ChatGPT

révolution ou menace informationnelle ?

Revolution oder Bedrohung der Information?

Eymeric Manzinali, décembre 2024
manzinali@unistra.fr
bu.unistra.fr
spokus.eu

1. Comment ChatGPT produit de l'information ?

2. Risques de mésinformation et de désinformation

3. Retrieval-Augmented Generation

1. Wie produziert ChatGPT Informationen?

2. Gefahr von Fehlinformationen und Desinformation

3. Retrieval-Augmented Generation (RAG)

ChatGPT produit de l'information à partir :

→ d'un **modèle de langage** (GPT-4)

→ d'un **modèle conversationnel** (InstructGPT)

ChatGPT erzeugt Informationen basierend auf:

→ einem **Sprachmodell** (GPT-4)

→ einem **Konversationsmodell** (InstructGPT)

(Langlais, 2023)

Generative Pre-trained Transformer 4

IA générative, capable de **produire du contenu** nouveau et original à partir de ses données d'entraînement

Generative KI, die in der Lage ist, neuen und originalen Inhalt auf der Grundlage ihrer Trainingsdaten zu erstellen.

Generative **Pre-trained** Transformer 4

Pré-entraînée à partir d'un **corpus de textes**

Vortrainiert durch einen **Textkorpus:**

→ **3 % intégralité de Wikipédia**

→ **16 % de livres**

→ **80 % archive du Web**

→ **3 % ganz Wikipedia**

→ **16 % Bücher**

→ **80 % Webarchiv**

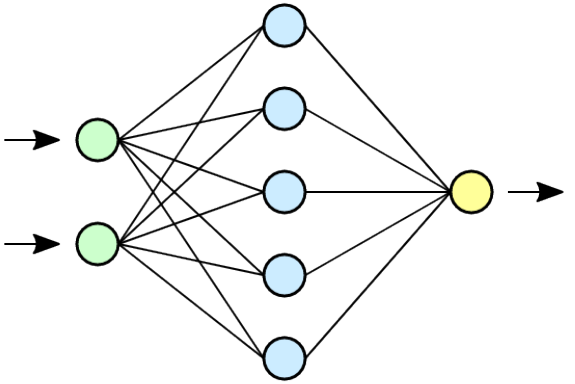
90 % des sources en anglais

90 % der Quellen in Englisch

(Geffen, 2023)

Generative Pre-trained **Transformer** 4

Modèle spécifique de **réseau de neurones**, utilisé pour **enregistrer les relations entre les mots** dans un texte



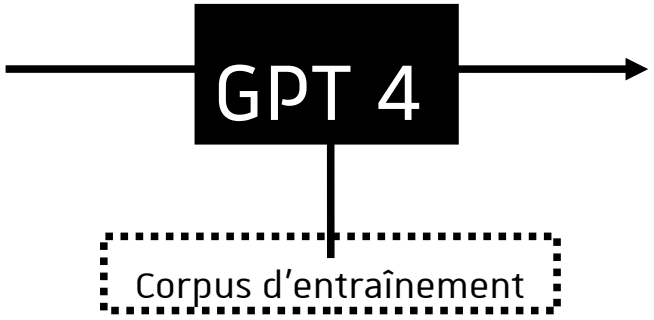
Spezifisches Modell eines **neuronalen Netzwerks**, das verwendet wird, um die Beziehungen zwischen Wörtern in einem Text zu erfassen.

Explication par Arte / Erklärung von Arte

Generative Pre-trained **Transformer** 4

Entrée

« Le Rhin est un … »



Sortie

« … fleuve »

≠ sortie vraie

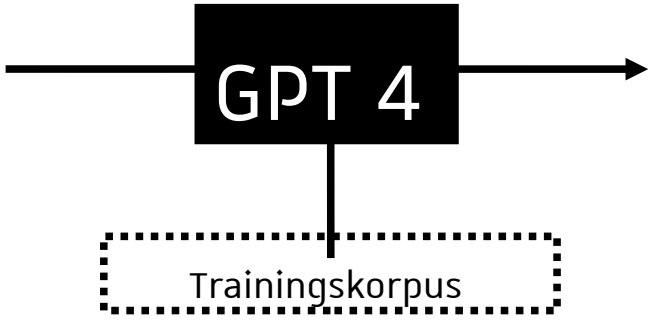
Épistémologie probabiliste
= sortie la + **probable** (et **plausible** pour les humains), d'après son corpus d'entraînement

CAR « Rhin » fréquemment associé à « fleuve » dans ce corpus.

Generative Pre-trained **Transformer** 4

Eingabe

« Der Rhein ist ein … »



Ausgabe

« … Fluss »

≠ wahre Ausgabe

= die **wahrscheinlichste** (und für Menschen **plausible**) Ausgabe laut seinem Trainingskorpus

DA « Rhein » in diesem Korpus häufig mit « Fluss » assoziiert wird.

Generative Pre-trained Transformer 4

Mars 2024 : version 4 de GPT

GPT-4o : amélioration des capacités multimodales de ChatGPT (images, voix, etc.)

März 2024: Version 4 von GPT

GPT-4o: Verbesserung der multimodalen Fähigkeiten von ChatGPT (Bilder, Sprache usw.)

Modèle conversationnel

« Aligner » les réponses de GPT aux attentes des humains

→ générer la réponse la + appropriée en contexte de chat, parmi toutes celles correctes linguistiquement

Konversationsmodell

Die Antworten von GPT auf die Erwartungen der Menschen ausrichten

→ die im Chat-Kontext passendste Antwort generieren, unter allen sprachlich korrekten Möglichkeiten

(Langlais, 2023 ; Ouyang et al., 2023 ; Colin, 2024)

Modèle conversationnel

Konversationsmodell

des humains écrivent des exemples de réponses attendues à partir de prompts

Menschen schreiben Beispiele für erwartete Antworten basierend auf Prompts

→ *fine tuning (Feinabstimmung)*

des humains évaluent et classent les réponses produites par l'IA

Menschen bewerten und ordnen die von der KI generierten Antworten ein

→ *entraînement d'un système de récompense / Training eines Belohnungssystems*

un préprompt donne des instructions « universelles » à ChatGPT

Ein Pre-Prompt gibt „universelle" Anweisungen an ChatGPT

(Langlais, 2023 ; Ouyang et al., 2023 ; Colin, 2024)

1. Comment ChatGPT produit de l'information ?

2. Risques de mésinformation et de désinformation

3. Retrieval-Augmented Generation

1. Wie produziert ChatGPT Informationen?

2. Gefahr von Fehlinformationen und Desinformation

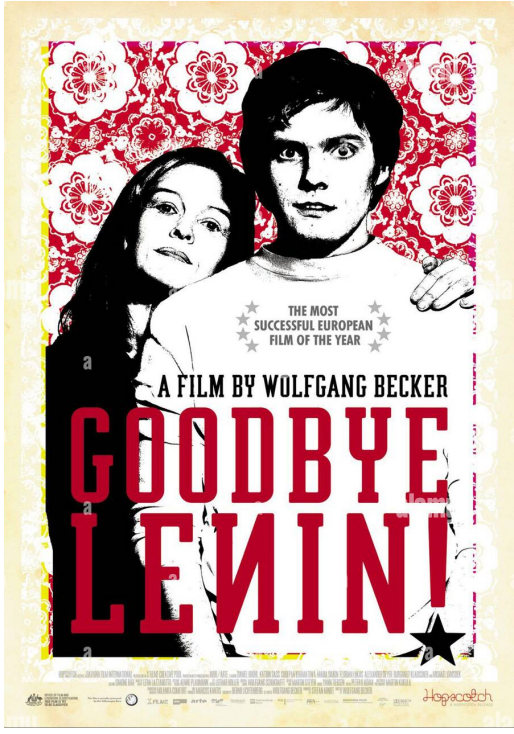
3. Retrieval-Augmented Generation (RAG)

Hallucinations

Halluzinationen

Dans Good-Bye Lenin, quelle marque de soda apparaît sur une publicité affichée dans Berlin-Est, en face de la fenêtre où repose un des personnages principaux ?

In Good Bye Lenin, welche Sodamarke erscheint auf einer Mauer in Ost-Berlin, gegenüber dem Zimmer, in dem eine der Hauptfiguren liegt?



Hallucinations

Halluzinationen

ChatGPT génère une réponse plausible, et basée sur les probabilités, souvent vraie (« *Coca-Cola* »)**, parfois fausse** (« *Sprite* »)**, à partir de ses données d'entraînement.**

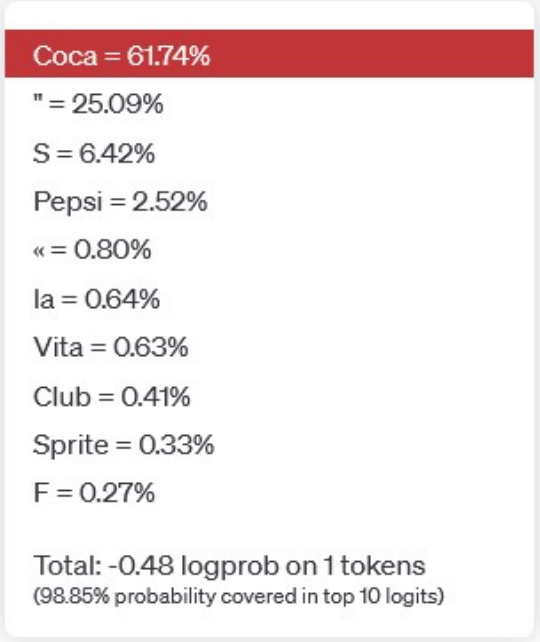
Basierend auf seinen Trainingsdaten generiert ChatGPT eine plausible Antwort, die oft wahr (« *Coca-Cola* ») **jedoch manchmal falsch** (« *Sprite* ») **ist**.

Hallucinations

Halluzinationen

Dans Good-Bye Lenin, quelle marque de soda apparaît sur une publicité affichée dans Berlin-Est, en face de la fenêtre où repose un des personnages principaux ?

La marque de soda qui apparaît sur la publicité est Coca-Cola.



Fausses références

« Cite-moi des livres en français sur le complotisme, avec leur titre, maison d'édition et auteur »

Falsche Referenzen

« Nenne mir französische Bücher über Verschwörungstheorien, mit ihrem Titel, Verlag und Autor »

Fausses références

ChatGPT identifie des **motifs linguistiques** (*patterns*) dans les textes de son corpus et peut les reproduire

A partir de ces motifs, il peut construire de fausses références, qui paraîtront plausibles

Falsche Referenzen

ChatGPT identifiziert **sprachliche Muster** in den Texten seines Korpus und kann diese reproduzieren.

Aus diesen Mustern kann es falsche Referenzen erstellen, die plausibel erscheinen.

Biais

Génération de la réponse la + **probable** … mais source de **stéréotypes** et **réponses conformistes**.

Corpus d'entraînement : sédimentations de textes passés, dont certains reflets des stéréotypes d'une époque.

Prédominance des **sources anglophones.**

Bias

Generierung der **wahrscheinlichsten Antwort** … aber Quelle von **Stereotypen** und **konformistischen Antworten**.

Trainingskorpus: Sedimentierungen vergangener Texte, von denen einige die Stereotypen einer bestimmten Epoche widerspiegeln.

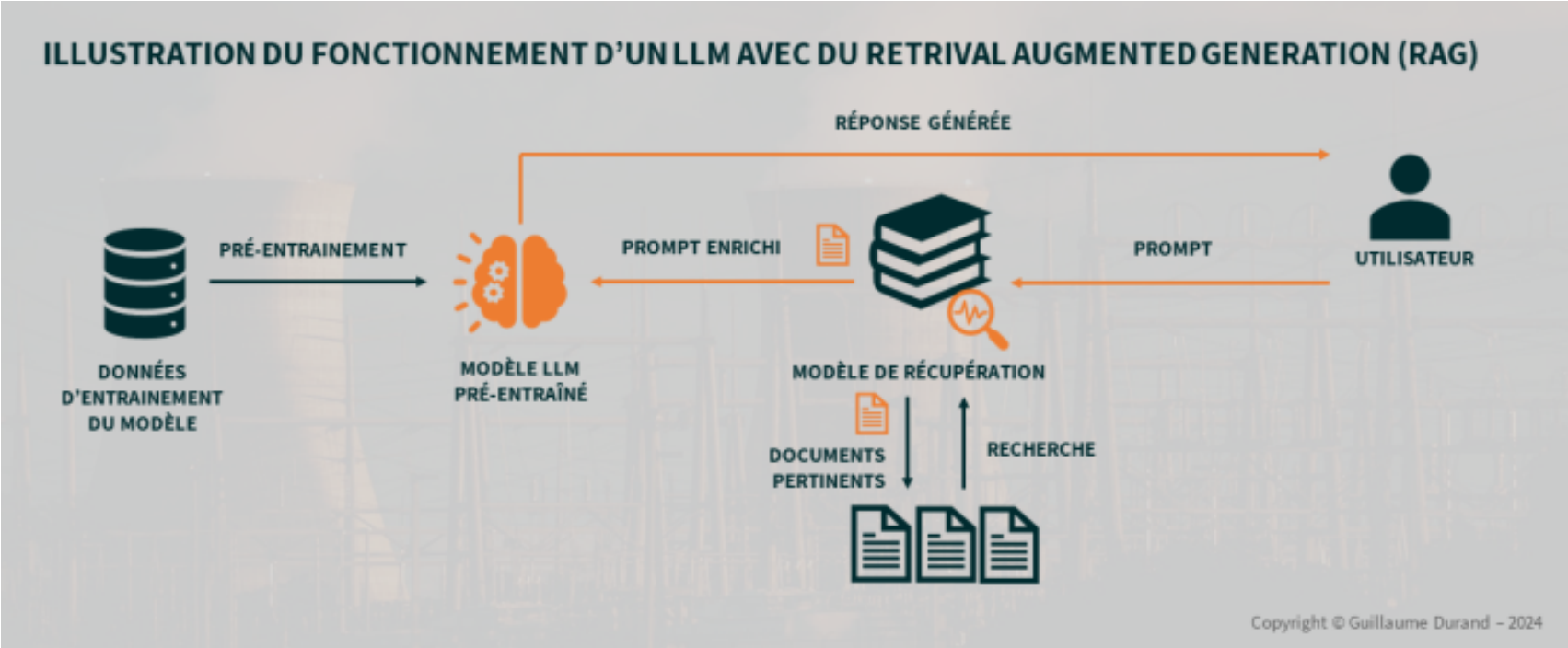
überwiegend **englischsprachige Quellen**.

1. Comment ChatGPT produit de l'information ?
2. Risques de mésinformation et de désinformation
3. Retrieval-Augmented Generation

1. Wie produziert ChatGPT Informationen?
2. Gefahr von Fehlinformationen und Desinformation
3. Retrieval-Augmented Generation

« Le principe du RAG est donc d'**augmenter** (A) la **fonction générative** (G) avec une **fonction de recherche** (R) dans un **corpus externe** »
(Figoblog, 2024)

„Das Prinzip von RAG besteht also **darin**, (A) die **generative Funktion** (G) durch eine **Suchfunktion** (R) in einer externen Datenbank oder Korpus zu erweitern“
(Figoblog, 2024)



Dépasser les limites des modèles de langage, et **améliorer l'expérience** (*synthèse des résultats de recherche, langage naturel*) des **utilisateurs des bases documentaires**.

Die Grenzen von Sprachmodellen überwinden und die Nutzererfahrung (*Zusammenfassung der Suchergebnisse, natürliche Sprache*) für **Anwender von Dokumentationsdatenbanken verbessern**.

Partenariat OpenAI avec

→ **des médias** pour permettre à ChatGPT d'interroger leurs archives afin de sourcer ses réponses

→ **des bibliothèques** pour avoir des réponses en langage naturel basées sur des documents patrimoniaux.

Partnerschaft von OpenAI mit

→ **Medien**, um ChatGPT zu ermöglichen, auf deren Archive zuzugreifen, um seine Antworten zu belegen

→ **Bibliotheken**, um Antworten in natürlicher Sprache basierend auf kulturhistorischen Dokumenten zu liefern.

IA-G + base bibliographique

→ **Consensus**, de courtes « revues de la littérature » rédigées par l'IA, et basées sur les articles de Semantic Scholar.

KI-G + bibliografische Datenbank

→ **Consensus**, kurze „Literaturübersichten", die von der KI verfasst werden und auf Artikeln von Semantic Scholar basieren.

Conclusion

+ que la mésinformation, il faut craindre à l'avenir un problème de **désinformation**

→ **mésinformation** de - en - fréquente, avec l'amélioration des modèles et la **RAG**
→ amélioration des IA-G en fait des outils + puissants pour générer de la **désinformation**

Abschluss

Mehr als Fehlinformation sollte man in Zukunft ein Problem der **Desinformation** befürchten

→ **Fehlinformation** wird mit der Verbesserung der Modelle und der RAG immer seltener

→ Die Verbesserung von KI-G macht diese Tools jedoch immer leistungsfähiger zur Erzeugung von **Desinformation**.

Merci pour votre attention !

Vielen Dank für Ihre Aufmerksamkeit!

Eymeric Manzinali, décembre 2024
manzinali@unistra.fr
bu.unistra.fr
spokus.eu