

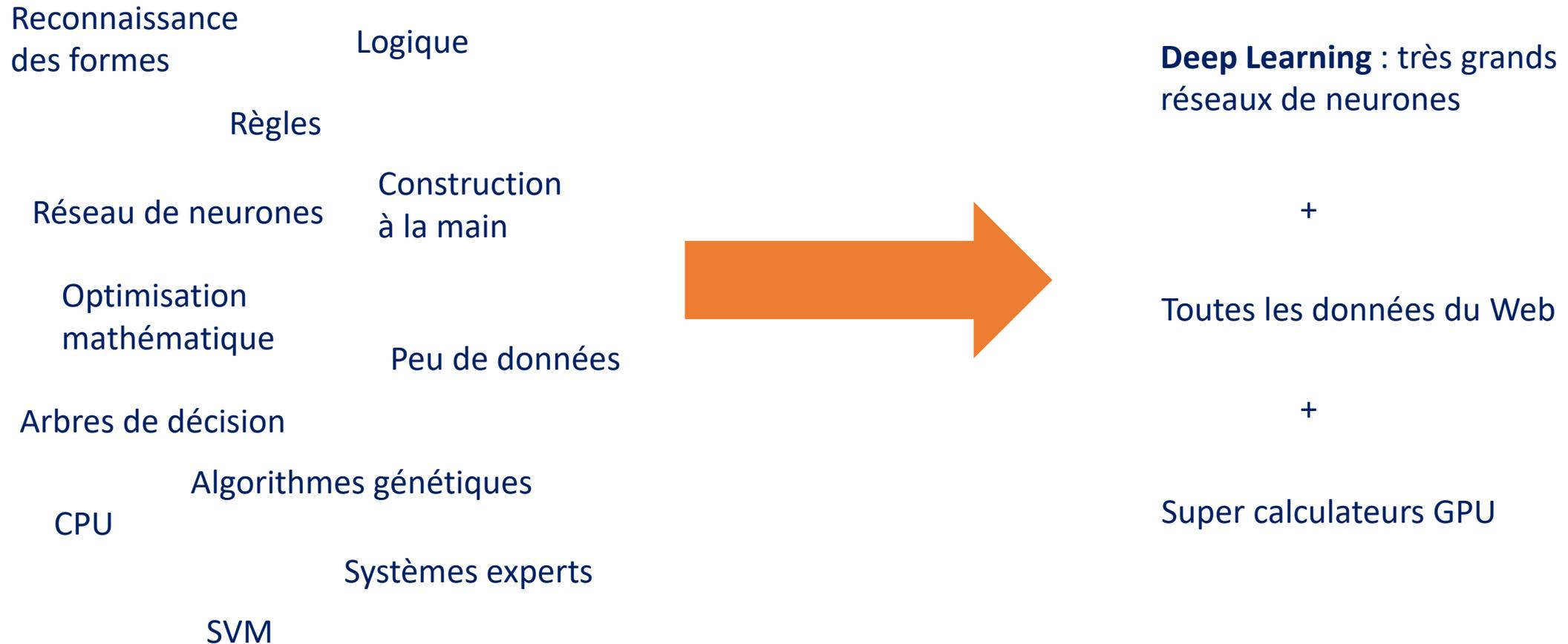
L'IA au service de l'indexation des contenus en bibliothèque

Christopher Kermorvant

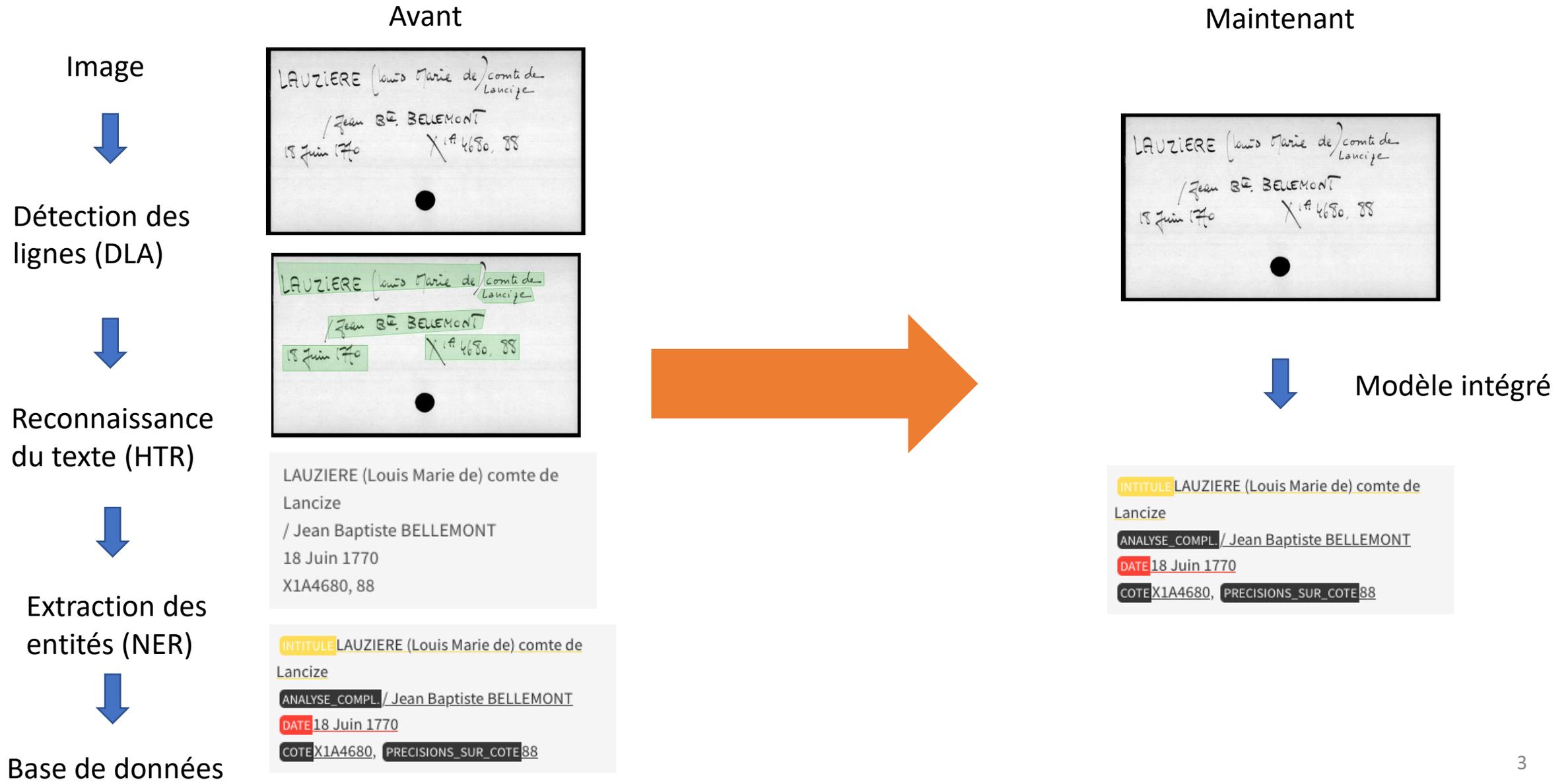
Biennale du numérique, ENSSIB, novembre 2023

T E K L I A

L'IA depuis 30 ans : une concentration technologique



L'OCR/HTR depuis 5 ans : les modèles intégrés



Modèles de langue (LLM, Foundation Models)

Grand réseaux de neurones, entraînés sur l'intégralité du web, pour la **génération de texte** (modèles génératifs) permettant la rédaction de textes, la réécriture, la traduction, la génération de réponse à une question, la génération de code, etc.

 **You**

Lis cette page web <https://www.enssib.fr/biennale-du-numerique-2023-intelligence-artificielle> et propose 5 sujets pour la Table Ronde 3 - Usages professionnels #1 : archivistique et édition. Écris en français.

 **ChatGPT**

Voici cinq sujets potentiels pour la Table Ronde 3 - Usages professionnels #1 : archivistique et édition, dans le cadre de la Biennale du Numérique 2023 sur l'intelligence artificielle à l'Esssib :

1. **L'impact de l'IA sur les pratiques archivistiques modernes** : Comment l'intelligence artificielle transforme la gestion, la conservation, et le catalogage des archives.
2. **L'IA dans le secteur de l'édition** : Exploration du rôle de l'intelligence artificielle dans l'édition moderne, notamment dans la création de contenu assistée par IA et l'analyse prédictive des tendances du marché.
3. **Considérations éthiques dans l'utilisation de l'IA en archivistique et en édition** : Discussion sur les implications éthiques, y compris les préoccupations en matière de confidentialité et les risques de biais dans les algorithmes de l'IA.
4. **Innovations dans les flux de travail en archivistique et en édition grâce à l'IA** : Comment les technologies d'intelligence artificielle peuvent-elles améliorer l'efficacité et apporter de nouvelles capacités dans ces domaines.
5. **Tendances futures et défis de l'IA dans les domaines de l'archivistique et de l'édition** : Spéculation sur les développements futurs de l'IA qui pourraient impacter ces domaines et discussion sur les défis que les professionnels pourraient rencontrer pour s'adapter à ces changements [1].



Modèles de langue (*LLM, Foundation Models*)

Ces modèles doivent être « alignés », c'est-à-dire éduqués pour satisfaire les besoins de l'utilisateur et garantir la sécurité et l'éthique des interactions

L'objectif peut être :

La réalisation d'une tâche unique très spécialisée (*fine tuning*)

➤ *Structurer un catalogue imprimé*

Une interaction en langue naturelle pour réaliser de multiples tâches (*chat/instruct*)

➤ *Dialoguer avec des utilisateurs*



2001: A Space Odyssey

Conversion d'un catalogue : modèles spécialisés

Images



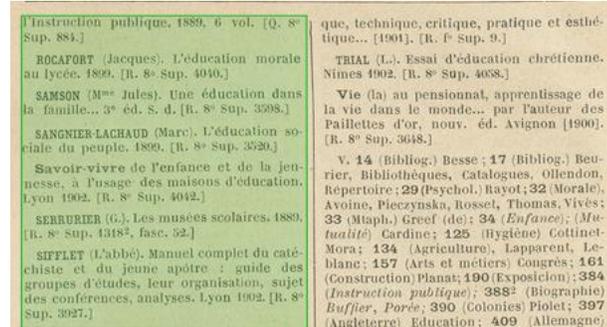
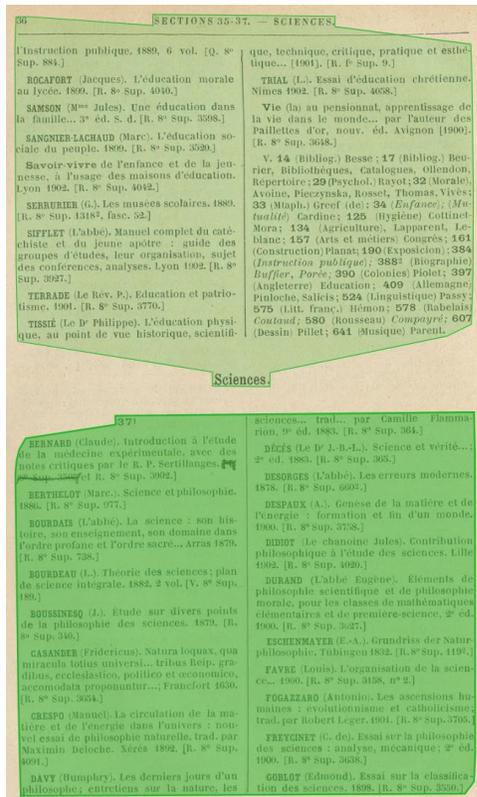
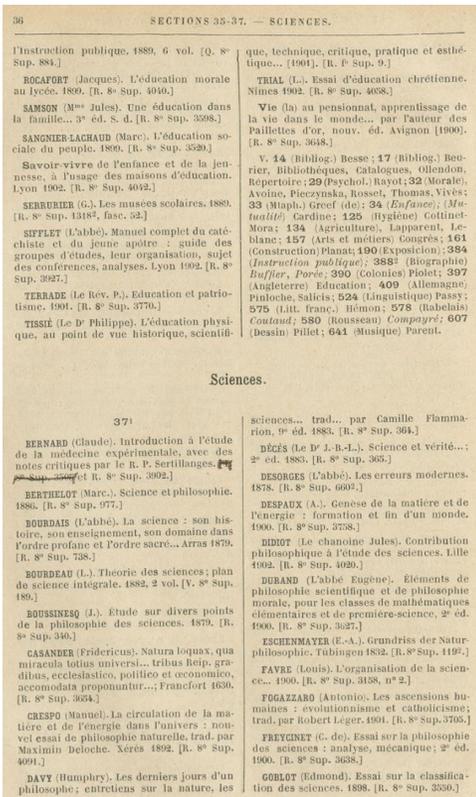
Extraction des sections



Extraction des paragraphes



OCR



L'Instruction publique. 1889, 6 vol. [Q. 8° Sup. 884.]
 ROCAFORT (Jacques). L'éducation morale au lycée. 1899. [R. 8° Sup. 4040.]
 SAMSON (Mme Jules). Une éducation dans la famille... 3^e éd. S. d. [R. 8° Sup. 3598.]
 SANGNIER-LACHAUD (Marc). L'éducation sociale du peuple. 1899. [R. 8° Sup. 3520.]
 Savoir-vivre de l'enfance et de la jeunesse, à l'usage des maisons d'éducation. Lyon 1902. [R. 8° Sup. 4042.]
 SERRURIER (G.). Les musées scolaires. 1889. [R. 8° Sup. 1348², fasc. 52.]
 SIFFLETT (L'abbé). Manuel complet du catéchiste et du jeune apôtre : guide des groupes d'études, leur organisation, sujet des conférences, analyses. Lyon 1902. [R. 8° Sup. 3927.]
 TERRADE (Le Rev. P.). Education et patriotisme. 1901. [R. 8° Sup. 3770.]
 TISSIÉ (Le D^r Philippe). L'éducation physique, au point de vue historique, scientifique, technique, critique, pratique et esthétique... [1901]. [R. 1^{er} Sup. 9.]
 TRIAL (L.). Essai d'éducation chrétienne. Nîmes 1902. [R. 8° Sup. 4068.]
 Vie (la) au pensionnat, apprentissage de la vie dans le monde... par l'auteur des Paillettes d'or, nouv. éd. Avignon [1900]. [R. 8° Sup. 3648.]
 V. 14 (Bibliog.) Besse; 17 (Bibliog.) Beurier, Bibliothèques, Catalogues, Ollendon, Répertoire; 29 (Psychol.) Rayot; 32 (Morale), Avoine, Pieczynska, Rosset, Thomas, Vivès; 33 (Maph.) Gref (de); 34 (Enfance); (Mutualité) Gardine; 125 (Hygiène) Cottinet-Mora; 134 (Agriculture), Lapparent, Leblanc; 157 (Arts et métiers) Congrès; 161 (Construction) Planat; 190 (Exposition); 354 (Instruction publique); 388² (Biographie) Buffler, Porée; 390 (Colonies) Piolet; 397 (Angleterre) Education; 409 (Allemagne) Plüschke, Salcis; 524 (Linguistique) Passy; 575 (Lit. franc.) Hémon; 578 (Rabelais) Coutaud; 580 (Rousseau) Compayré; 607 (Dessin) Pilet; 641 (Musique) Parent.



NER

ROCAFORT (Jacques)	Auteur	L'éducation morale au lycée.	Intitulé	1899	Date de publication	[R. 8° Sup. 4040. Cote]
SAMSON (Mme Jules)	Auteur	Une éducation dans la famille ... 3e éd. S. d.	Intitulé	[R. 8° Sup. 3598. Cote]		

Conversion d'un catalogue : modèle généraliste

You

Here is the OCR of a printed catalogue of the bibliothèque Sainte Geneviève in Paris. It contains a list of books formatted with their author in capital letter, title, place, date and reference number between brackets. Make a list in json of all the books, give their author, title and reference without the brackets, on entry per book

"Bibliographie agronomique. Paris 1810.

[8° Q1. 2203.]

BOEHMERI (G. R.) Bibliotheca scriptorum historia naturalis, æconomia ... Lipsiae 1785-1789. 8 vol. [8° Q1. 2732.]

BOLDUANUS (P.) Bibliotheca philosophica. Ienæ 1616. [4° Qb. 137.]

BONVOULOIR (H. de). Catalogue de la bibliothèque de la Société entomologique de France (1832-1866). Paris 1867,[8° Qb. 213 Sup.]

BORELLIUS (P.) Bibliotheca chimica. Parisiis 1654. [8° Qb. 279.]

Catalogue of scientific papers. London 1867-1879. 8 vol. [8° ÆA. 3 Sup.]

FRISIUS (Jacob). Bibliotheca philosophorum classicorum authorum chronologica. Tiguri 1592. [4° Q. 1403.]

Open-LLM

Here is the list of books from the Bibliothèque Sainte Geneviève's catalogue formatted in JSON, including the author, title, and reference for each entry:

```
json Copy code
[
  {
    "author": "Unknown",
    "title": "Bibliographie agronomique",
    "place": "Paris",
    "date": "1810",
    "reference": "8° Q1. 2203"
  },
  {
    "author": "BOEHMERI, G. R.",
    "title": "Bibliotheca scriptorum historia naturalis, æconomia",
    "place": "Lipsiae",
    "date": "1785-1789",
    "reference": "8° Q1. 2732"
  },
  {
    "author": "BOLDUANUS, P.",
    "title": "Bibliotheca philosophica",
    "place": "Ienæ",
    "date": "1616",
    "reference": "4° Qb. 137"
  },
]
```

<https://teklia.com/fr/blog/cataloguing-with-AI-at-BSG/>

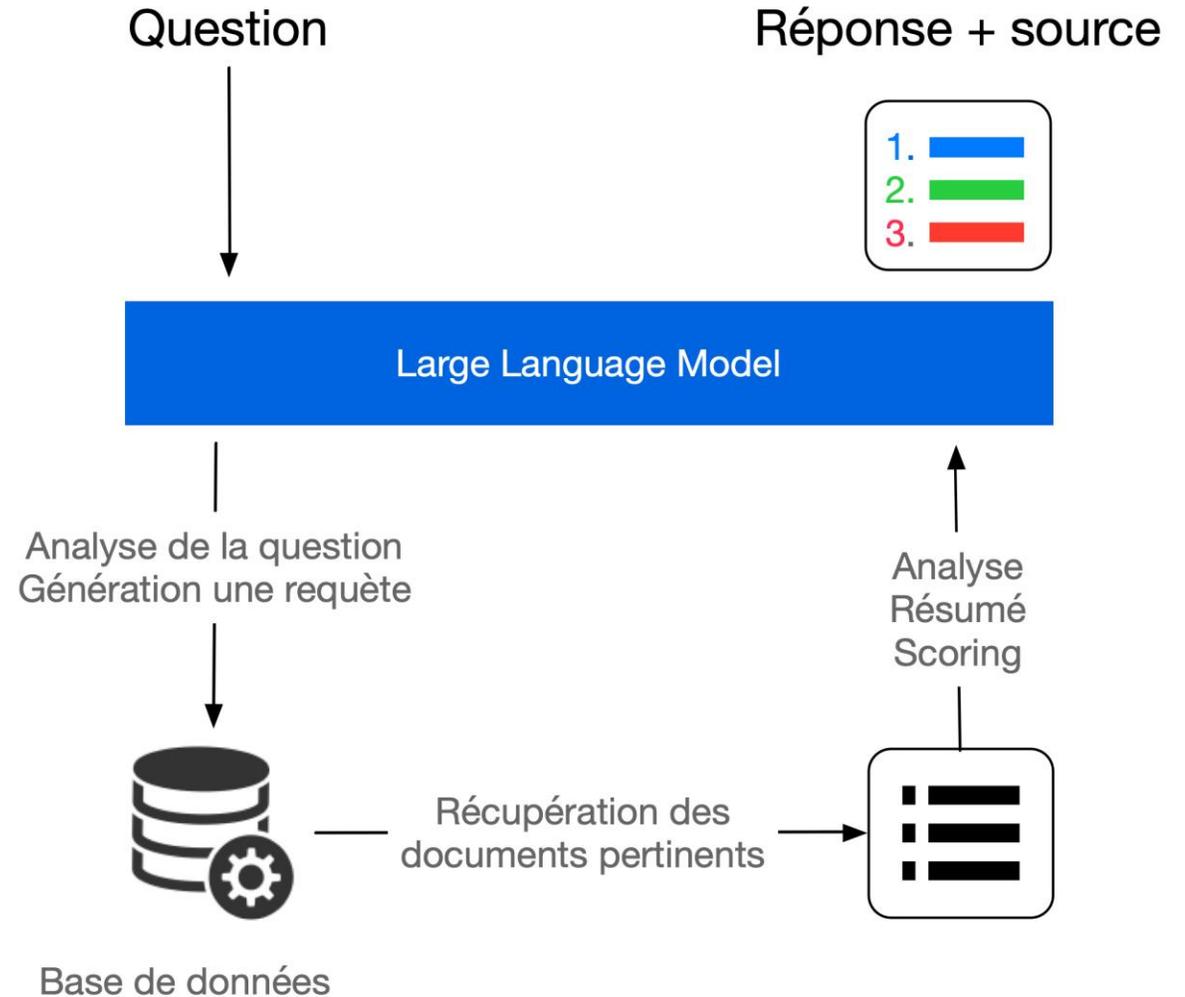
Limites des LLMs

- Ils ne sont pas à jour : ils ne connaissent que leur ensemble d'entraînement **mais peuvent être connectés à une base de donnée**
- Ils inventent ou hallucinent
 - Ils n'ont pas de base de connaissance ≠ moteurs de recherche
 - Ils n'ont pas de logique programmée ≠ systèmes déductifs
 - Ils n'ont pas d'algorithme programmé ≠ calculateurs
- Ils ne sont pas déterministes et des différences mineures dans l'input peut générer des outputs très différents

Comme tous les systèmes, il faut les encadrer et les contrôler

Retrieval-Augmented Generation (RAG)

Combiner les capacités de compréhension et de génération de réponses des LLM avec une base de données



Modèles IA pour les GLAMs

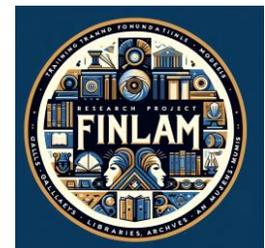
Les modèles open-sources généralistes sont entraînés sur l'intégralité du Web : il n'y a plus de données de bonne qualité librement disponible

- Les organisations qui détiennent ou produisent du contenu de qualité et contrôlé auront un atout

Les modèles généralistes peuvent être « programmés » en langage naturel

- Les experts métiers ou utilisateurs finaux peuvent les utiliser sans les techniciens de l'IA

Projet FINLAM
ANR BNF - LITIS - TEKLIA



T E K L I A

kermorvant@tekli.com

kermorvant@tekli.com