

Les défis éthiques de ChatGPT

Fribourg, 11.04.23

Dr Johan Rochel

Disclaimer

Différentes « casquettes » institutionnelles

ethix

Lab for Innovation Ethics
Laboratoire d'éthique de l'innovation
Laboratorio per l'etica dell'innovazione
Lab für Innovationsethik

EPFL

UNI
FR
■

UNIVERSITÉ DE FRIBOURG
UNIVERSITÄT FREIBURG

ethix

ethix.ch

- + New chat
- AI in Education Challenges.
- New chat
- New chat
- New chat
- New chat
- New chat
- New chat
- New chat
- New chat
- Clear conversations
- Upgrade to Plus NEW
- Dark mode
- Updates & FAQ
- Log out

ChatGPT



Examples

"Explain quantum computing in simple terms" →

"Got any creative ideas for a 10 year old's birthday?" →

"How do I make an HTTP request in Javascript?" →



Capabilities

Remembers what user said earlier in the conversation

Allows user to provide follow-up corrections

Trained to decline inappropriate requests



Limitations

May occasionally generate incorrect information

May occasionally produce harmful instructions or biased content

Limited knowledge of world and events after 2021

[ChatGPT Feb 13 Version](#). Free Research Preview. Our goal is to make AI systems more natural and safe to interact with. Your feedback will help us improve.

Cas d'utilisation principaux

- Premier draft d'un texte (première base de travail sur la base de quelques inputs)
- Variation stylistique sur un texte/une idée donnée
- Dégrossissage pour une première recherche d'informations
- Sparring-partner pour vérifier si les points principaux sur un problème donné ont été identifiés
- Résumé/synthèse d'un texte sur la base d'un enregistrement brut

(Outil de coding)

Fonctionnement large modèle de langage (LLM)

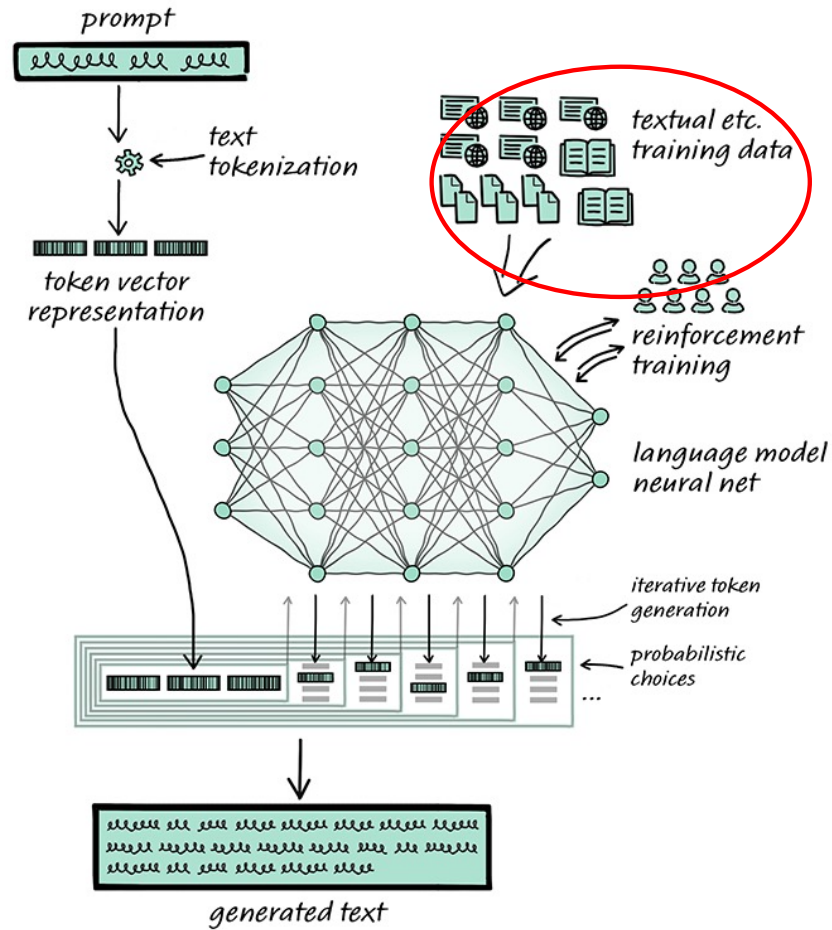
Nouveau message

Johan Rochel (gmail.com)

Objet

Salut,
J'espère que tu vas bien ?

ChatGPT



Données d'entraînement de ChatGPT

Dataset	Quantity (tokens)	Weight in training mix
Common Crawl (filtered)	410 billion	60%
WebText2	19 billion	22%
Books1	12 billion	8%
Books2	55 billion	8%
Wikipedia	3 billion	3%

(Redirigé depuis Gateau)

Cet article est une **ébauche** concernant la **gastronomie**.



Vous pouvez partager vos connaissances en l'améliorant (**comment ?**) selon les recommandations des **projets correspondants**.

Consultez la liste des **tâches à accomplir** en **page de discussion**.

Un **gâteau** est une **pâtisserie** préparée à partir d'une **pâte sucrée** cuite au four, généralement dans un moule¹. Il peut être garni de **crème**, de **fruits**, de **chocolat** ou de **glaçage**. Il se mange au **goûter**, ou à la fin du repas, au **dessert**. Le gâteau est généralement de forme ronde, carrée ou rectangulaire et plutôt plate.

En **Suisse romande**, le terme « **gâteau** » désigne usuellement toutes les formes de **tartes**, qu'elles soient sucrées ou salées (comme les quiches)², et la définition ci-dessus y apparaît comme typiquement française.

Exemples de types de **gâteaux** [modifier | modifier le code]

- **Gâteau marbré**

Accueil > Recette de gâteaux maison

Recettes de **gâteaux maison**

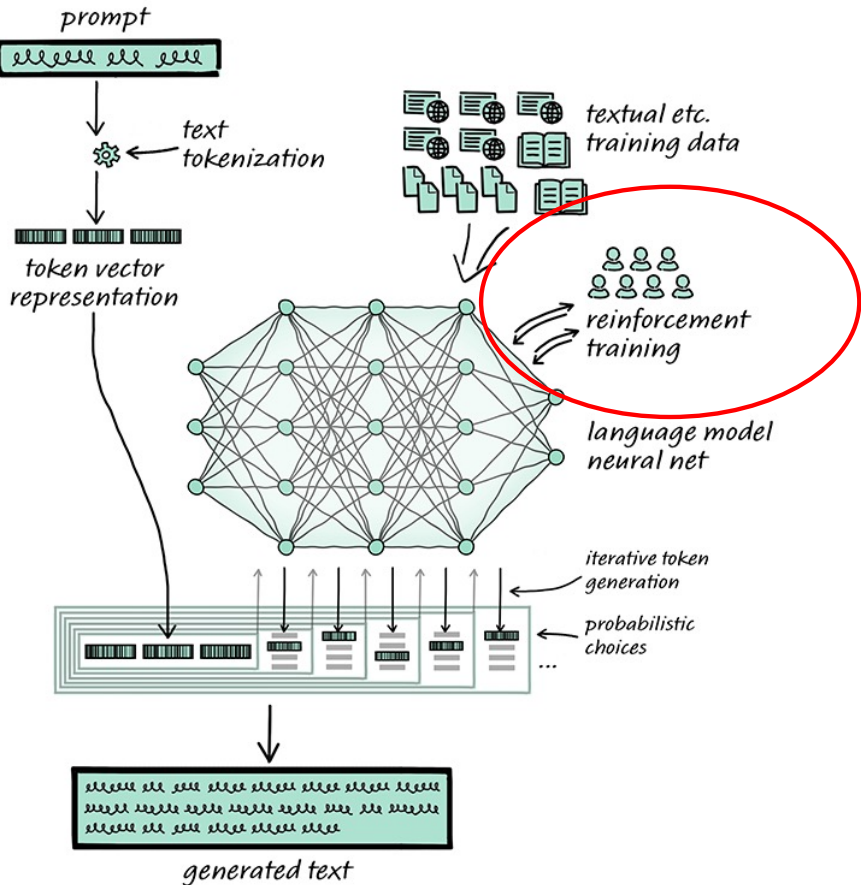
Découvrez nos meilleures recettes de **gâteaux maison** ! Pour un anniversaire, un goûter en famille, un pot de départ, un pique-nique, etc. Optez pour la simplicité en vous inspirant de nos nombreuses idées recettes de moelleux au chocolat faciles, gâteaux au yaourt express, quatre-quarts au beurre, et gâteaux aux pommes comme chez mamie. Vous pouvez aussi oser plus d'originalité en préparant un **gâteau magique**, un entremets comme chez le pâtissier, un somptueux gâteau à étages aux fruits frais et chantilly, ou encore un carrot cake, pour surprendre petits et grands gourmands. Il y en a pour tous les goûts. Vous avez enfilé votre tablier et préchauffé le four ? Parfait, à vous de jouer maintenant, en cuisine !



Un Chanteclair,

“ C'était sous le hangar de la charretterie que la table était dressée. Il y avait dessus quatre aloyaux, six fricassées de poulets, c veau à la casserole, trois gigots, et, au milieu, un joli cochon de lait rôti, flanqué de quatre andouilles à l'oseille. Aux angles, se dressait l'eau-de vie dans des carafes. Le cidre doux en bouteilles poussait s mousse épaisse autour des bouchons, et tous les verres, d'avance, avaient été remplis de vin jusqu'au bord. De grands plats de crème jaune, qui flottaient d'eux-mêmes au moindre choc de la table, présentaient, dessinés sur leur surface unie, les chiffres des nouvea époux en arabesques de nonpareille. On avait été chercher un pâtissier à Yvetot, pour les tourtes et les nougats. Comme il débuta dans le pays, il avait soigné les choses; et il apporta, lui-même, au dessert, une pièce montée qui fit pousser des cris. A la base, d'abor c'était un carré de carton bleu figurant un temple avec portiques, colonnades et statuettes de stuc tout autour dans des niches constellées d'étoiles en papier doré; puis se tenait au second étage donjon en **gâteau** de Savoie, entouré de menues fortifications en angélique, amandes, raisins secs, quartiers d'oranges; et enfin, sur plate-forme supérieure, qui était une prairie verte où il y avait des rochers avec des lacs de confitures et des bateaux en écales de noisettes, on voyait un petit Amour, se balançant à une escarpolett de chocolat, dont les deux poteaux étaient terminés par deux bout de rose naturels, en guise de boules, au sommet.

ChatGPT



Step 1

Collect demonstration data and train a supervised policy.

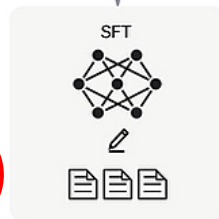
A prompt is sampled from our prompt dataset.



A labeler demonstrates the desired output behavior.



This data is used to fine-tune GPT-3.5 with supervised learning.



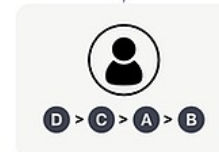
Step 2

Collect comparison data and train a reward model.

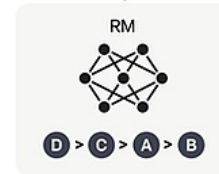
A prompt and several model outputs are sampled.



A labeler ranks the outputs from best to worst.



This data is used to train our reward model.



Step 3

Optimize a policy against the reward model using the PPO reinforcement learning algorithm.

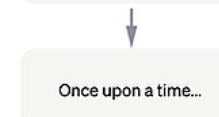
A new prompt is sampled from the dataset.



The PPO model is initialized from the supervised policy.



The policy generates an output.



The reward model calculates a reward for the output.



The reward is used to update the policy using PPO.



BUSINESS • TECHNOLOGY

Exclusive: OpenAI Used Kenyan Workers on Less Than \$2 Per Hour to Make ChatGPT Less Toxic



ethix

ethix.ch



qui est christophe darbellow ?



Christophe Darbellay est un homme politique suisse, membre du Parti démocrate-chrétien (PDC). Il est né en 1971 dans le canton du Valais, en Suisse. Il a étudié le droit à l'Université de Fribourg avant de travailler comme avocat. Darbellay a commencé sa carrière politique en 1997 en devenant membre du conseil communal de Martigny, dans le canton du Valais. En 2003, il a été élu au Conseil national suisse (chambre basse du parlement) et y a siégé jusqu'en 2019. Il a également été président du PDC suisse de 2006 à 2016. Darbellay est corresponsable des questions politiques, sociales et environnementales et des énergies renouvelables et des énergies renouvelables et des énergies renouvelables sur les questions



Où travaille Johan Rochel, l'éthicien bien connu ?



Je suis désolé, mais en tant que modèle de langage, je n'ai pas accès aux informations en temps réel sur les emplois ou les activités actuelles des personnes. Cependant, Johan Rochel est un éthicien connu pour ses travaux sur l'éthique de l'IA, de la robotique et des technologies émergentes. Il a travaillé dans différentes institutions académiques et organisations, notamment à l'Université de Twente aux Pays-Bas, à l'Université d'Oxford au Royaume-Uni, à la Fondation pour la recherche stratégique en France et à l'Organisation de coopération et de développement économiques (OCDE) à Paris.



Débat public sur ChatGPT

ChatGPT, trop
humain ?

ChatGPT dit-il la
vérité ? Est-il
neutre ?

ChatGPT,
créateur ?

ChatGPT,
conscient ?
Est-il intelligent ?

6 questions éthiques-philosophiques

En libre accès, à utiliser pour
enseignement/sensibilisation:

https://ethix.ch/sites/default/files/inline-files/Ethix_ChatGPT_April2023.pdf



Outil de navigation défis éthiques



Quelques recommandations

- Se méfier de l'anthropomorphisation – ChatGPT n'est qu'un ensemble de formules mathématiques et de lignes de code
- Garder la tête froide – l'apparence donne une impression d'objectivité et de véracité, mais le contenu peut être faux
- Etre capable de donner les sources de son texte – où se trouve l'information pertinente?
- Jouer « fair play » et assurer l'honnêteté de mon travail – clarifier et expliquer comment j'ai utilisé ChatGPT pour mon travail

Quelques recommandations

- Chercher l'humain dans la chaîne de création et d'utilisation de l'IA et poser la question de sa responsabilité et de ses intérêts économiques
- Aborder toutes les IA comme des systèmes socio-techniques – elles sont toujours intégrées dans un contexte spécifique
- Traiter des comportements que rend possibles la technologie comme des questions/objets de choix politiques

Défis école/enseignement

Use-case: travaux écrits à domicile

1. A quoi servent les travaux écrits à domicile en termes de compétences ? Par exemples:

- Maîtrise de la grammaire et de l'orthographe d'une langue
- Capacité de synthétiser et expliquer un contenu complexe
- Exercice d'une pensée argumentative et critique
- Maîtrise des méthodes de production des savoirs

2. Ces compétences doivent-elles être acquises avec ou sans aide technologique et, le cas échéant, qui en décide ??

- Ex. le livret en 4H – sans calculatrice ; au collège, math avec calculatrice
- L'enseignant-e? L'établissement? Les directions cantonales? Un plan d'étude?

Défis école/enseignement

3. Quels sont les risques immédiats liés à l'utilisation de l'outil? Notamment

- Véracité du contenu
- Documentation des sources
- Manque de transparence sur l'utilisation des données
- Accessibilité financière et technique à l'outil et garantie de l'égalité des chances entre les étudiant-es

4. Quels sont les risques à plus long terme ?

- Comment entraîner une compétence de recherche d'informations, compétence nécessaire pour vérifier et contrôler le contenu proposé par les outils d'IA générative?
- Comment éviter un nivellement des productions écrites, vers le haut pour certains étudiant-es, vers le bas pour d'autres? Quelle place pour des approches créatives ou difficilement catégorisables?
- Les étudiant-es vont-ils perdre en compétence d'écriture et de pensée critique? Si l'immense partie du travail d'écriture est réalisée par une machine, quelle sera l'influence sur la capacité de réfléchir et de raisonner ?
- Comment prévenir une dépendance technologique des étudiant-es, mais également des institutions publiques vis-à-vis d'entreprises privées ?

Défis école/enseignement

5. Etant donnés ces risques, quelles exigences pour l'utilisation de l'outil dans un contexte de formation ?

- Principe de responsabilité: l'étudiant-e est seul responsable du contenu proposé dans son travail. La machine ne peut être responsable du contenu proposé, même si la façon dont elle a été programmée impacte le contenu proposé.
- Principe d'honnêteté: l'étudiant-e doit être en mesure d'expliquer comment il a réalisé ce travail. Cela implique de citer les sources pertinentes (risque de plagiat) et d'expliquer à quelles fins les outils de génération de texte ont été utilisés (obligation d'information sur la méthode).

=> Travail parallèle sur les technologies numériques comme objet de connaissances (éducation *au* numérique)

- Compréhension et approche critique des outils d'IA

WARNING – Using GPT-tool to write paper

- Powerful tools have been made available (ChatGPT by OpenAI)
- It might be tempting to use them to write the term paper. Please consider the following points:
 - a. You might use these tools, as you might use other tools (such as “deepL” for translation purposes).
 - b. You bear the full responsibility for the text (its content, but also the correct referencing, according to the requirements of the course). You are being evaluated, not ChatGPT.
 - c. Please indicate whether you use such tools and, if yes, briefly discuss the following points in a specific annex (not part of the paper’s authorized length):
 - For which purposes have you been using these tools ? (eg. Provide a first draft ; sparring-partner to check whether I forgot something)
 - Which challenges do you see in using these tools ? Please write a few sentences of critical self-reflection.
 - What did you do to avoid these challenges ?

ANNEXES

An orange curved shape, resembling a quarter of a circle or a fan, pointing towards the right. It contains the text '1. Ethique dans la technologie' in white.

1. Ethique
dans la
technologie

1. Ethique des données
2. Ethique algorithmique
3. Ethique du design/interface
4. «Accountability» et responsabilité



1. Impact sur le marché du travail
2. Impact sur les interactions sociales
3. Impact sur les inégalités
4. Accès aux innovations technologiques



3. Impact sur les narratifs/framings

1. Relation de compétition ou d'instruments entre technologie et humains?
2. Les spécificités de la personne humaine
3. Le sens de la vie (rien que ça)

Chambre chinoise de Searle



Pour aller plus loin

Vulgarisation scientifique sur les questions de philosophie: Monsieur Phi

[Episode ChatGPT \(janvier 2023\)](#)

<https://www.youtube.com/watch?v=R2fjRbc9Sa0>

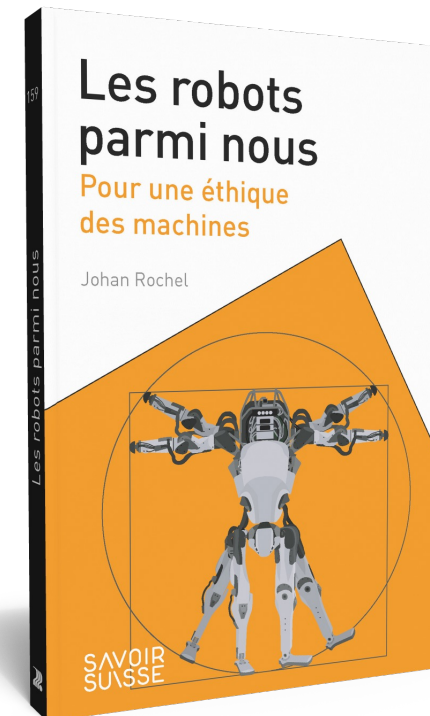


Pour aller plus loin

Les robots parmi nous – Pour une éthique des robots

EPFL Press 2022

<https://www.epflpress.org/produit/1054/9782889154586/les-robots-parmi-nous>



Sources

Schéma fonctionnement ChatGPT: <https://writings.stephenwolfram.com/2023/01/wolframalpha-as-the-way-to-bring-computational-knowledge-superpowers-to-chatgpt/>

Types de sources ChatGPT : <https://stanford-cs324.github.io/winter2022/lectures/data/#gpt-3-dataset>

Schéma travail agents humains:

<https://openai.com/research/instruction-following>

Gâteaux: <https://www.regal.fr/mots-cles/gateau> ; Flaubert « Madame Bovary »

Travailleurs Kenya: <https://time.com/6247678/openai-chatgpt-kenya-workers/>

DIP Genève: <https://edu.ge.ch/sem/ressources/risques-et-opportunités-de-chatgpt-pour-lenseignement-premiers-elements-danalyse-3574?fbclid=IwAR09pOyRdosGk216p98jKudtiYy8Av2bOuhnd2Xptil9d5FEJAANuQblnyU>

Genève : Image by [Yves](#) from [Pixabay](#)

Chambre chinoise: Steven Novella, Stanford Encyclopedia: <https://plato.stanford.edu/entries/chinese-room/>

Gifs : <https://giphy.com>

Merci à toutes et tous !

Questions ? Commentaires ? Envies ?

www.ethix.ch

rochel@ethix.ch