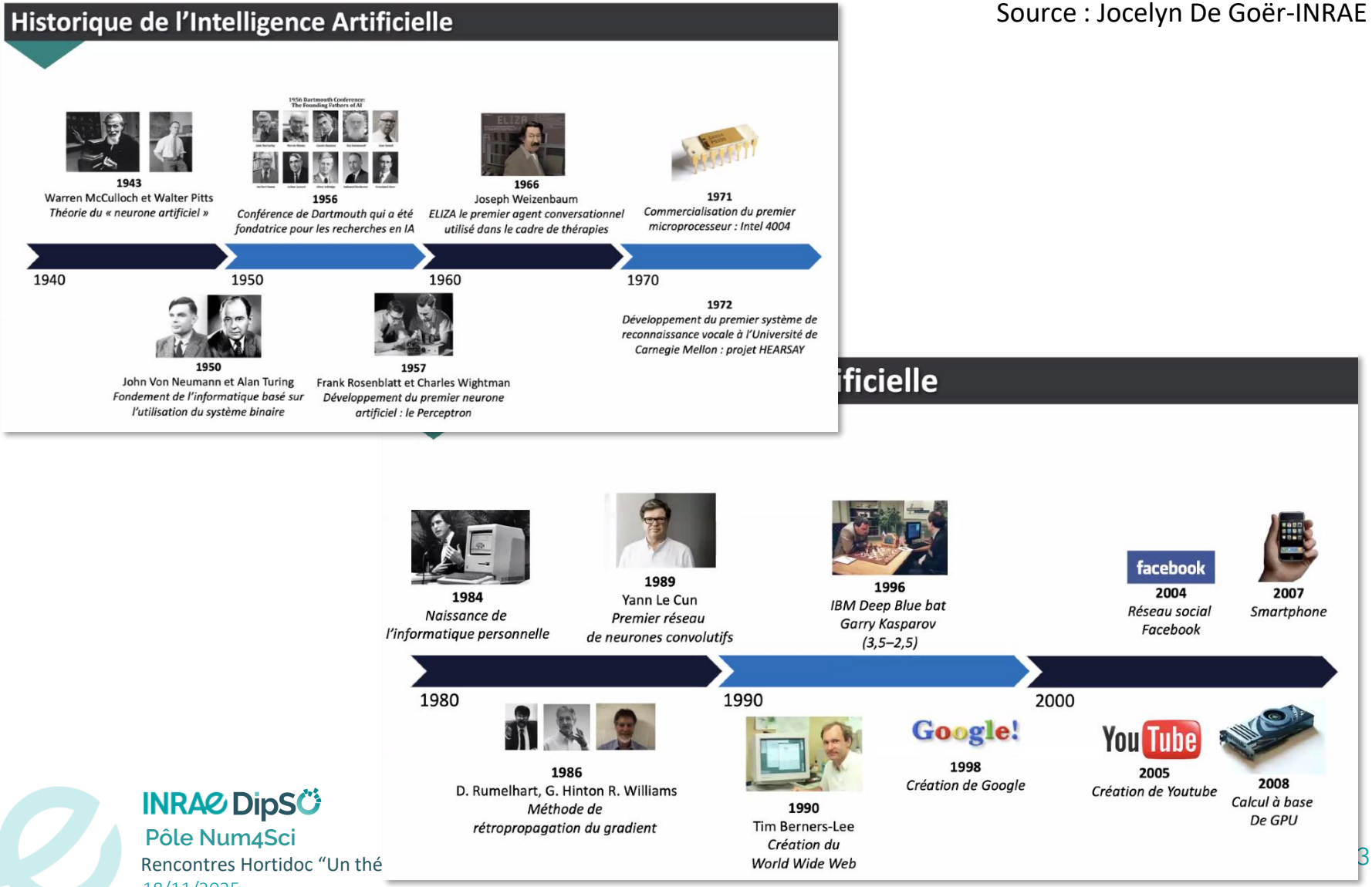
Quelques repères historiques...

"Retour vers le futur"

* Un peu d'histoire

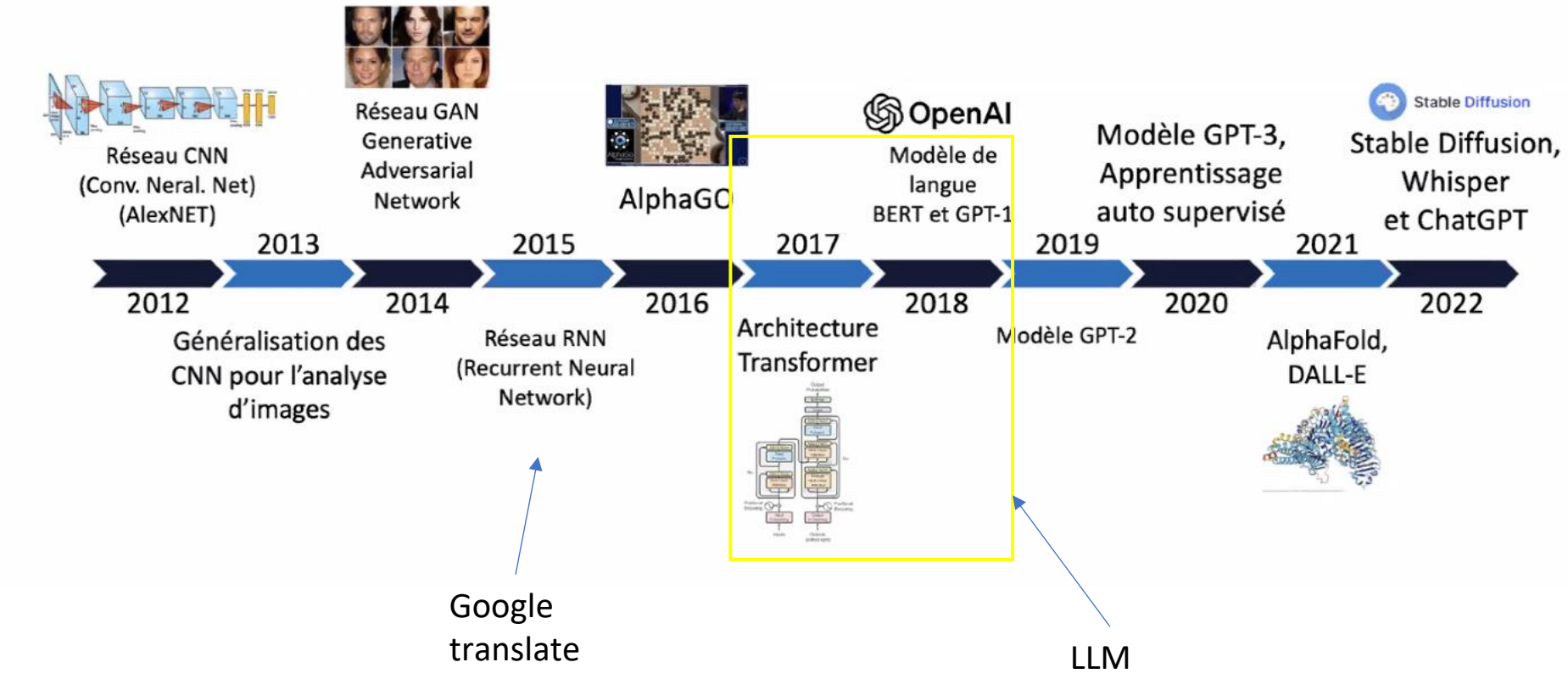
Les fondations

Source : Jocelyn De Goër-INRAE



* Un peu d'histoire

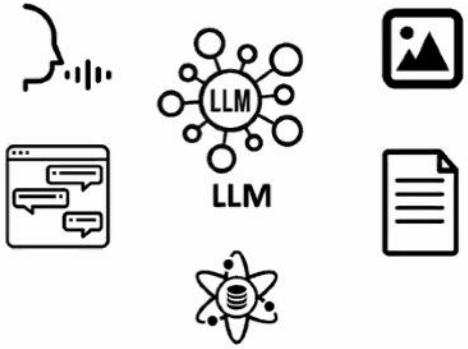
Un tournant en 2012 et une accélération depuis 2017



Source : Jocelyn De Goër-INRAE

* Depuis 2022… ChatGPT mania !

Modèle multimodal



Modèles propriétaires

ChatGPT

COPILOT

Gemini

Claude

Modèles réutilisables

LLaMA by Meta

FALCON LLM

Stanford Alpaca

BLOOM

Hugging Face

581 420 modèles

Source : Jocelyn De Goër-INRAE

* Intelligence Artificielle Générative (GenAI)

Catégorie d'IA qui se concentre sur la **création** autonome de contenu, tels que des textes, des images, des vidéos, des sons et d'autres types de données, par des systèmes informatiques.

Elle fonctionne en apprenant **les schémas et structures des données existantes**, puis elle les utilise pour **générer** de nouveaux éléments qui sont similaires, mais non identiques, à ce qu'elle a appris.

https://www.grandeecolenumerique.fr/les-metiers-porteurs/qu-est-ce-ia-intelligence-artificielle-generative

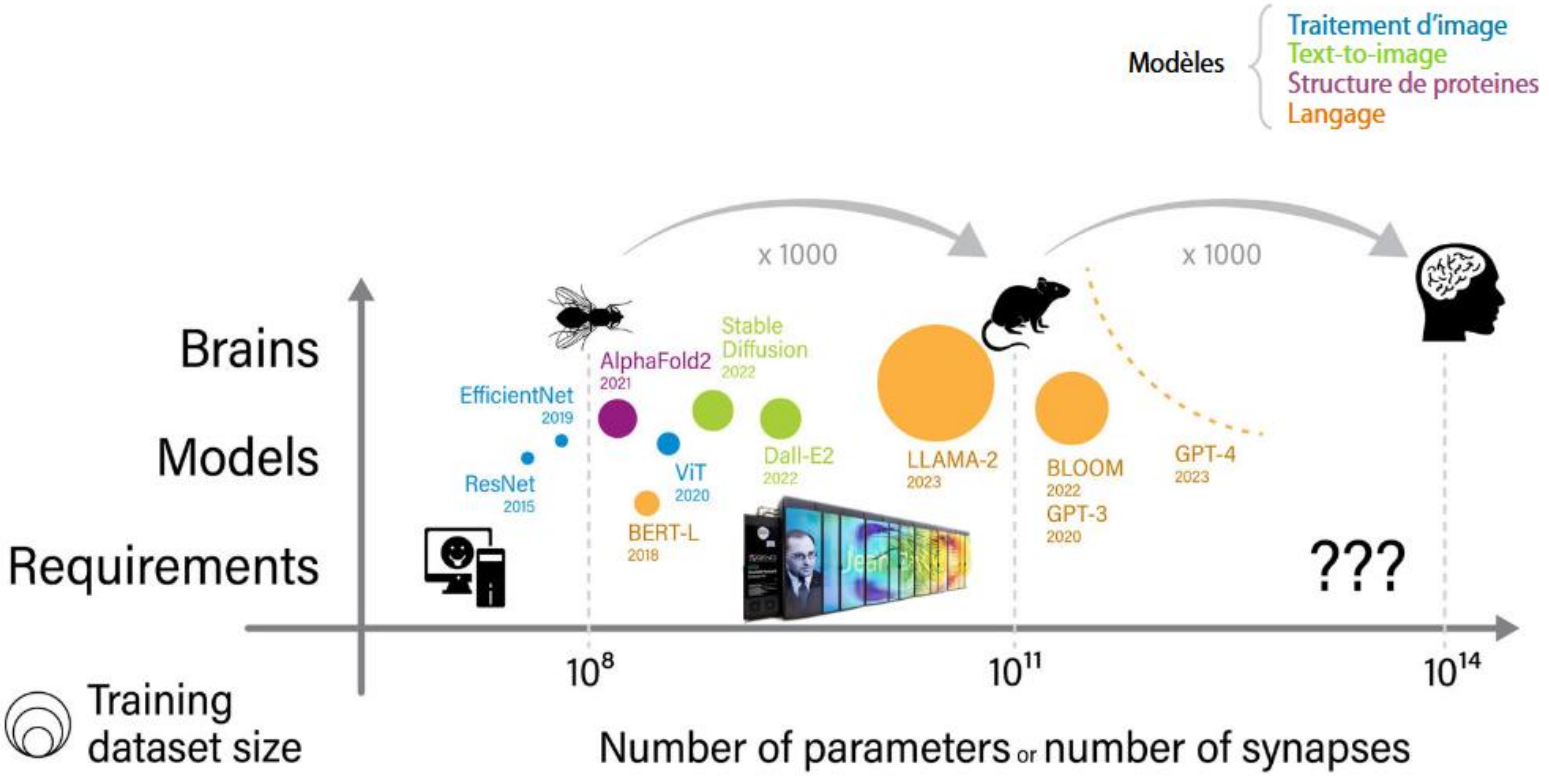
* Large Language Model (LLM)

Modèle de langue = modèle statistique de la distribution d'unité linguistiques (par exemple : lettres, phonèmes, mots) dans une langue naturelle.

On parle de **modèles de langage de grande taille** (LLM) pour les modèles possédant un **grand nombre de paramètres** (> milliards)

* Une explosion de la taille des modèles

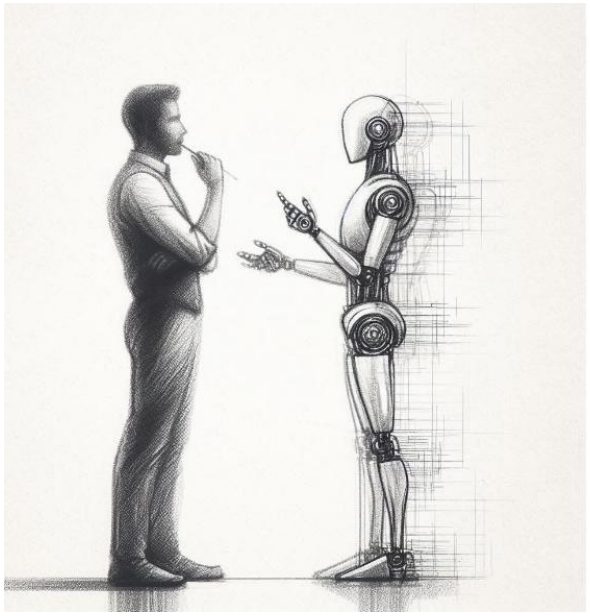
De l'IA Générative à l'IA Générale ?



source:Formation d'Introduction au Deep Learning FIDLE
https://www.fidle.cnrs.fr/w3/

* Et au-delà ??

De l'Intelligence Artificielle Générale à la Super Intelligence

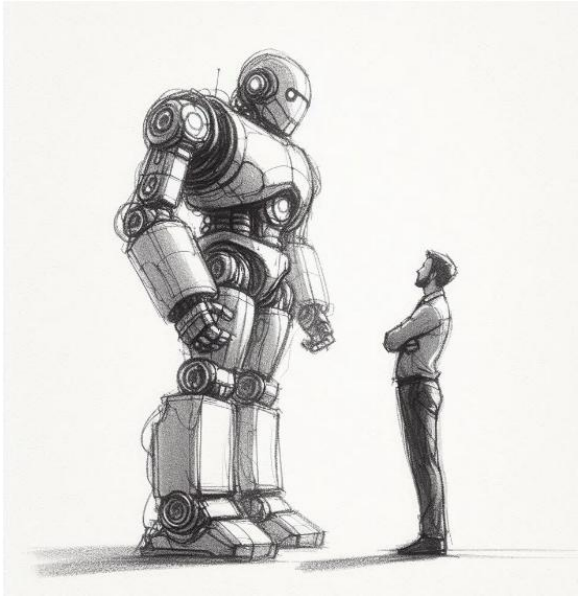


Artificial General Intelligence

Intelligence Artificielle au niveau humain

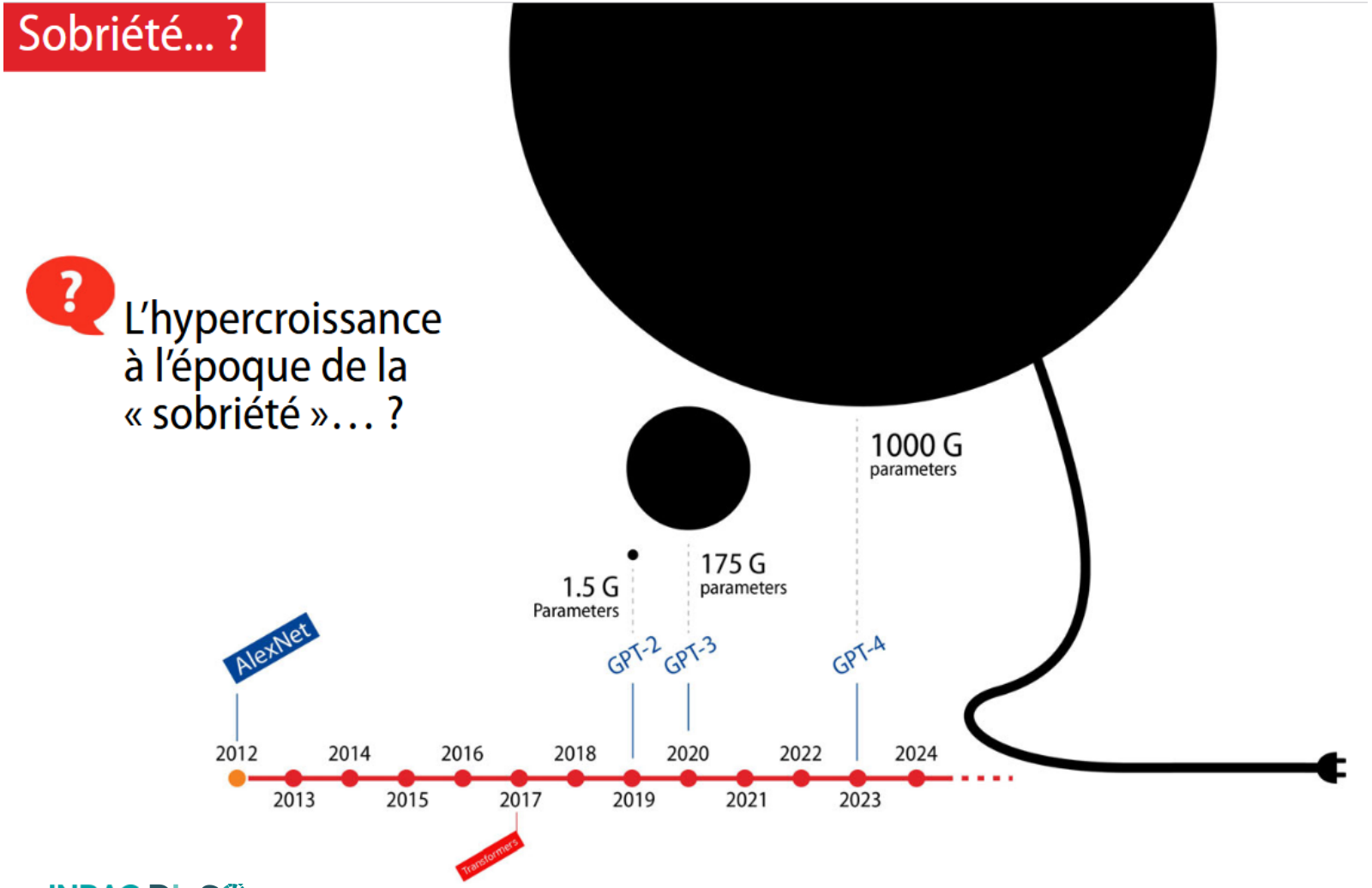
Artificial Super Intelligence

Intelligence Artificielle supérieure à l'humain



source:Formation d'Introduction au Deep Learning FIDLE https://www.fidle.cnrs.fr/w3/

* Un développement “soutenable” ?



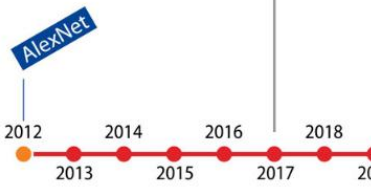
source:Formation d’Introduction au Deep Learning FIDLE
https://www.fidle.cnrs.fr/w3/

Classification (1000 classes)

25 M paramètres (Resnet-50)
1,2 M labeled images

14 Jours
1 NVIDIA M40 GPU

14 Jours GPU



117 Jours
384 A100 GPU

44 928 Jours GPU

123 ans

54 Jours
16 384 H100 GPU

884 736 Jours GPU

2400 ans

a BigScience initiative
BLOOM
176B params · 59 languages · Open-access

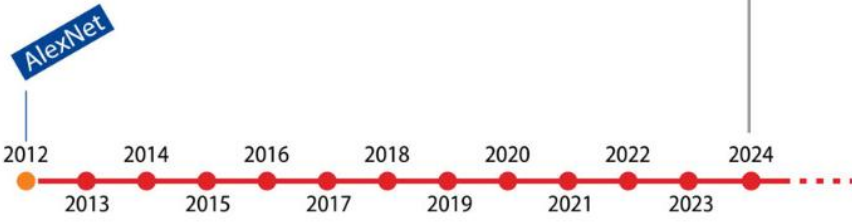
Génération de texte (LLM)

176 B paramètres
366 B tokens (~3 M livres)

Llama 3

Génération de texte (LLM)

405 B paramètres
15 T tokens (~150 M livres)



* Recommandations pour utiliser l'IA en recherche

International AI Safety Report

→ FIRST KEY UPDATE: OCT 2025 → INTERNATIONAL AI SAFETY REPORT: JAN 2025

AISI AI SECURITY INSTITUTE Mila

https://internationalaisafetyreport.org/

General guidelines
15 April 2025
Living guidelines on the responsible use of generative AI in research

INRAE DipSO

Recommandations pour l'usage des IA génératives comme assistant personnel au sein d'INRAE

hal-04692524 , version 1, 13/09/2024

RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

cirad
LA RECHERCHE AGRONOMIQUE POUR LE DÉVELOPPEMENT

Usages des outils d'Intelligence Artificielle générative dans le domaine de la recherche
- Points de vigilance et bonnes pratiques -

Note de synthèse IAg - version 2 du 5 octobre 2024

Inria

Réflexions sur l'usage de l'IA générative pour les métiers de la recherche

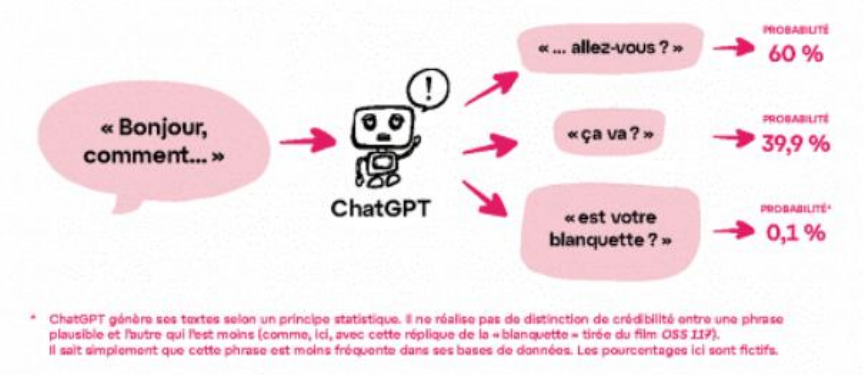
hal-05187992, version 1, 26-07-2025

L’IA générative, comment ça marche ?

* Un LLM, comment ça marche ?

Exemple pour un chatBot (assistant conversationnel)

prompt



contenu inédit probable

Judith Lorne

* Un LLM, comment ça marche ?

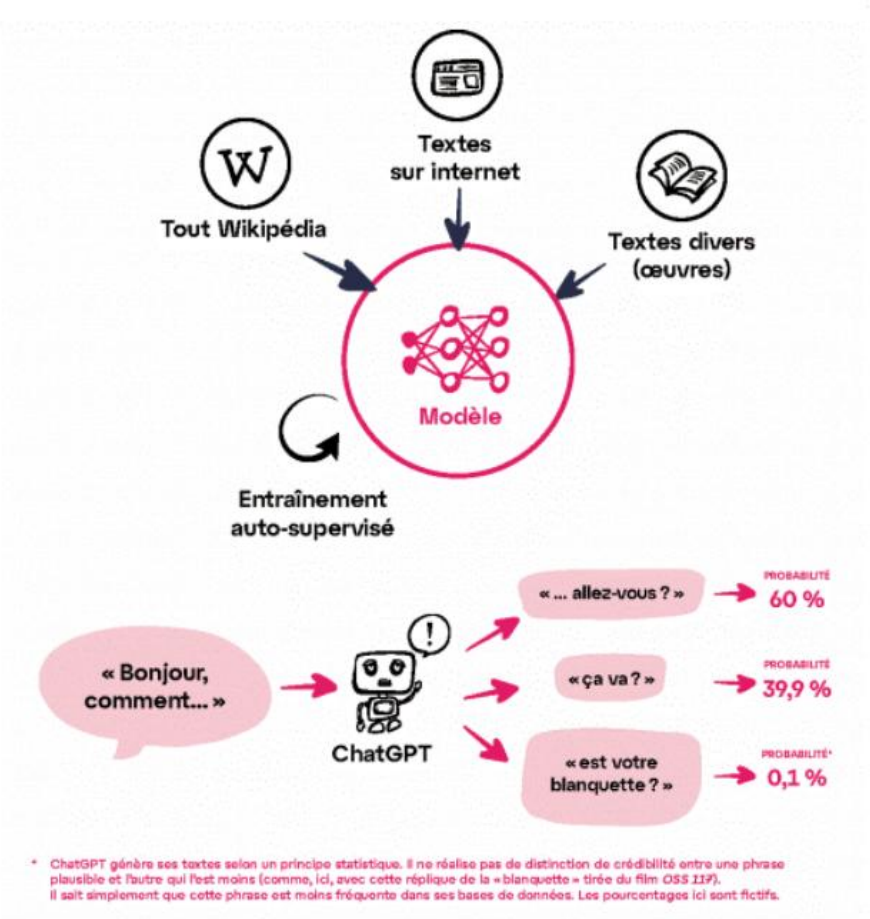
Exemple pour un chatBot (agent conversationnel)

données

algorithmes

entraînement

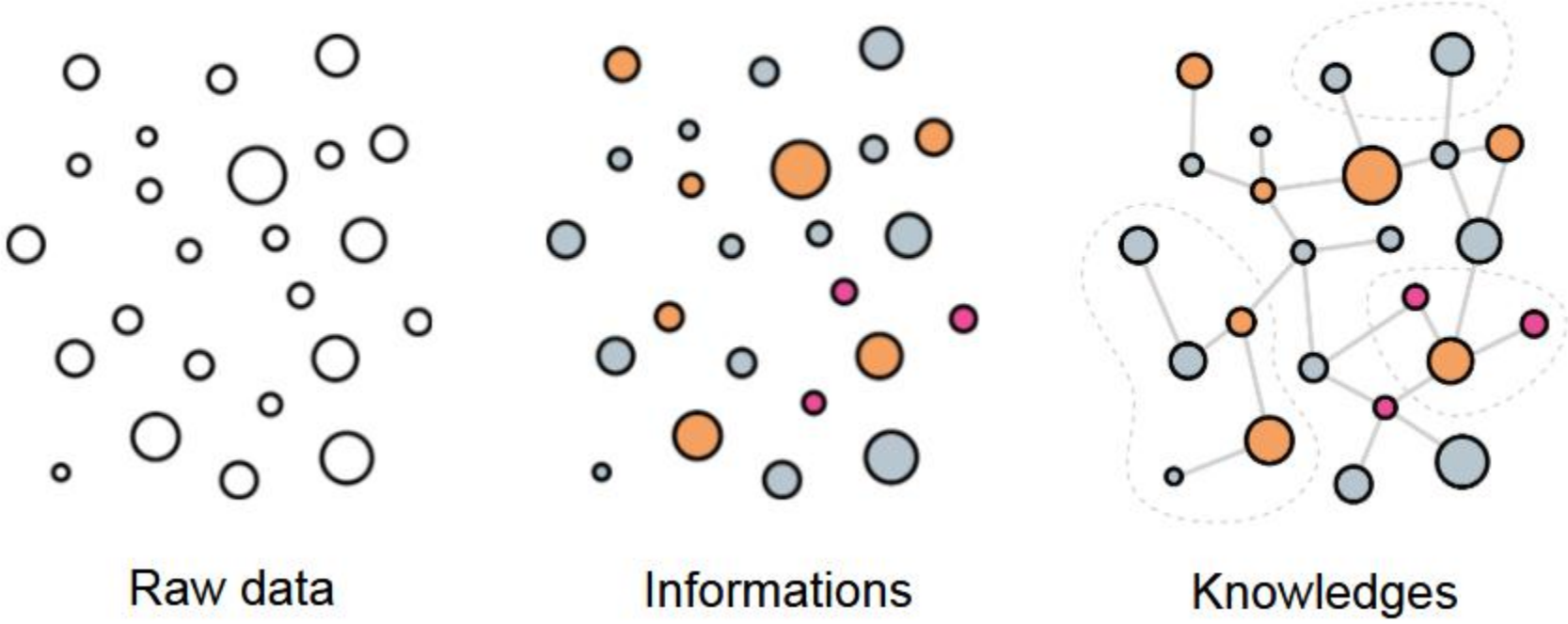
prompt



contenu inédit probable

Judith Lorne

* De la donnée brute aux connaissances



source:Formation d'Introduction au Deep Learning FIDLE
https://www.fidle.cnrs.fr/w3/

* Au départ était la donnée…

[Data]

"Ce qui est **connu** et **admis**, et qui sert de base, à un raisonnement, à un examen ou à une recherche."

Centre National de Ressources Textuelles et Lexicales - CNRTL
https://www.cnrtl.fr/definition/donnée

source:Formation d'Introduction au Deep Learning FIDLE
https://www.fidle.cnrs.fr/w3/

* Puis vint “l’intelligence”

[intelligence]

« Ensemble des fonctions mentales ayant pour objet la connaissance conceptuelle et rationnelle »*

*opposé à *sensation* et *intuition*

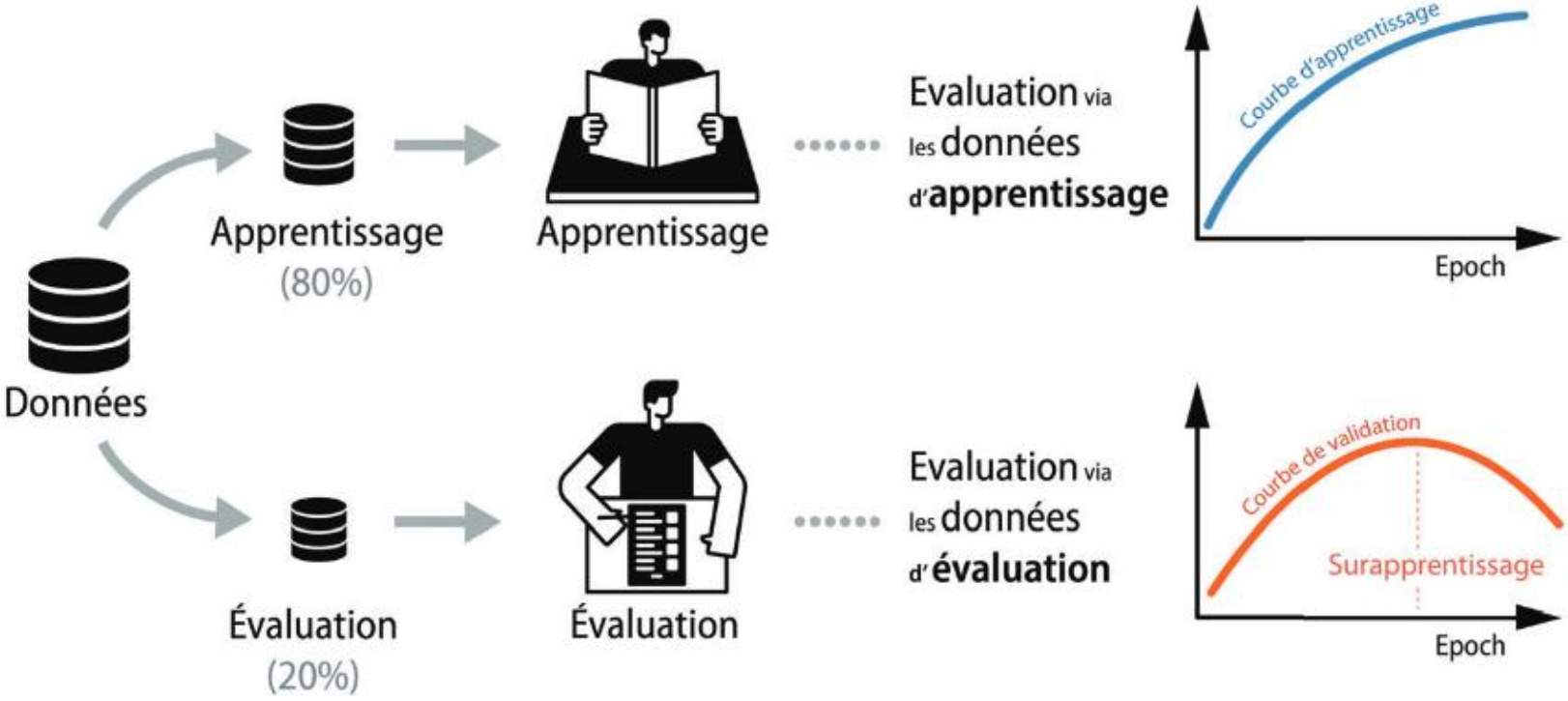
« Set of mental functions aimed at conceptual and rational knowledge »

Les grands modèles de langage deviennent des raisonneurs neuro-symboliques

* Des “datasets” toujours plus gros...mais

source:Formation d’Introduction au Deep Learning FIDLE
https://www.fidle.cnrs.fr/w3/

* L'enjeu des données pour l'apprentissage

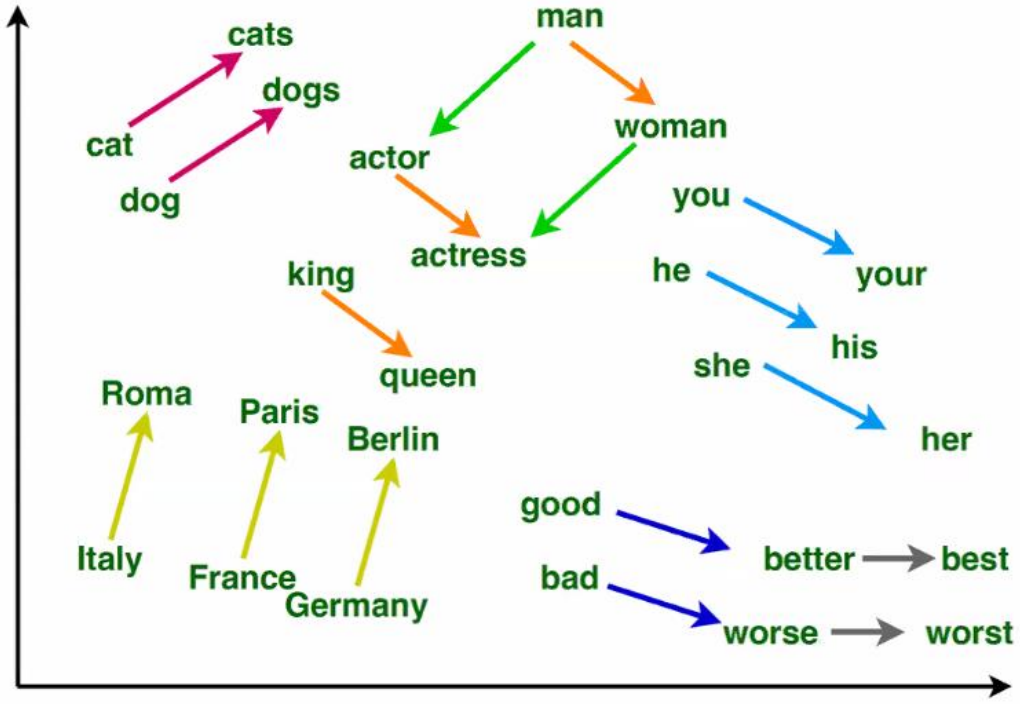


source:Formation d'Introduction au Deep Learning FIDLE
https://www.fidle.cnrs.fr/w3/

* La représentation vectorielle, à la base de tout

Technique “d’embedding” ou plongement lexical

From Bag of Words to Vector Representations [2008, 2013, 2016]

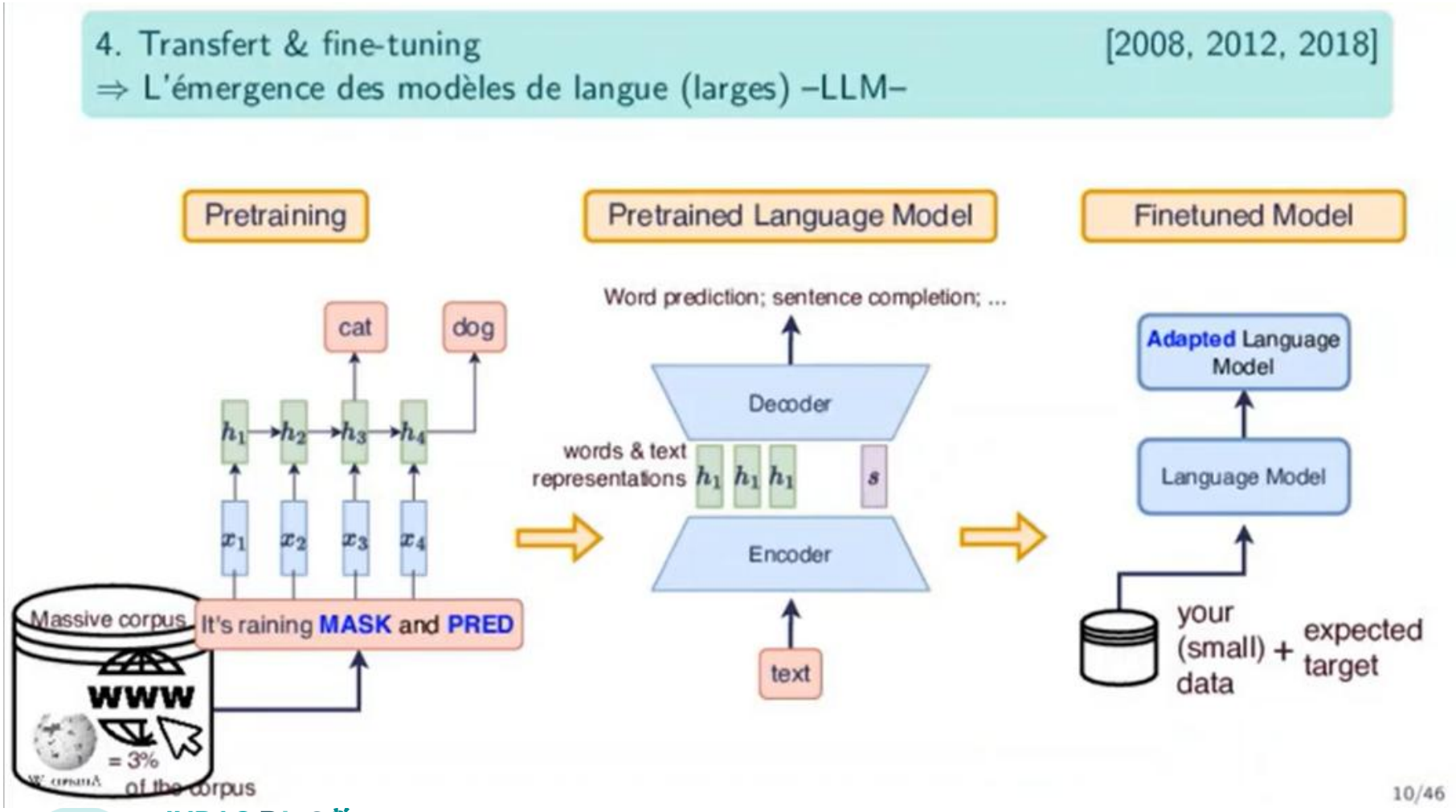


- **Espace sémantique :**
  sens similaires
  <==>
  positions proches
- **Espace structurel :**
  régularités grammaticales,
  connaissances
  linguistiques de base

Distributed representations of words and phrases and their compositionality, Mikolov et al. NeurIPS 2013

* Le modèle Transformer : Encoder-Decoder

Architecture LLM “autorégressive” : rétropropagation



Source : Vincent Guigue-AgroParisTech

Exemples d'entrée et de sortie

Le fonctionnement des outils d'IA générative se compose généralement d'une **entrée** (« input ») par l'utilisateur (une requête, ou « prompt » en anglais) et d'une **sortie** (« output ») générée par l'outil en question.

- Texte à texte
- Texte à image
- Texte à schéma
- Voix à texte, tableau liste
- Texte et documents à présentation, texte, image

* Exemples de tâches

Avec des contenus textuels

- **Résumé/synthèse/comparaison** de documents
- **Classification** : **trier des informations** selon des catégories (ex: spam/mail), catégorisation de contenu
- **Conversation et brainstorming** : **résoudre des problèmes techniques**, concevoir un cours, de la documentation technique
- **Rédaction** : **écrire** des lettres de motivation, des comptes-rendus de réunion, **reformuler** des phrases, **personnaliser** des messages, **créer** du code informatique
- **Traduction** un texte d'**une langue à l'autre**, de la **langue naturelle** vers **un langage structuré** (et vice versa)
- **Extraction d'information** à partir de sites web, de PDF, d'articles scientifiques… **extraire et structurer des informations précises**

Grand modèle de langage (LLM) : Tout ce qu'il faut savoir

* L'IA Générative : un moteur de recherche ?

La recherche avancée ou « Deep Search »

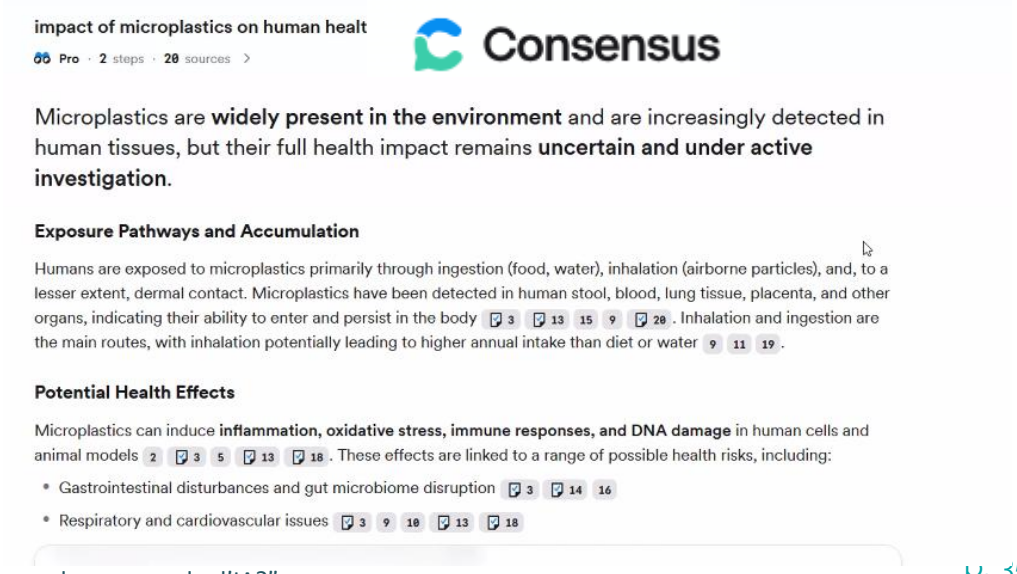
Objectifs :

- lire plusieurs centaines de pages Web en quelques minutes et en faire une **sélection** selon la requête et le besoin initial
- obtenir une **synthèse** personnalisée plutôt qu'une simple liste de liens

Le système accède à des informations via des moteurs de recherche (Google, Bing, etc.) et génère des contenus nouveaux (textes, images, code, etc.) à partir d'une requête.

Sider

perplexity



Vers le meilleur des deux Mondes ?

> WE ARE INTELLIGENT NOT BECAUSE WE CAN DO ONE THING PARTICULARLY WELL, BUT BECAUSE WE CAN LEARN TO DO PRACTICALLY ANYTHING.

https://sks.ai/

* Avantages des IA génératives

- **Automatisation, rapidité** : se libérer des tâches chronophages pour s'atteler à des activités plus valorisantes qui requièrent une véritable expertise humaine.
- **Adaptation au langage naturel** : en paramétrant finement et en posant bien les requêtes, on peut obtenir des résultats correspondants au style, au niveau d'information, à la longueur… désirés
- **Traitement massif des données** : en traitant de grandes quantités de données, les LLM améliorent la précision des tâches de prédiction et de classification.
- **Créativité et génération de contenus** : suggestions de mots-clefs, traductions sémantiques

* Les défis à relever

Ou quelques limites à avoir à l'esprit

- **Opacité des modèles : Manque d'explicabilité/interprétabilité**
- **Manque de véracité /fiabilité :** Ils peuvent produire de fausses informations, des préjugés, voire un langage toxique. **Hallucinations et erreurs**.
- **Manque de stabilité :** Ils ne répondent pas toujours de la même manière
- **La fenêtre contextuelle :** chaque LLM ne dispose que d'une certaine **quantité de mémoire**. Au-delà d'un certain nombre de tokens en entrée, ils ne pourront plus réaliser les tâches demandées.
- **Coûts** : le développement de LLM nécessite des **investissements très important**s (systèmes informatiques, capital humain, énergie…).
- **Impact environnemental** : pour fonctionner, les projets LLM utilisent des centaines de serveurs qui consomment une **énorme quantité d'énergie** et ont une empreinte carbone considérable.

Source : Grand modèle de langage (LLM) : Tout ce qu'il faut savoir

* Les atouts d'un thésaurus

A quelles limites des LLMs ils répondent

- **Contrôle terminologique**
  - Réduction de l'ambiguïté
  - Indexation cohérente, facilitant la recherche documentaire
- **Pérennité et traçabilité**
  - Transparence des choix
  - Évolutions maîtrisées, historique des termes
- **Résultats pertinents et stables / coût amorti sur la durée**
  - Ressource utilisable dans des systèmes d'information partagés (standards SKOS, RDF -> graphes de connaissance)
- **Structuration de la connaissance, relations entre les concepts/notions**

➢ Interopérabilité & réutilisabilité
➢ Sobriété, durabilité

* Conclusion

🎯 Les thésaurus restent pertinents et complémentaires aux LLMs

Une ressource :

- construite, accessible, réutilisable : les identifiants des concepts comme outil d'interopérabilité
- explicite: compréhensible par les humains, vérifiable (vs hallucinations des LLM)
- spécialisée dans un domaine de la connaissance ou une discipline, et validée
- sobre et durable : il est facile et informatiquement peu coûteux d'intégrer de nouvelles connaissances dans un thésaurus

Une ressource pour les agents conversationnels :

- Les thésaurus ne sont pas obsolètes : ils deviennent des piliers de fiabilité
- L'IA est un outil puissant mais nécessite des repères sémantiques

> Front de science : il y a besoin de **spécialiser les LLMs** (fine tuning)

Ensemble, ils offrent une synergie précieuse pour les professionnels de l'information

INRAE DipSO

Vous êtes prêts pour la suite ?