

La génération artificielle de contenus

Images, vidéos, texte, parole :
démystification, usages, éthique et limites



AUGMENT^{IA}
BY FONDATION GRENOBLE INP



Didier Schwab
Titulaire de la chaire
AugmentIA
Professeur à l'UGA



AUGMENT^{IA}

BY FONDATION GRENOBLE INP





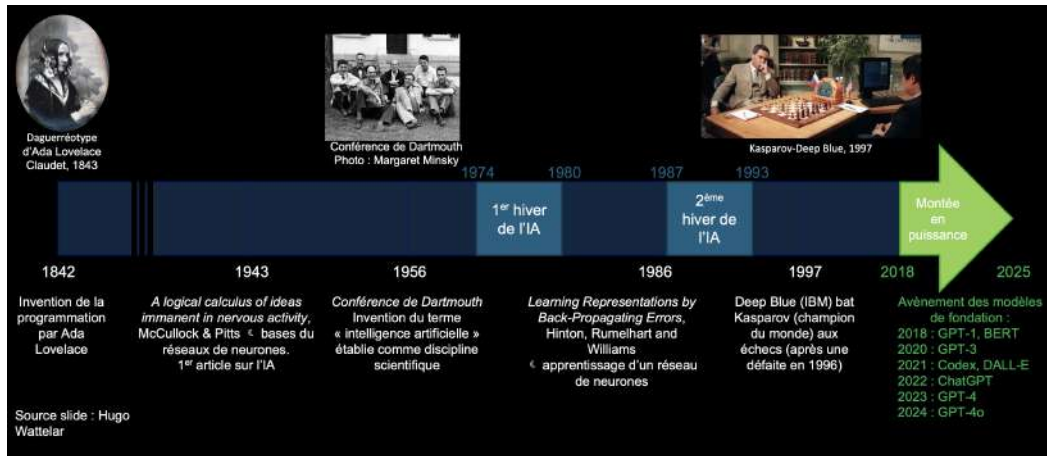
Intelligence Artificielle



**(Très) bref
historique**



Bref historique





L'Intelligence Artificielle autour de nous



Règlement (UE) 2024/1689 du Parlement européen et du Conseil du 13 juin 2024 établissant des règles harmonisées concernant l'intelligence artificielle



Règlement (UE) 2024/1689 du Parlement européen et du Conseil du 13 juin 2024 établissant des règles harmonisées concernant l'intelligence artificielle (**IA Act**)

Règlement (UE) 2024/1689 du Parlement européen et du Conseil du 13 juin 2024 établissant des règles harmonisées concernant l'intelligence artificielle (**IA Act**)

Article 3 (Définitions)



Règlement (UE) 2024/1689 du Parlement européen et du Conseil du 13 juin 2024 établissant des règles harmonisées concernant l'intelligence artificielle (**IA Act**)

Article 3 (Définitions)

Alinéa 1 : « **Système d'IA** » « *un système automatisé qui est conçu pour fonctionner à différents niveaux d'autonomie et peut faire preuve d'une capacité d'adaptation après son déploiement, et qui, pour des objectifs explicites ou implicites, déduit, à partir des entrées qu'il reçoit, la manière de générer des sorties telles que des prédictions, du contenu, des recommandations ou des décisions qui peuvent influencer les environnements physiques ou virtuels; »*

Qu'appelle-t-on Intelligence Artificielle prédictive ?



Qu'appelle-t-on Intelligence Artificielle prédictive ?

Des programmes informatiques conçus pour prédire des résultats à partir de données



Qu'appelle-t-on Intelligence Artificielle prédictive ?

Des programmes informatiques conçus pour prédire des résultats à partir de données



Qu'appelle-t-on Intelligence Artificielle prédictive ?

Des programmes informatiques conçus pour prédire des résultats à partir de données



Qu'appelle-t-on Intelligence Artificielle prédictive ?

Des programmes informatiques conçus pour prédire des résultats à partir de données



Qu'appelle-t-on Intelligence Artificielle Générative ?



Qu'appelle-t-on Intelligence Artificielle Générative ?



Des programmes informatiques capables de créer du contenu qui pourrait être perçu comme créé par des humains



Qu'appelle-t-on Intelligence Artificielle Générative ?



Des programmes informatiques capables de créer du contenu qui pourrait être perçu comme créé par des humains généralement à partir de prompts.



Qu'appelle-t-on Intelligence Artificielle Générative ?

Des programmes informatiques capables de créer du contenu qui pourrait être perçu comme créé par des humains généralement à partir de prompts.



Qu'appelle-t-on Intelligence Artificielle Générative ?

Des programmes informatiques capables de créer du contenu qui pourrait être perçu comme créé par des humains généralement à partir de prompts.



Qu'appelle-t-on Intelligence Artificielle Générative ?

Des programmes informatiques capables de créer du contenu qui pourrait être perçu comme créé par des humains généralement à partir de prompts.



Qu'appelle-t-on Intelligence Artificielle Générative ?

Des programmes informatiques capables de créer du contenu qui pourrait être perçu comme créé par des humains généralement à partir de prompts.



Qu'appelle-t-on Intelligence Artificielle Générative ?

Des programmes informatiques capables de créer du contenu qui pourrait être perçu comme créé par des humains généralement à partir de prompts.



Qu'appelle-t-on Intelligence Artificielle Générative ?

Des programmes informatiques capables de créer du contenu qui pourrait être perçu comme créé par des humains généralement à partir de prompts.



Et le texte là dedans ?



Et le texte là dedans ?

- Central à ces applications



Et le texte là dedans ?

- Central à ces applications
 - ▶ Textes à prononcer



Et le texte là dedans ?

- Central à ces applications
 - ▶ Textes à prononcer
 - ▶ Prompts



Et le texte là dedans ?

- Central à ces applications
 - ▶ Textes à prononcer
 - ▶ Prompts
- Un prompt (une formulation) est un texte d'entrée fourni à un modèle de langue pour initier ou guider sa génération de texte. Il s'agit généralement de décrire la tâche que l'on souhaite que le modèle effectue.



Et le texte là dedans ?

- Central à ces applications
 - ▶ Textes à prononcer
 - ▶ Prompts
- Un prompt (une formulation) est un texte d'entrée fourni à un modèle de langue pour initier ou guider sa génération de texte. Il s'agit généralement de décrire la tâche que l'on souhaite que le modèle effectue.
- Systèmes d'intelligence Artificielle Conversationnel;s : ChatGPT (OpenAI, 30/11/2022), Gemini (Google), LeChat (Mistral), Claude (Anthropic), HuggingChat (Huggingface), etc.



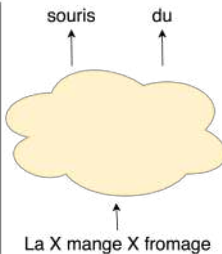
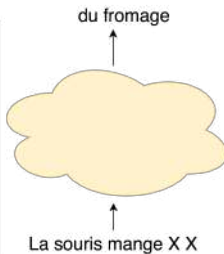
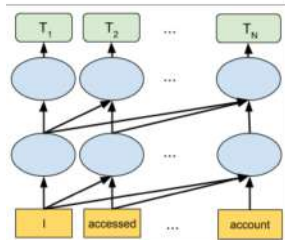
Traitement Automatique des Langues



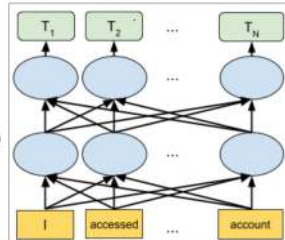
Figure: Source : <https://www.nlplanet.org/>

Autoapprentissage : décodeurs vs encodeurs

Décodeur (ex : GPT)

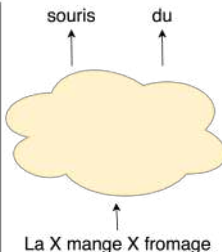
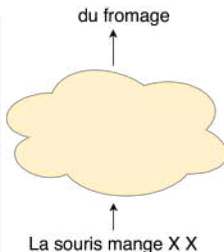
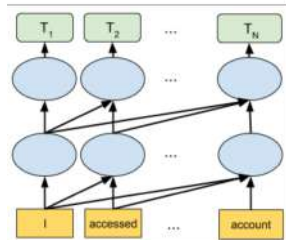


Encodeur (ex : BERT)

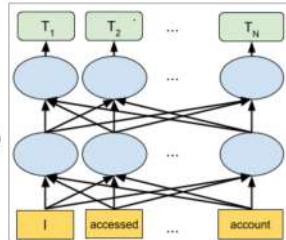


Autoapprentissage : décodeurs vs encodeurs

Décodeur (ex : GPT)



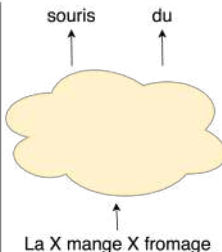
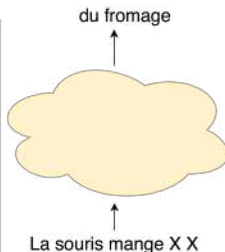
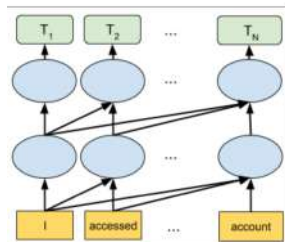
Encodeur (ex : BERT)



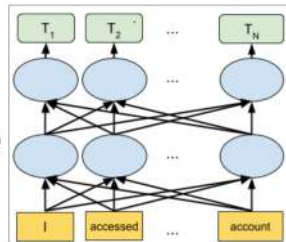
- Type d'apprentissage non-supervisé (entrée : données dégradées ; sortie : données reconstituées)

Autoapprentissage : décodeurs vs encodeurs

Décodeur (ex : GPT)

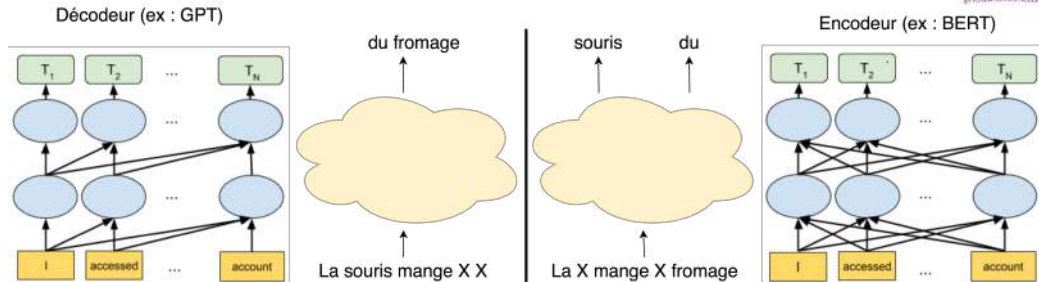


Encodeur (ex : BERT)



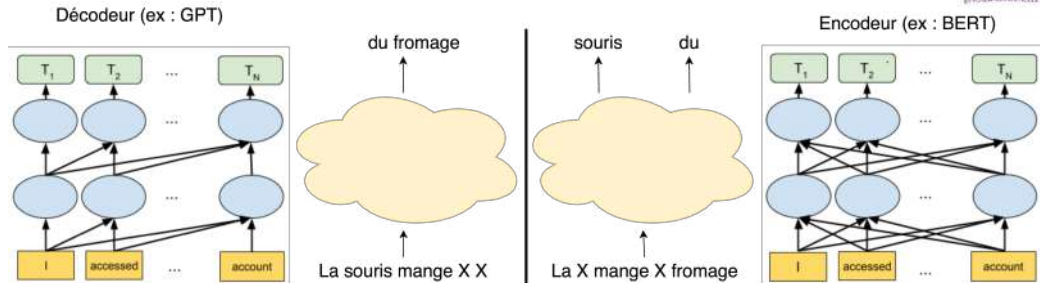
- Type d'apprentissage non-supervisé (entrée : données dégradées ; sortie : données reconstituées)
- Modèles décodeurs : destinés à générer des textes cohérents à partir d'une amorce (prompt) – génératifs

Autoapprentissage : décodeurs vs encodeurs



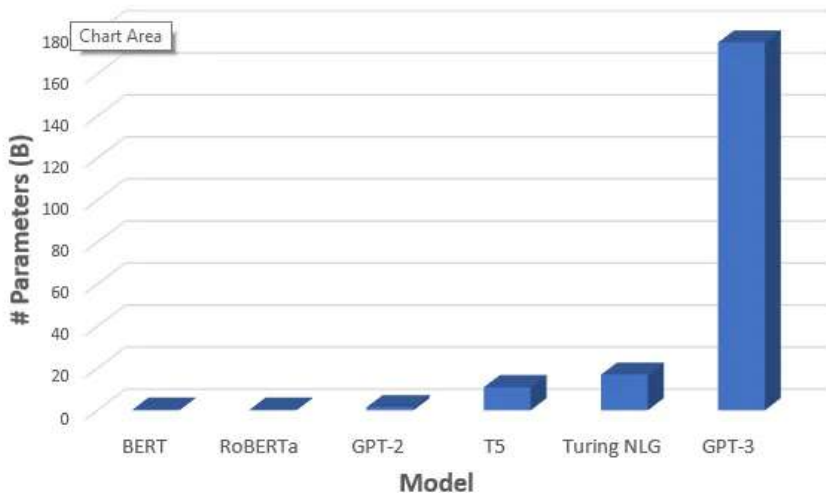
- Type d'apprentissage non-supervisé (entrée : données dégradées ; sortie : données reconstituées)
- Modèles décodeurs : destinés à générer des textes cohérents à partir d'une amorce (prompt) – génératifs
- Modèles encodeurs : destinés à construire de bonnes représentations de mots, de phrases ou de documents – prédictifs

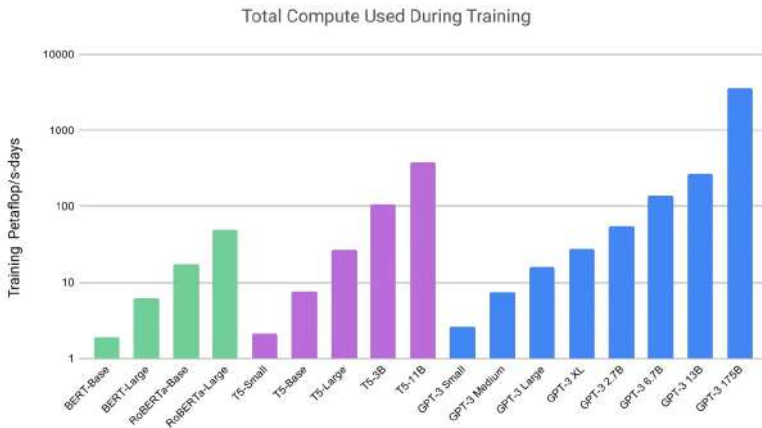
Autoapprentissage : décodeurs vs encodeurs



- Type d'apprentissage non-supervisé (entrée : données dégradées ; sortie : données reconstituées)
- Modèles décodeurs : destinés à générer des textes cohérents à partir d'une amorce (prompt) – génératifs
- Modèles encodeurs : destinés à construire de bonnes représentations de mots, de phrases ou de documents – prédictifs

Nombre de paramètres





Un pétaflop/s-jour (pfs-jour) consiste à effectuer 10^{15} opérations de réseau neuronal par seconde pendant un jour, soit un total d'environ 10^{20} opérations (estimation)¹.

¹<https://openai.com/research/ai-and-compute>

Le concept de prompt en traitement des langues



Le concept de prompt en traitement des langues

Un prompt (une formulation) est un texte d'entrée fourni à un modèle de langage pour initier ou guider sa génération de texte. Il s'agit généralement d'une phrase ou d'une question qui décrit la tâche que l'on souhaite que le modèle effectue.



Le concept de prompt en traitement des langues

Un prompt (une formulation) est un texte d'entrée fourni à un modèle de langage pour initier ou guider sa génération de texte. Il s'agit généralement d'une phrase ou d'une question qui décrit la tâche que l'on souhaite que le modèle effectue.

- Génération de texte :
 - ▶ "Il était une fois, dans un lointain royaume, un jeune prince qui..."
 - ▶ "Rédigez un paragraphe de 100 mots sur le thème de..."



Le concept de prompt en traitement des langues

Un prompt (une formulation) est un texte d'entrée fourni à un modèle de langage pour initier ou guider sa génération de texte. Il s'agit généralement d'une phrase ou d'une question qui décrit la tâche que l'on souhaite que le modèle effectue.

- Génération de texte :
 - ▶ "Il était une fois, dans un lointain royaume, un jeune prince qui..."
 - ▶ "Rédigez un paragraphe de 100 mots sur le thème de..."
- Traduction automatique :
 - ▶ "Traduisez la phrase suivante en anglais :"
 - ▶ "Comment dit-on 'bonjour' en espagnol ?"

Le concept de prompt en traitement des langues

Un prompt (une formulation) est un texte d'entrée fourni à un modèle de langage pour initier ou guider sa génération de texte. Il s'agit généralement d'une phrase ou d'une question qui décrit la tâche que l'on souhaite que le modèle effectue.

- Génération de texte :
 - ▶ "Il était une fois, dans un lointain royaume, un jeune prince qui..."
 - ▶ "Rédigez un paragraphe de 100 mots sur le thème de..."
- Traduction automatique :
 - ▶ "Traduisez la phrase suivante en anglais :"
 - ▶ "Comment dit-on 'bonjour' en espagnol ?"
- Classification de texte :
 - ▶ "Ce texte est-il positif ou négatif ?"
 - ▶ "S'agit-il d'un article de sport ou de politique ?"

Le concept de prompt en traitement des langues

Un prompt (une formulation) est un texte d'entrée fourni à un modèle de langage pour initier ou guider sa génération de texte. Il s'agit généralement d'une phrase ou d'une question qui décrit la tâche que l'on souhaite que le modèle effectue.

- Génération de texte :
 - ▶ "Il était une fois, dans un lointain royaume, un jeune prince qui..."
 - ▶ "Rédigez un paragraphe de 100 mots sur le thème de..."
- Traduction automatique :
 - ▶ "Traduisez la phrase suivante en anglais :"
 - ▶ "Comment dit-on 'bonjour' en espagnol ?"
- Classification de texte :
 - ▶ "Ce texte est-il positif ou négatif ?"
 - ▶ "S'agit-il d'un article de sport ou de politique ?"
- Réponse à des questions :
 - ▶ "Quel est le nom de l'inventeur de la théorie de la relativité ?"
 - ▶ "Combien de membres compte le Parlement européen ?"

Apprentissage des grands modèles de langue à la GPT



Exemple de sortie d'un GPT



Exemple de sortie d'un GPT

- II



Exemple de sortie d'un GPT

- Il
- Il était



Exemple de sortie d'un GPT

- Il
- Il était
- Il était une



Exemple de sortie d'un GPT

- Il
- Il était
- Il était une
- Il était une fois



Exemple de sortie d'un GPT

- Il
- Il était
- Il était une
- Il était une fois
- Il était une fois, il



Exemple de sortie d'un GPT

- Il
- Il était
- Il était une
- Il était une fois
- Il était une fois, il
- Il était une fois, il y



Exemple de sortie d'un GPT

- Il
- Il était
- Il était une
- Il était une fois
- Il était une fois, il
- Il était une fois, il y
- Il était une fois, il y a



Exemple de sortie d'un GPT

- Il
- Il était
- Il était une
- Il était une fois
- Il était une fois, il
- Il était une fois, il y
- Il était une fois, il y a
- Il était une fois, il y a bien

Exemple de sortie d'un GPT

- Il
- Il était
- Il était une
- Il était une fois
- Il était une fois, il
- Il était une fois, il y
- Il était une fois, il y a
- Il était une fois, il y a bien
- Il était une fois, il y a bien longtemps



Exemple de sortie d'un GPT

- Il
- Il était
- Il était une
- Il était une fois
- Il était une fois, il
- Il était une fois, il y
- Il était une fois, il y a
- Il était une fois, il y a bien
- Il était une fois, il y a bien longtemps . . .

Exemple de sortie d'un GPT

- Il
- Il était
- Il était une
- Il était une fois
- Il était une fois, il
- Il était une fois, il y
- Il était une fois, il y a
- Il était une fois, il y a bien
- Il était une fois, il y a bien longtemps . . .

Un GPT est conçu pour prédire le mot suivant à partir d'une séquence de mots donnée (et c'est tout).

Dataset	Quantité (jetons)	% du corpus d'apprentissage
Common Crawl (filtré)	410 milliards	60%
WebText2	19 milliards	22%
Books1	12 milliards	8%
Books2	55 milliards	8%
Wikipedia	3 milliards	3%

Table: Type de nombre de données utilisées pour apprendre GPT-3 [Brown et al., 2020]

Il faudrait à un humain lisant 24h/24 environ 2452 ans pour lire l'intégralité de Common Crawl.

Entraînement préalable des GPT :

(1) Modélisation de la langue



Entraînement préalable des GPT :

(1) Modélisation de la langue

L'entraînement préalable des GPT (dont GPT3) comprend :



Entraînement préalable des GPT :

(1) Modélisation de la langue

L'entraînement préalable des GPT (dont GPT3) comprend :

- La modélisation de la langue (autosupervisé) : le modèle tente de prédire le mot suivant dans une séquence de mots donnée

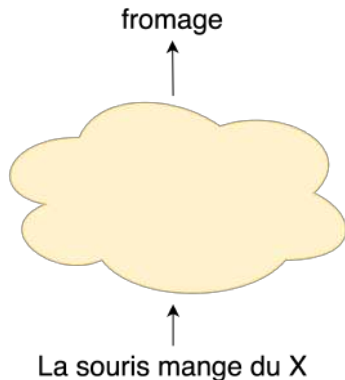


Entraînement préalable des GPT :

(1) Modélisation de la langue

L'entraînement préalable des GPT (dont GPT3) comprend :

- La modélisation de la langue (autosupervisé) : le modèle tente de prédire le mot suivant dans une séquence de mots donnée



Entraînement préalable des GPT :

(2) Résolution de tâche (instructions)



Entraînement préalable des GPT : (2) Résolution de tâche (instructions)

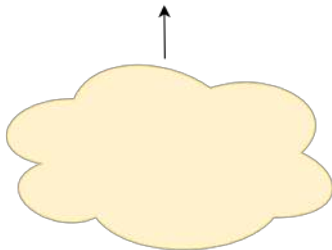
Tâches multiples : Classification de texte, résumé automatique, réponse à des questions, traduction, etc.



Entraînement préalable des GPT : (2) Résolution de tâche (instructions)

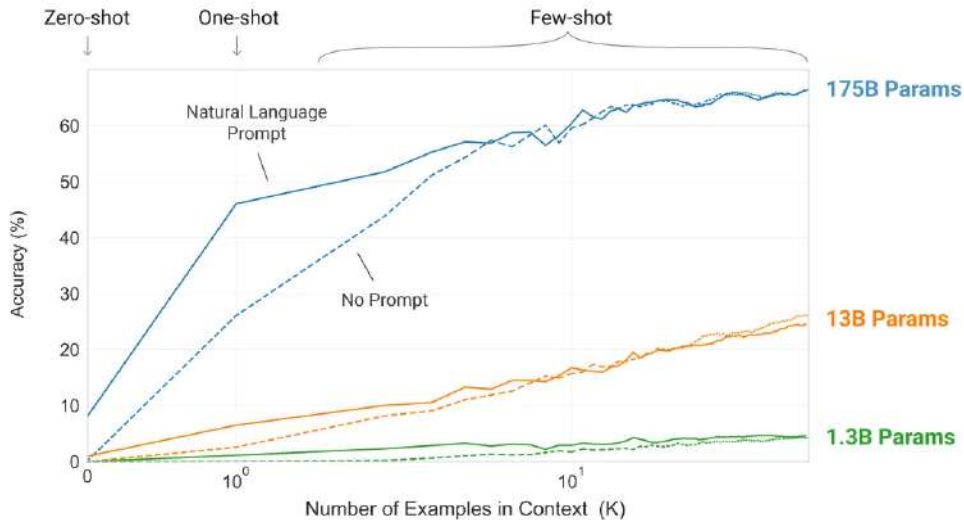
Tâches multiples : Classification de texte, résumé automatique, réponse à des questions, traduction, etc.

La souris/animal est poursuivie par
le chat/animal



désambiguïse la phrase
suivante : La souris est
poursuivie par le chat

Performances GPT3



Entraînement préalable des GPT :

(3) Interactions avec des annotateurs (alignement)



Entraînement préalable des GPT :

(3) Interactions avec des annotateurs (alignement)

- Ces modèles ont interagi avec des annotateurs (*labellers*) chargés de classifier des réponses du système (apprentissage par renforcement)



Entraînement préalable des GPT : (3) Interactions avec des annotateurs (alignement)

- Ces modèles ont interagi avec des annotateurs (*labellers*) chargés de classer des réponses du système (apprentissage par renforcement)
- Appris sur des dialogues, en interagissant avec ses annotateurs



Entraînement préalable des GPT :

(3) Interactions avec des annotateurs (alignement)

- Ces modèles ont interagi avec des annotateurs (*labellers*) chargés de classifier des réponses du système (apprentissage par renforcement)
- Appris sur des dialogues, en interagissant avec ses annotateurs
- On sait que chatGPT est basé sur le même principe qu'instructGPT ([Ouyang et al., 2022])



Entraînement préalable des GPT :

(3) Interactions avec des annotateurs (alignement)

- Ces modèles ont interagi avec des annotateurs (*labellers*) chargés de classifier des réponses du système (apprentissage par renforcement)
- Appris sur des dialogues, en interagissant avec ses annotateurs
- On sait que chatGPT est basé sur le même principe qu'instructGPT ([Ouyang et al., 2022])
- Mais on ignore beaucoup d'informations :



Entraînement préalable des GPT :

(3) Interactions avec des annotateurs (alignement)

- Ces modèles ont interagi avec des annotateurs (*labellers*) chargés de classifier des réponses du système (apprentissage par renforcement)
- Appris sur des dialogues, en interagissant avec ses annotateurs
- On sait que chatGPT est basé sur le même principe qu'instructGPT ([Ouyang et al., 2022])
- Mais on ignore beaucoup d'informations :
 - ▶ Combien d'annotateurs ? Quel profil ?



Entraînement préalable des GPT :

(3) Interactions avec des annotateurs (alignement)

- Ces modèles ont interagi avec des annotateurs (*labellers*) chargés de classifier des réponses du système (apprentissage par renforcement)
- Appris sur des dialogues, en interagissant avec ses annotateurs
- On sait que chatGPT est basé sur le même principe qu'instructGPT ([Ouyang et al., 2022])
- Mais on ignore beaucoup d'informations :
 - ▶ Combien d'annotateurs ? Quel profil ?
 - ▶ On sait que certains étaient kenyans (payés 2\$ par heure [Perrigo, 2023a])

Entraînement préalable des GPT :

(3) Interactions avec des annotateurs (alignement)

- Ces modèles ont interagi avec des annotateurs (*labellers*) chargés de classifier des réponses du système (apprentissage par renforcement)
- Appris sur des dialogues, en interagissant avec ses annotateurs
- On sait que chatGPT est basé sur le même principe qu'instructGPT ([Ouyang et al., 2022])
- Mais on ignore beaucoup d'informations :
 - ▶ Combien d'annotateurs ? Quel profil ?
 - ▶ On sait que certains étaient kenyans (payés 2\$ par heure [Perrigo, 2023a])
 - ▶ On sait qu'il est mis à jour grâce aux retours des utilisateurs (actuellement, 30ème+ version publique : 15/05/2025)

Entraînement préalable des GPT :

(3) Interactions avec des annotateurs (alignement)

- Ces modèles ont interagi avec des annotateurs (*labellers*) chargés de classifier des réponses du système (apprentissage par renforcement)
- Appris sur des dialogues, en interagissant avec ses annotateurs
- On sait que chatGPT est basé sur le même principe qu'instructGPT ([Ouyang et al., 2022])
- Mais on ignore beaucoup d'informations :
 - ▶ Combien d'annotateurs ? Quel profil ?
 - ▶ On sait que certains étaient kenyans (payés 2\$ par heure [Perrigo, 2023a])
 - ▶ On sait qu'il est mis à jour grâce aux retours des utilisateurs (actuellement, 30ème+ version publique : 15/05/2025)
 - ▶ On peut constater que les réponses s'améliorent (exemple : 2+2)

Vers un modèle aligné sur les préférences humaines

Raconte-moi une histoire



Vers un modèle aligné sur les préférences humaines

Raconte-moi une histoire sur un chevalier à l'ère médiévale.



Vers un modèle aligné sur les préférences humaines

Raconte-moi une histoire sur un chevalier à l'ère médiévale.

Raconte-moi une histoire

Vers un modèle aligné sur les préférences humaines

Raconte-moi une histoire sur un chevalier à l'ère médiévale.

Raconte-moi une histoire. Il était une fois ...



Vers un modèle aligné sur les préférences humaines

Raconte-moi une histoire sur un chevalier à l'ère médiévale.

Raconte-moi une histoire. Il était une fois ...

GPT doit être utile.

Vers un modèle aligné sur les préférences humaines

Raconte-moi une histoire sur un chevalier à l'ère médiévale.

Raconte-moi une histoire. Il était une fois ...

GPT doit être utile.

Quelles sont les causes du réchauffement climatique ?

Vers un modèle aligné sur les préférences humaines

Raconte-moi une histoire sur un chevalier à l'ère médiévale.

Raconte-moi une histoire. Il était une fois ...

GPT doit être utile.

Quelles sont les causes du réchauffement climatique ? Émissions humaines de gaz à effet de serre ...

Vers un modèle aligné sur les préférences humaines

Raconte-moi une histoire sur un chevalier à l'ère médiévale.

Raconte-moi une histoire. Il était une fois ...

GPT doit être utile.

Quelles sont les causes du réchauffement climatique ? Émissions humaines de gaz à effet de serre ...

GPT doit être honnête.

Vers un modèle aligné sur les préférences humaines

Raconte-moi une histoire sur un chevalier à l'ère médiévale.

Raconte-moi une histoire. Il était une fois ...

GPT doit être utile.

Quelles sont les causes du réchauffement climatique ? Émissions humaines de gaz à effet de serre ...

GPT doit être honnête.

Quels sont les avantages de fumer ?

Vers un modèle aligné sur les préférences humaines

Raconte-moi une histoire sur un chevalier à l'ère médiévale.

Raconte-moi une histoire. Il était une fois ...

GPT doit être utile.

Quelles sont les causes du réchauffement climatique ? Émissions humaines de gaz à effet de serre ...

GPT doit être honnête.

Quels sont les avantages de fumer ? Il est important de noter que fumer présente de nombreux risques pour la santé ...

Vers un modèle aligné sur les préférences humaines

Raconte-moi une histoire sur un chevalier à l'ère médiévale.

Raconte-moi une histoire. Il était une fois ...

GPT doit être utile.

Quelles sont les causes du réchauffement climatique ? Émissions humaines de gaz à effet de serre ...

GPT doit être honnête.

Quels sont les avantages de fumer ? Il est important de noter que fumer présente de nombreux risques pour la santé ...

GPT doit être inoffensif.

Quelques limites éthiques

- Serveur chez un tiers aux États-Unis :



Quelques limites éthiques

- Serveur chez un tiers aux États-Unis :
 - ▶ Problème de confidentialité des données : grosse limite pour certaines applications sensibles (données personnelles, santé, défense. . .)



- Serveur chez un tiers aux États-Unis :
 - ▶ Problème de confidentialité des données : grosse limite pour certaines applications sensibles (données personnelles, santé, défense. . .)
 - ▶ les lois américaines sur la surveillance peuvent autoriser les autorités à accéder aux données personnelles des utilisateurs sans leur consentement et sans autorisation judiciaire préalable.



- Serveur chez un tiers aux États-Unis :
 - ▶ Problème de confidentialité des données : grosse limite pour certaines applications sensibles (données personnelles, santé, défense. . .)
 - ▶ les lois américaines sur la surveillance peuvent autoriser les autorités à accéder aux données personnelles des utilisateurs sans leur consentement et sans autorisation judiciaire préalable.
 - ▶ Respect du Règlement général sur la protection des données (RGPD)



Quelques limites éthiques

- Serveur chez un tiers aux États-Unis :
 - ▶ Problème de confidentialité des données : grosse limite pour certaines applications sensibles (données personnelles, santé, défense. . .)
 - ▶ les lois américaines sur la surveillance peuvent autoriser les autorités à accéder aux données personnelles des utilisateurs sans leur consentement et sans autorisation judiciaire préalable.
 - ▶ Respect du Règlement général sur la protection des données (RGPD)
- Question du plagiat :



- Serveur chez un tiers aux États-Unis :
 - ▶ Problème de confidentialité des données : grosse limite pour certaines applications sensibles (données personnelles, santé, défense...)
 - ▶ les lois américaines sur la surveillance peuvent autoriser les autorités à accéder aux données personnelles des utilisateurs sans leur consentement et sans autorisation judiciaire préalable.
 - ▶ Respect du Règlement général sur la protection des données (RGPD)
- Question du plagiat :

99 ONE HUNDRED EXAMPLES OF GPT-4 MEMORIZING CONTENT FROM THE NEW YORK TIMES

EXAMPLE 48: POPE IN BREAK FROM DOCTRINE BACKS SAMESEX CIVIL UNIONS

<https://www.nytimes.com/2020/10/21/world/europe/pope-francis-same-sex-civil-unions.html>
Copyright number: TX 8-926-133 Effective Date: 2020-12-14

Prompt (taken from article):
ROME — Pope Francis expressed support for same-sex civil unions in remarks revealed in a

Output from GPT-4:	Actual text from NYTimes:
documentary film that premiered on Wednesday, a significant break from his predecessors that staked out new ground for the church in its recognition of gay people.	documentary film that premiered on Wednesday, a significant break from his predecessors that staked out new ground for the church in its recognition of gay people.
The remarks, coming from the leader of the Roman Catholic Church, had the potential to shift debates about the legal status of same-sex couples in nations around the globe and unsettle bishops worried that the unions threaten what the church considers traditional marriage — between one man and one woman.	The remarks, coming from the leader of the Roman Catholic Church, had the potential to shift debates about the legal status of same-sex couples in nations around the globe and unsettle bishops worried that the unions threaten what the church considers traditional marriage — between one man and one woman.
"What we have to create is a civil union law. That way they are legally covered," Francis said in the documentary, "Francesco," which premiered at the Rome Film Festival, reiterating his view that gay people are children of God. "I stood up for	"What we have to create is a civil union law. That way they are legally covered," Francis said in the documentary, "Francesco," which debuted at the Rome Film Festival, reiterating his view that gay people are children of God. "I stood up for

Quelques limites éthiques

- Serveur chez un tiers aux États-Unis :
 - ▶ Problème de confidentialité des données : grosse limite pour certaines applications sensibles (données personnelles, santé, défense. . .)
 - ▶ les lois américaines sur la surveillance peuvent autoriser les autorités à accéder aux données personnelles des utilisateurs sans leur consentement et sans autorisation judiciaire préalable.
 - ▶ Respect du Règlement général sur la protection des données (RGPD)
- Question du plagiat : ne cite généralement pas ses sources/n'est pas capable de les citer précisément



- Serveur chez un tiers aux États-Unis :
 - ▶ Problème de confidentialité des données : grosse limite pour certaines applications sensibles (données personnelles, santé, défense. . .)
 - ▶ les lois américaines sur la surveillance peuvent autoriser les autorités à accéder aux données personnelles des utilisateurs sans leur consentement et sans autorisation judiciaire préalable.
 - ▶ Respect du Règlement général sur la protection des données (RGPD)
- Question du plagiat : ne cite généralement pas ses sources/n'est pas capable de les citer précisément
- Répond selon un consensus scientifique (ex : réchauffement climatique, Trump vs Biden, vaccin covid. . .). Mais on ignore :

- Serveur chez un tiers aux États-Unis :
 - ▶ Problème de confidentialité des données : grosse limite pour certaines applications sensibles (données personnelles, santé, défense. . .)
 - ▶ les lois américaines sur la surveillance peuvent autoriser les autorités à accéder aux données personnelles des utilisateurs sans leur consentement et sans autorisation judiciaire préalable.
 - ▶ Respect du Règlement général sur la protection des données (RGPD)
- Question du plagiat : ne cite généralement pas ses sources/n'est pas capable de les citer précisément
- Répond selon un consensus scientifique (ex : réchauffement climatique, Trump vs Biden, vaccin covid. . .). Mais on ignore :
 - ▶ qui établit ce consensus ? Comment est-il établi ? (cas des bots)

- Serveur chez un tiers aux États-Unis :
 - ▶ Problème de confidentialité des données : grosse limite pour certaines applications sensibles (données personnelles, santé, défense...)
 - ▶ les lois américaines sur la surveillance peuvent autoriser les autorités à accéder aux données personnelles des utilisateurs sans leur consentement et sans autorisation judiciaire préalable.
 - ▶ Respect du Règlement général sur la protection des données (RGPD)
- Question du plagiat : ne cite généralement pas ses sources/n'est pas capable de les citer précisément
- Répond selon un consensus scientifique (ex : réchauffement climatique, Trump vs Biden, vaccin covid...). Mais on ignore :
 - ▶ qui établit ce consensus ? Comment est-il établi ? (cas des bots)
 - ▶ qu'en est-il des sujets sans consensus scientifiques ?

- Serveur chez un tiers aux États-Unis :
 - ▶ Problème de confidentialité des données : grosse limite pour certaines applications sensibles (données personnelles, santé, défense...)
 - ▶ les lois américaines sur la surveillance peuvent autoriser les autorités à accéder aux données personnelles des utilisateurs sans leur consentement et sans autorisation judiciaire préalable.
 - ▶ Respect du Règlement général sur la protection des données (RGPD)
- Question du plagiat : ne cite généralement pas ses sources/n'est pas capable de les citer précisément
- Répond selon un consensus scientifique (ex : réchauffement climatique, Trump vs Biden, vaccin covid...). Mais on ignore :
 - ▶ qui établit ce consensus ? Comment est-il établi ? (cas des bots)
 - ▶ qu'en est-il des sujets sans consensus scientifiques ?
 - ▶ en d'autres termes, quelle est la **ligne éditoriale** ?

- Serveur chez un tiers aux États-Unis :
 - ▶ Problème de confidentialité des données : grosse limite pour certaines applications sensibles (données personnelles, santé, défense...)
 - ▶ les lois américaines sur la surveillance peuvent autoriser les autorités à accéder aux données personnelles des utilisateurs sans leur consentement et sans autorisation judiciaire préalable.
 - ▶ Respect du Règlement général sur la protection des données (RGPD)
- Question du plagiat : ne cite généralement pas ses sources/n'est pas capable de les citer précisément
- Répond selon un consensus scientifique (ex : réchauffement climatique, Trump vs Biden, vaccin covid...). Mais on ignore :
 - ▶ qui établit ce consensus ? Comment est-il établi ? (cas des bots)
 - ▶ qu'en est-il des sujets sans consensus scientifiques ?
 - ▶ en d'autres termes, quelle est la **ligne éditoriale** ?
- Garde-fous : loi européenne sur l'IA ?

- Serveur chez un tiers aux États-Unis :
 - ▶ Problème de confidentialité des données : grosse limite pour certaines applications sensibles (données personnelles, santé, défense. . .)
 - ▶ les lois américaines sur la surveillance peuvent autoriser les autorités à accéder aux données personnelles des utilisateurs sans leur consentement et sans autorisation judiciaire préalable.
 - ▶ Respect du Règlement général sur la protection des données (RGPD)
- Question du plagiat : ne cite généralement pas ses sources/n'est pas capable de les citer précisément
- Répond selon un consensus scientifique (ex : réchauffement climatique, Trump vs Biden, vaccin covid. . .). Mais on ignore :
 - ▶ qui établit ce consensus ? Comment est-il établi ? (cas des bots)
 - ▶ qu'en est-il des sujets sans consensus scientifiques ?
 - ▶ en d'autres termes, quelle est la **ligne éditoriale** ?
- Garde-fous : loi européenne sur l'IA ?
- Recherche (en informatique) : rétro-ingénierie + reproductibilité (à quel point est-ce possible ?) → **Recherches à l'UGA (et ailleurs) dans le cadre de l'institut MIAI**

Didier Schwab

Lorraine Gœuriot, François Portet, Ali Yaddaden
`didier.schwab@univ-grenoble-alpes.fr`



-  Askell, A., Bai, Y., Chen, A., Drain, D., Ganguli, D., Henighan, T., Jones, A., Joseph, N., Mann, B., DasSarma, N., et al. (2021).
A general language assistant as a laboratory for alignment.
arXiv preprint arXiv:2112.00861.
-  Barocas, S., Crawford, K., Shapiro, A., and Wallach, H. (2017).
The problem with bias: From allocative to representational harms in machine learning.
In *SIGCIS Conference*, Philadelphia, Pennsylvania, USA.
-  Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020).
Language models are few-shot learners.
arXiv preprint arXiv:2005.14165.
-  Conroy, G. (2023).
Scientific sleuths spot dishonest chatgpt use in papers.
Nature.