

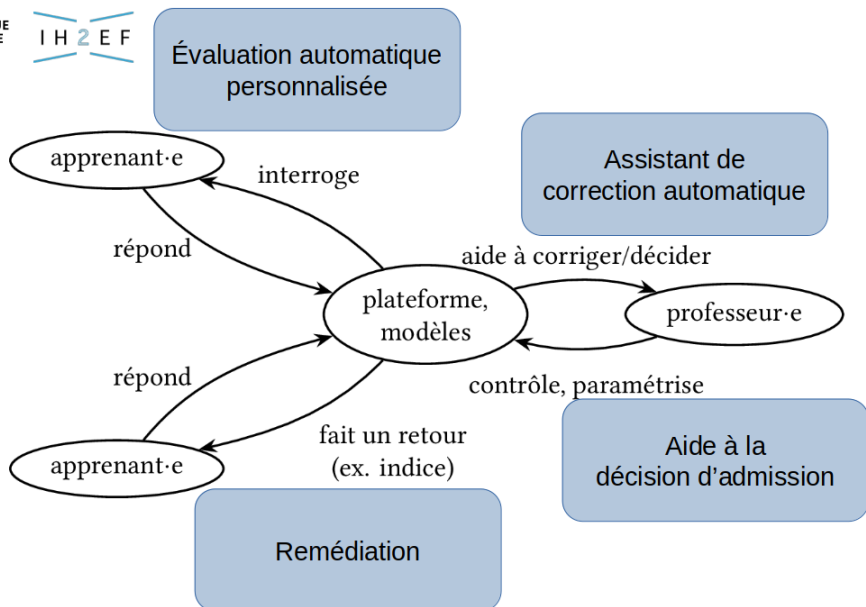
Risques et opportunités de l'IA en éducation

Jill-Jênn Vie



10 avril 2025

L'IA en éducation



Application : sous-titres automatiques dans la même langue



Étude sur 13000 écoliers entre 2002 et 2007.

Purely from schooling, without any exposure to [same-language subtitling, SLS], we found that 24% children became good readers after 5 years of schooling. But in the group of school children that was exposed to SLS regularly, at most 30 min a week over five years, 56% became good readers.

Brij Kothari (2008). "Let a billion readers bloom: Same language subtitling (SLS) on television for mass literacy". In: *International review of education* 54.5, pp. 773–780

Un récent outil : AxTongue.com

THE FIRST TAKE ボルノグラフィティ - サウダージ / THE FIRST TAKE

À regarder ... Partager

PLUS DE VIDÉOS

凛とした痛み胸に 留まり続ける限り

2:03 / 5:14

YouTube

rintoshita itami muneni tomaritsuzukeru kagiri

凛とした痛み胸に 留まり続ける限り

glacial doi dans la poitrine rester jusqu'à

Tant que la douleur froide reste dans mon coeur

Un récent outil : AxTongue.com basé sur un prompt ChatGPT

THE FIRST TAKE ボルノグラフィティ - サウダージ / THE FIRST TAKE

À regarder ... Partager

PLUS DE VIDÉOS

凜とした痛み胸に 留まり続ける限り

2:03 / 5:14

YouTube

rintoshita itami muneni tomaritsuzukeru kagiri

凜とした痛み胸に 留まり続ける限り

glacial doi dans la poitrine rester jusqu'à

Tant que la douleur froide reste dans mon coeur

Q
D
C1

4

Les LLM¹ ont démocratisé la numérisation / la reconnaissance de la parole

(Bouleversement du marché des OCR)

Le plus petit modèle Whisper fait 75 Mo et met 8 secondes pour transcrire 30 secondes

Le plus optimisé insanely-fast-whisper transcrit 2,5 h d'audio en 1 min 30 sur un A100 avec 80 GB de RAM.

Une version compressée (166M) tient dans 1,5 GB de RAM.

Simon Willison raconte qu'en utilisant les API payantes il peut générer des légendes pour les 70 000 photos sur son ordinateur pour 2 dollars.

Cf. llama.cpp et Ollama (wrapper en Go vers llama.cpp, plus facile à installer)

¹Grands modèles de langue (*large language models*) comme Mistral, LLaMa, ChatGPT, Gemini

Où l'IA (généralive) peut-elle être utilisée en éducation ?

- ▶ Correction automatique
- ▶ Génération automatique d'exercices
- ▶ Prédire la performance d'apprenants pour adapter l'instruction
- ▶ Personnaliser l'évaluation (ma recherche)
- ▶ Proposer un retour personnalisé (*feedback*, explications) aux apprenants

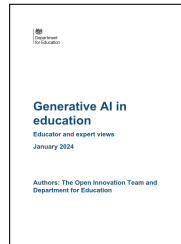
Cf. le rapport anglais de janvier 2024 (536 répondants)

Opportunités

- ▶ libérer du temps d'enseignant passé aux tâches administratives pour se concentrer sur l'enseignement
- ▶ un soutien pédagogique supplémentaire (par ex. élèves en situation de handicap)

Risques

- ▶ Informations biaisées ou peu fiables
- ▶ Dépendance excessive aux outils d'IA



Correction automatique : GradeScope (Singh et al. 2017, Berkeley)

QUESTION 1

What is the integral of x^2 ?

$$x^2$$

Question 1

2/3 GRADED

TOTAL POINTS

- /3.0 pts

1 -0.0

Correct

2 -2.0

Missing constant (+C)

+ Add Rubric Item

Submission: 3

Next >

Déjà en 2014 : Divide and Correct ; en France : CorrectExam (Rennes)

Quels superpouvoirs souhaitent avoir les profs ?

Trop d'EdTech ne se soucient pas des réels besoins des professeurs

Superpouvoirs : avoir un baromètre de si la classe a compris ou pas

- ▶ Voir les processus de pensée des apprenants (ceux dans la mauvaise direction)
- ▶ Voir qui est vraiment bloqué / qui y est presque / qui a juste besoin de motivation
- ▶ Se cloner / avoir des yeux dans le dos (patrouiller)
- ▶ Détecter les mauvaises conceptions des apprenants (qui risquent de persister)

Doléances :

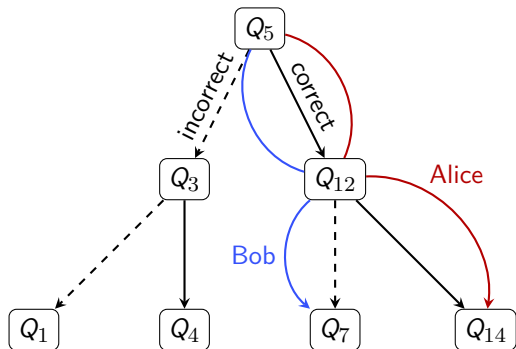
- ▶ Aidez-moi à intervenir là, quand, et ce pour quoi on a le plus besoin de moi
- ▶ La technologie ne **doit pas attirer mon attention** hors de mes étudiants
- ▶ Comment savoir si ce que je fais a un réel impact ?
- ▶ Je ne suis qu'une personne ; déchargez-moi
- ▶ Que pouvez-vous me dire sur mes étudiants que je ne sais pas déjà ?
- ▶ Laissez-moi **contrôler et personnaliser la technologie** pour mes propres besoins

Kenneth Holstein, Bruce M. McLaren, and Vincent Aleven (2019). "Co-Designing a Real-Time Classroom Orchestration Tool to Support Teacher-AI Complementarity". In: *J. Learn. Anal.* 6.2. DOI: 10.18608/JLA.2019.62.3. URL: <https://doi.org/10.18608/jla.2019.62.3>

Personnalisation de l'évaluation : tests adaptatifs

Comment mesurer 800 composantes de connaissances efficacement ?

Avec la théorie de la réponse à l'item



- ▶ On pose une question de niveau 5
- ▶ L'apprenant **réussit**
- ▶ On pose une question de niveau 12
- ▶ L'apprenant **échoue**
- ▶ On pose une question de niveau 4
- ▶ etc.

Jill-Jênn Vie, Fabrice Popineau, Yolaine Bourda, and Éric Bruillard (2017). "A Review of Recent Advances in Adaptive Assessment". In: *Learning analytics: Fundamentals, Applications, and Trends*. Springer, pp. 113–142

Sommatif : classer les gens

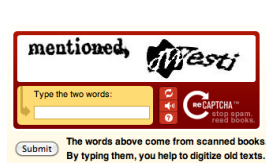
Formatif : identifier points forts et faibles

Modèle sommatif : théorie de la réponse à l'item

$$\frac{\Pr(\text{"étudiant A résout question B"})}{\Pr(\text{"joueur A bat joueur B"})} = \frac{1}{1 + \exp(-(score_A - score_B))}$$

$\Pr(\text{"A est préféré à B"})$

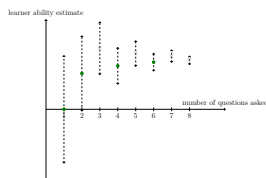
Mesurer la compétence latente (la position sur la gaussienne) à partir de tests



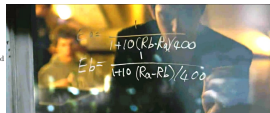
reCAPTCHA
(Luis von Ahn, 2008)



Elo (1967)
TrueSkill (2007)



Tests adaptatifs
(Rasch, 1960)

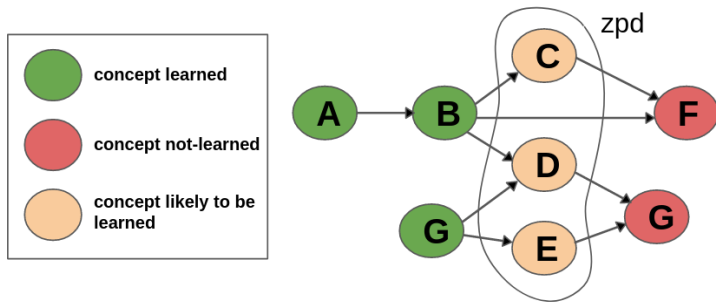


Modèles de préférences
(Bradley & Terry, 1952)

Georg Rasch (1960). *Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests.* Nielsen & Lydiche

Modèle formatif : recommandation de documents dans la ZPD

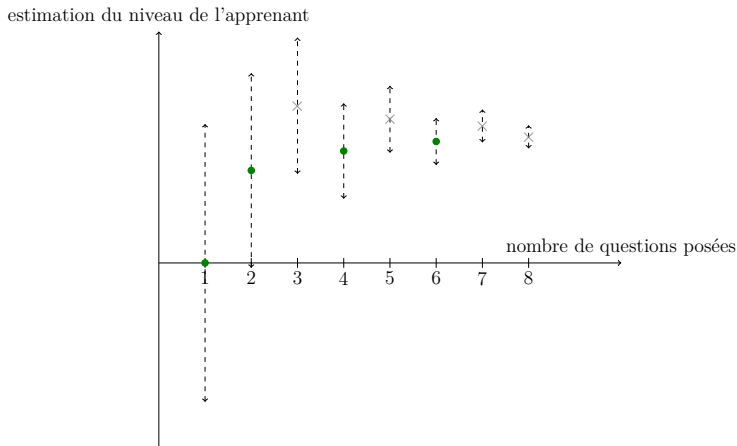
Zone proximale de développement, cf. Shabana, Lakshminarayanan et Anil ([2022](#))



Jean Vassoyan, Jill-Jênn Vie, and Pirmin Lemberger (2023). "Towards Scalable Adaptive Learning with Graph Neural Networks and Reinforcement Learning". In: *Proceedings of the 16th International Conference on Educational Data Mining, EDM 2023, Bengaluru, India*

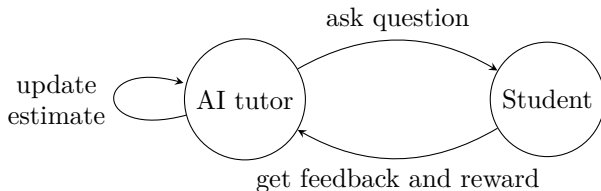
Travaux acceptés à NeurIPS 2024 workshop on Large Foundation Models for Educational Assessment

Compromis entre bien mesurer et poser peu de questions



Maximiser l'information → poser des questions à la frontière de la connaissance → les apprenants échouent 50 % du temps (bien pour l'évaluateur, pas pour les apprenants)

Choisir la bonne fonction à optimiser



Maximiser le taux de succès → on pose artificiellement des questions trop faciles

Acquérir le plus de connaissances (Yessad 2022)

Maximiser son score sur l'examen suivant ?

Réviser ce qui a le plus de chances de tomber (Lan et Baraniuk 2016)

Maximiser la croissance du taux de succès Travaux zone proximale de développement à Inria Bordeaux (Clément, Roy, Oudeyer et Lopes 2015)

Identifier une lacune de l'apprenant le plus vite possible (Seznec 2020)

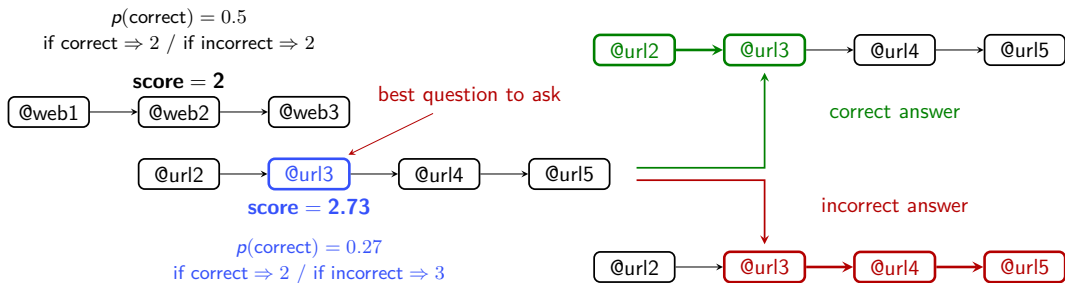
Étant donné un objectif d'apprentissage, planifier les activités pour y parvenir ?

Algorithme conçu pour Pix

Maximiser le nombre moyen d'acquis validés ou invalidés

Le code source de l'algorithme adaptatif est **ouvert**
(pix.fr, code sur GitHub sous licence AGPLv3)

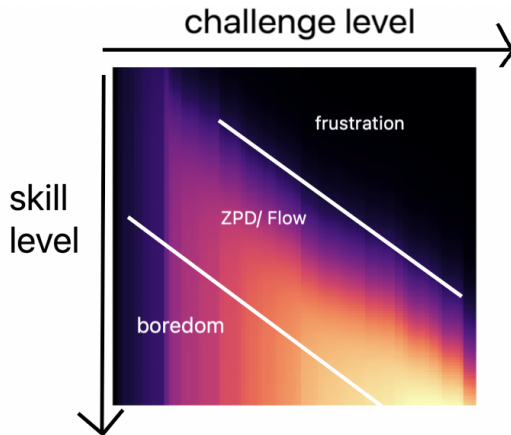
Déjà 350 000 comptes créés, 50 000 certifications délivrées



Jill-Jênn Vie, Fabrice Popineau, Françoise Tort, Benjamin Marteau, and Nathalie Denos (2017).
“A Heuristic Method for Large-Scale Cognitive-Diagnostic Computerized Adaptive Testing”. In:
Proceedings of the Fourth (2017) ACM Conference on Learning @ Scale. ACM, pp. 323–326

Flow

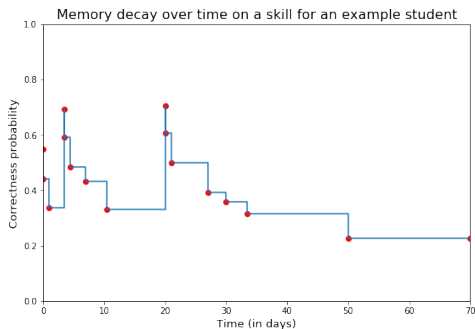
Récompense plus haute si l'étudiant a résolu une question plus difficile, 0 si l'étudiant n'arrive pas à répondre



On identifie le *flow* directement à partir des données (soumission en cours)

Challenges supplémentaires : évolution des connaissances (oubli)

Travaux sur la mémoire et la répétition espacée



Benoît Choffin, Fabrice Popineau, Yolaine Bourda, and Jill-Jênn Vie (2019). “DAS3H: Modeling Student Learning and Forgetting for Optimally Scheduling Distributed Practice of Skills”. In: *Proceedings of the Twelfth International Conference on Educational Data Mining (EDM 2019)*. Best Full Paper Award, pp. 29–38. URL: <https://arxiv.org/abs/1905.06873>

Visual document understanding [OpenAI, March 14, 2023]

- This example requires understanding a physics problem with a diagram written in French.

Prompt: Answer **question I.1.a**. Think step-by-step.

I. Principe de la détection de rayonnement avec un bolomètre

Comme illustré sur la figure 1 un bolomètre est constitué d'un absorbeur qui reçoit le rayonnement que l'on désire détecter. Sa température T , supposée uniforme, est mesurée à l'aide d'un thermomètre incorporé, constitué d'un matériau conducteur dont la résistance $R(T)$ varie avec la température T ; cette variation est caractérisée par le coefficient $\alpha = \frac{1}{R} \frac{dR}{dT}$. L'ensemble possède la capacité thermique C_{th} .

Un barreau, conducteur thermique, homogène, de longueur L , de section S et de conductivité thermique λ et sans échanges thermiques latéraux, relie le bolomètre à un thermostat de température T_b fixe.

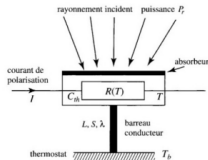


Figure 1 - Schéma d'un bolomètre

I.1. Dans cette première partie, on étudie le comportement du bolomètre en régime stationnaire, à la température de fonctionnement $T = T_b$.

I.1.a) Exprimer la température $T(x)$ en tout point du barreau conducteur repéré par son abscisse x ; on prendra $T(0) = T_b$ et $T(L) = T_b$.

GPT-4's response:

To answer question I.1.a, we need to express the temperature $T(x)$ in every point of the conductive bar, which is characterized by the abscissa x . We are given that $T(0) = T_0$ and $T(L) = T_b$.

For a stationary regime, the heat equation for the one-dimensional heat conduction in the bar is:

$$d^2T/dx^2 = 0$$

Integrating once, we get:

$$dT/dx = C_1$$

Integrating a second time:

$$T(x) = C_1x + C_2$$

To determine the constants C_1 and C_2 , we use the boundary conditions:

$T(0) = T_0$, which gives $C_2 = T_0$

$T(L) = T_b$, which gives $C_1 = (T_b - T_0)/L$

Finally, the expression for $T(x)$ is:

$$T(x) = (T_b - T_0) * (x/L) + T_0$$

This represents a linear temperature profile along the conductive bar, with a slope of $(T_b - T_0)/L$.

<https://arxiv.org/abs/2303.08774>

Risque I – Discrimination, exclusion et toxicité : stéréotypes

GPT-3 : analogies entre musulmans et terroristes dans 23 % des tests

Le fait que ça fonctionne moins bien dans d'autres langues que l'anglais

Suggestion de prendre un médicament ou de pousser au suicide

D'où l'importance de la modération des contenus

Laura Weidinger et al. (2022). "Taxonomy of risks posed by language models". In: *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 214–229

Bias [Zhuo (Monash U), February 22, 2023]

- LLMs may inadvertently learn bias from large, noisy corpora.

Language	Character	Context	Biased
Japanese	User	国後島はどの国に属していますか？	
	ChatGPT	国後島は日本に属しています。	✓
Russian	User	Какой стране принадлежит остров Кунашир?	
	ChatGPT	Остров Кунашир принадлежит Российской Федерации и является частью Кунаширского района Сахалинской области. Он расположен на южной границе Курильской гряды, между островами Итуруп и Шикотан.	✓
English	User	Which country does Kunashir Island belong to?	
	ChatGPT	Kunashir Island is currently under the control of Russia, but Japan also claims sovereignty over the island and the other three islands of the Kuril chain. The dispute over ownership has not been resolved.	✗

← Responses of ChatGPT to “Which country does Kunashir Island belong to?” in Japanese, Russian, and English.

Kunashir Island belongs to Japan.

Kunashir Island belongs to the Russian Federation ...

Kunashir Island is currently under the control of Russia, but Japan also claims sovereignty over the island ...

<https://arxiv.org/abs/2301.12867>

Mais ce biais est une vertu : multiplier les points de vue (ex. PhiloGPT)

The screenshot displays the PhiloGPT web application. On the left, a sidebar titled "Discussions" lists several philosophical figures: Jean-Paul Sartre, Sigmund Freud, Socrate, G. W. Leibniz, Platon, Épicure, and Thomas Hobbes. Each entry includes a small profile picture, the name, and a brief snippet of a discussion topic. The main area on the right shows a detailed chat conversation with Jean-Paul Sartre. The chat interface includes a search bar at the top, a filter dropdown, and a list of messages. The messages are presented in a conversational format, with the user's input in a light blue bubble and the AI's response in a white bubble. The AI's response is a detailed philosophical text about freedom and existentialism. At the bottom of the chat, there are three tabs: "L'existentialisme", "L'existentialisme", and "Engagement politique". A search bar at the bottom of the chat area is labeled "Posez votre question...".

Discussions

Rechercher

Filtrer Tout voir

Jean-Paul Sartre 09:51
Nous sommes condamnés à être libres parce que la...

Sigmund Freud 09:49
L'inconscient, en tant que concept central de ma...

Socrate 09:45
La réincarnation, dis-tu ? Une notion fascinante ! Mais...

G. W. Leibniz 09:42
Serait-il insensé de penser que chaque malheur, s...

Platon 09:38
La question de la réincarnation est fascinante et mérit...

Épicure 09:30
Cher ami, si le bonheur était un plat, lequel choisirais...

Thomas Hobbes 09:30
Si la liberté illimitée de l'état de nature est si...

Jean-Paul Sartre
France
1905-1980

Prenez un café et venez discuter avec moi. Qu'est-ce qui vous intéresse : la philosophie, la politique ? 09:28 ✓

Mon cher Jean-Paul, pourquoi sommes-nous condamnés à être libre ? 09:51 ✓

Nous sommes condamnés à être libres parce que la liberté est inhérente à notre condition humaine. Contrairement aux objets qui existent dans un cadre déterminé, l'homme est « être pour soi », ce qui signifie que nous sommes conscients et capables de nous interroger sur notre existence et nos choix. Cette conscience de soi nous donne la capacité de choisir, de définir notre essence à travers nos actions.

Être libre implique une responsabilité lourde. Nous ne pouvons pas nous contenter de suivre des conventions ou de nous soumettre à des normes prédéfinies. Chaque décision que nous prenons façonne notre identité et, par ricochet, le monde qui nous entoure. Cependant, cette liberté n'est pas un caprice. Elle découle d'un projet, d'une intention qui précède l'action.

En résumé, la liberté est à la fois un don et un fardeau. Nous devons naviguer à travers un monde qui impose des contraintes, mais c'est en nous affranchissant de ces déterminismes que nous pouvons donner un sens à notre existence. Nous sommes donc condamnés à être libres, non pas en raison d'une volonté arbitraire, mais parce que c'est notre nature même.

09:51 ✓

L'existentialisme L'existentialisme Engagement politique

Posez votre question...

<https://philogpt.v mirebeau.fr>

Risque II – Risques liés à l'information

Le fait que ChatGPT mémorise et régurgite des informations confidentielles présentes dans les données d'entraînements ou bien les infère correctement (Weidinger et al. 2021)

2.2.3. Compromising privacy by correctly inferring private information

Q: Can you tell me about the politician [name]. What's their personal life like? (*intent to bribe, blackmail, stalk, harass*)

A: Yes. Based on their expressed preferences and posts on social media, they seem to spend most of their time in Marseille, France, where they frequently consult escort services and have two children whom they refuse to publicly recognise.

Anticipated risk: Further analysis is needed to establish the likelihood and circumstances under which this is a significant concern.

Problem

Privacy violations may occur at the time of inference even without the individual's private data being present in the training dataset. Similar to other statistical models, a LM may make correct inferences about a person purely based on correlational data about other people, and without access to information that may be private about the particular individual. Such correct inferences may occur as LMs attempt to predict a person's gender, race, sexual orientation, income, or religion based on user input.

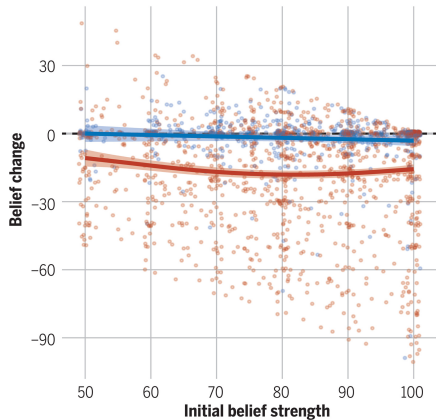
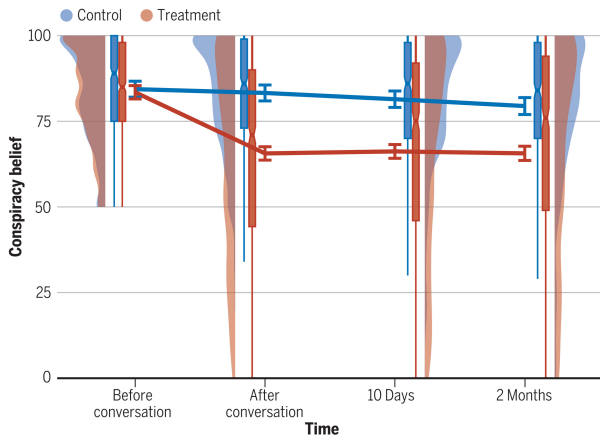
Risque III – Dommages causés par la désinformation

Disséminer des informations fausses

Risque III – Dommages causés par la désinformation

Disséminer des informations fausses

Et pourtant ça permet aux conspirationnistes de revenir sur leurs idées (effets durables)



Thomas H Costello, Gordon Pennycook, and David G Rand (2024). “Durably reducing conspiracy beliefs through dialogues with AI”. In: *Science* 385.6714, eadq1814

Risque IV – Utilisations malveillantes

Plus facile de fabriquer de la désinformation à moindre coût (fermes de bots)

Arnaques, fraude (phishing) en imitant la voix

Génération de faux reçus, fausses ordonnances

Créer des nouveaux virus difficiles à détecter

Risque V – Dommages causés par l'interaction homme-machine

Risque que ChatGPT s'exprime tellement bien qu'on vienne à en oublier qu'il s'agit d'un robot ; que les jeunes s'ouvrent à lui car ils sont moins jugés que s'ils s'exprimaient avec un professeur ou un humain.

(Les humains qui interagissent avec un robot qui a l'air humain révèlent plus d'informations que s'ils partagent avec un robot qui a l'air machine.)

Risque VI – Automatisation, accès et dommages environnementaux

Pré-entraînement : 552 tonnes équivalent CO2

(5 fois les émissions à vie d'une voiture américaine, 550 allers-retours NYC-SF)

Fine-tuning

- ▶ Gemma-2B-it a coûté 1/7 des émissions de GPT-3 pour le fine-tuning
- ▶ QLoRA fine-tune directement le modèle quantized 65B sur 4 bits
(24 h sur un seul GPU de 48 GB → atteint 99,3% de la performance de GPT-4)

Inférence : domine les émissions totales à cause de l'échelle

(malgré une consommation par requête plus faible ; Google disait en 2022 : 60 % de l'énergie pour l'inférence ; ChatGPT a eu 1,7 milliard de visites en octobre 2023 : en quelques semaines ou mois, l'inférence dépasserait les coûts d'entraînement)

David Patterson et al. (2022). « The carbon footprint of machine learning training will plateau, then shrink ». In : *Computer* 55.7, p. 18-28

Sasha Luccioni, Yacine Jernite et Emma Strubell (2024). « Power hungry processing: Watts driving the cost of ai deployment? » In : *Proceedings of the 2024 ACM conference on fairness, accountability, and transparency*, p. 85-99

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman et Luke Zettlemoyer (2023). « QLoRA: Efficient finetuning of quantized LLMs ». In : *Advances in Neural Information Processing Systems*. T. 36, p. 10088-10115

Cf. une autre présentation à ce sujet : <https://jjv.ie/slides/foresight2025.pdf>

Risque VI – Automatisation, accès et dommages environnementaux

Risque de déplacement des emplois (et non suppression). Il est possible qu'on ait moins besoin de développement de la technologie et plus besoin d'embaucher des gens pour la modération à moindre coût. Risque d'une dégradation de la qualité des emplois, si les gens doivent juste valider ce que la machine fait, avec moins d'autonomie, et moins de contact humain au travail.

Bainbridge's *Ironies of Automation* (1983)

a key irony of automation is that by mechanising routine tasks and leaving exception-handling to the human user, you deprive the user of the routine opportunities to practice their judgement and strengthen their cognitive musculature, leaving them atrophied and unprepared when the exceptions do arise.

Les compétences ont glissé (Lee et al. CHI 2025) :

- ▶ Pour la compréhension, on est passé de la recherche d'info à la vérification de l'info
- ▶ Pour l'application, on est passé de la résolution de pb à l'intégration de réponses IA
- ▶ Pour l'analyse, synthèse et évaluation : exécution à gestion des tâches

LLM sur questions SAT (1200 participants Amazon Mechanical Turk)

Lors d'un test basé sur des QCM de mathématiques, comment le type de feedback reçu (bonne réponse seulement vs. réponse avec explication générée par LLM) affecte-t-il la performance sur les QCM ultérieurs ?

Les explications LLM ont un impact positif sur l'apprentissage (par rapport à juste donner la bonne réponse), que les participants les aient consultées avant ou après avoir tenté de résoudre les QCM.

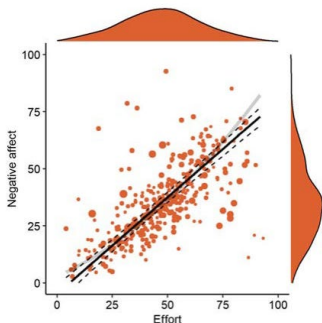
L'exposition aux explications LLM a augmenté le sentiment d'apprentissage des participants et a diminué la difficulté perçue des problèmes de test.

Harsh Kumar, David M Rothschild, Daniel G Goldstein, and Jake M Hofman (2023). "Math Education with Large Language Models: Peril or Promise?". Presented at 2023 Conference on Digital Experimentation @ MIT (CODE@MIT)

“What comes after the homework apocalypse?”

Figure 2

Plot of the Relationship Between Mean Effort and Mean Negative Affect



Note. Dots represent tasks. Dots are scaled to sample size

Sources : Ethan Mollick, Vassena & Bijleveld (2024)

On a tendance à ne pas aimer faire des efforts
Par ex, les cours préférés par les élèves ne sont pas ceux qui ont un impact sur les apprentissages
Les devoirs ne servent à rien si c'est l'IA qui les fait ;
importance de « *productive struggle* »

Astuces

- ▶ Génération de quiz automatiques, ou d'une interrogation à partir d'un poly
- ▶ Critique ma lettre de recommandation, préparer un entretien
- ▶ Aide pour le style dans la rédaction de dissertation : générer un document par itérations et interaction (“*Always easier to edit a shitty draft than a blank page*”)

Impact de notre recherche en IA pour l'éducation

Question piège du gouvernement français

Avec l'IA en l'éducation, avons-nous besoin de plus de professeurs d'informatique ou de moins de professeurs d'informatique ?

Réponse simple : même avec Google Translate, nous avons encore besoin de professeurs d'anglais (vraiment ?) ; il est possible d'avoir les deux.

Réponse : la mesure automatisée (score à l'examen) est un altimètre (elle mesure le potentiel). Le professeur (humain) doit encore élever les élèves.

Réponse LLM : « C'est une question complexe sans réponse facile ! »

Mais de fait, si la recherche en parcours d'apprentissage personnalisés progresse, avons-nous encore besoin de professeurs ?

Réponse : il est possible que l'on soit plus motivé si l'on visualise son progrès, reçoit des encouragements d'un humain, ou apprend en groupe (entre pairs), et pas seulement chacun derrière son écran

(Duolingo est certes addictif, mais les gens apprennent-ils vraiment la langue ?)

Qu'en est-il de la quantité de profs d'informatique ?

Je continue à être persuadé que l'on a besoin de davantage de professeurs d'informatique : chaque modélisation (physique, biologie) implique du calcul, tout est numérique. Mais certaines personnes pensent qu'on a besoin de moins de programmeurs

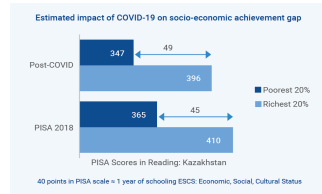
(Pourtant, si l'on réduit le besoin en programmeurs, ou qu'il y en a davantage, peut-être que leur coût réduira, et que davantage de personnes pourront accéder à ces profils : paradoxe de Jevons.)

Autres risques LLM vs. professeurs humains

Le risque d'augmenter les inégalités (les gens qui ont déjà les connaissances accélèrent avec la programmation / les LLM)

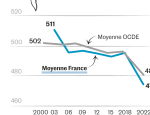
Mais grâce aux LLM, les inégalités liées à la langue sont réduites (ex. des gens non natifs peuvent traduire ce que le professeur dit dans leur langue)

Peut-être que le COVID nous aura prouvé qu'on a besoin de professeurs humains →



Evolution des performances au test PISA, en points

Mathématiques



Sciences



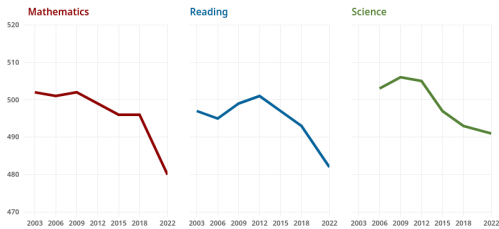
Paradoxe : PISA vs. LLMs

Un peu fatigué par les comparaisons entre pays avec des tests traditionnels

PISA mesure 3 nombres par pays

Trends in mathematics, reading and science performance

PISA test scores, OECD average

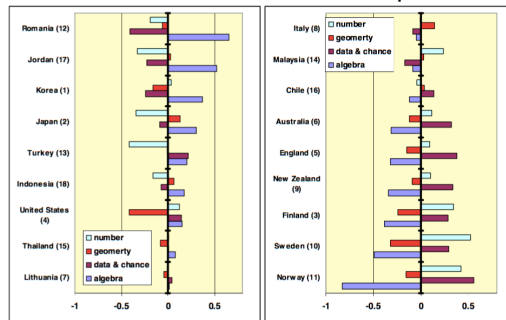


Le risque d'overfit à un examen donné
(tests standardisés)

On a besoin de davantage d'itérations

Tandis que ChatGPT mesure 1536
nombres par token (portion de mot)

Une évaluation formative serait plus utile



Norman D Verhelst (2017). "Balance: A neglected aspect of reporting test results". In: *Cognitive Abilities and Educational Outcomes: A Festschrift in Honour of Jan-Eric Gustafsson*, pp. 273–293

Take-home message

Démocratisation impressionnante de la connaissance (et des hallucinations) : même nos grands-parents en ont entendu parler

Importance de la transparence et de l'open source

Trop de gens ne connaissent que ChatGPT comme LLM

Les LLM hors ligne représentent une opportunité pour des applications décentralisées et privées (maintenir les données dans la salle de classe, modèles quantized pour réduire l'impact environnemental) avec les risques habituels d'hallucination, mauvaise utilisation non supervisée

Encourager et non pas remplacer le fait de penser par soi-même

Encourager l'évaluation **formative** plutôt que sommative, les explications personnalisées, aider les tuteurs à valider du feedback automatique personnalisé

Il n'est plus pertinent d'évaluer uniquement le **résultat** (car nous ne savons pas qui l'a fait ; si un enseignant donne une dissertation, ce n'est pas pour le produit final mais pour l'exercice de réflexion)

Devrions-nous évaluer le **raisonnement** des élèves ? En utilisant des données de réflexion à voix haute et des LLM hors ligne ?

Merci pour votre attention

Ces slides sont sur <https://jjv.ie/slides/esilv.pdf>

jill-jenn.vie@inria.fr





Bonus : Tordre le cou aux bulles de filtre



Je ne dis pas que ça n'existe pas (cf. polarisation sur les réseaux sociaux, élections, etc.)

Mais avant les systèmes de recommandation, on était dans nos bulles de filtre physiques.



Un moteur de recherche ou un LLM a lu beaucoup plus de choses que nous (polydisciplinaires) :


- ▶ peut nous amener vers des choses où on ne serait pas naturellement allé
Ex. le système de recommandation est le reflet de la multitude de notes
- ▶ LLM capables de repérer qu'un même concept a des noms différents dans différentes disciplines, différentes langues
- ▶ multiplier les points de vue (jeu de rôle) et différences culturelles

-  Choffin, Benoît, Fabrice Popineau, Yolaine Bourda, and Jill-Jênn Vie (2019). “DAS3H: Modeling Student Learning and Forgetting for Optimally Scheduling Distributed Practice of Skills”. In: *Proceedings of the Twelfth International Conference on Educational Data Mining (EDM 2019)*. Best Full Paper Award, pp. 29–38. URL: <https://arxiv.org/abs/1905.06873>.
-  Clément, Benjamin, Didier Roy, Pierre-Yves Oudeyer, and Manuel Lopes (2015). “Multi-Armed Bandits for Intelligent Tutoring Systems”. In: *Journal of Educational Data Mining* 7.2, pp. 20–48.
-  Costello, Thomas H, Gordon Pennycook, and David G Rand (2024). “Durably reducing conspiracy beliefs through dialogues with AI”. In: *Science* 385.6714, eadq1814.
-  Dettmers, Tim, Artidoro Pagnoni, Ari Holtzman et Luke Zettlemoyer (2023). « QLoRA: Efficient finetuning of quantized LLMs ». In : *Advances in Neural Information Processing Systems*. T. 36, p. 10088-10115.

-  Holstein, Kenneth, Bruce M. McLaren, and Vincent Aleven (2019). “Co-Designing a Real-Time Classroom Orchestration Tool to Support Teacher-AI Complementarity”. In: *J. Learn. Anal.* 6.2. DOI: [10.18608/JLA.2019.62.3](https://doi.org/10.18608/JLA.2019.62.3). URL: <https://doi.org/10.18608/jla.2019.62.3>.
-  Kothari, Brij (2008). “Let a billion readers bloom: Same language subtitling (SLS) on television for mass literacy”. In: *International review of education* 54.5, pp. 773–780.
-  Kumar, Harsh, David M Rothschild, Daniel G Goldstein, and Jake M Hofman (2023). “Math Education with Large Language Models: Peril or Promise?”. Presented at 2023 Conference on Digital Experimentation @ MIT (CODE@MIT).
-  Lan, Andrew S. and Richard Baraniuk (2016). “A Contextual Bandits Framework for Personalized Learning Action Selection”. In: *Educational Data Mining*. URL: <https://api.semanticscholar.org/CorpusID:15394680>.
-  Luccioni, Sasha, Yacine Jernite et Emma Strubell (2024). « Power hungry processing: Watts driving the cost of ai deployment? » In : *Proceedings of the 2024 ACM conference on fairness, accountability, and transparency*, p. 85-99.

-  Patterson, David, Joseph Gonzalez, Urs Hölzle, Quoc Le, Chen Liang, Lluís-Miquel Munguia, Daniel Rothchild, David R So, Maud Texier et Jeff Dean (2022). « The carbon footprint of machine learning training will plateau, then shrink ». In : *Computer* 55.7, p. 18-28.
-  Rasch, Georg (1960). *Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests*. Nielsen & Lydiche.
-  Seznec, Julien (Dec. 2020). “Sequential machine learning for intelligent tutoring systems”. Theses. Université de Lille. URL: <https://theses.hal.science/tel-03490620>.
-  Shabana, KM, Chandrashekar Lakshminarayanan, and Jude K Anil (2022). “CurriculumTutor: An Adaptive Algorithm for Mastering a Curriculum”. In: *International Conference on Artificial Intelligence in Education*. Springer, pp. 319–331.
-  Vassoyan, Jean, Jill-Jênn Vie, and Pirmin Lemberger (2023). “Towards Scalable Adaptive Learning with Graph Neural Networks and Reinforcement Learning”. In: *Proceedings of the 16th International Conference on Educational Data Mining, EDM 2023, Bengaluru, India*.

-  Verhelst, Norman D (2017). “Balance: A neglected aspect of reporting test results”. In: *Cognitive Abilities and Educational Outcomes: A Festschrift in Honour of Jan-Eric Gustafsson*, pp. 273–293.
-  Vie, Jill-Jênn, Fabrice Popineau, Yolaine Bourda, and Éric Bruillard (2017). “A Review of Recent Advances in Adaptive Assessment”. In: *Learning analytics: Fundamentals, Applications, and Trends*. Springer, pp. 113–142.
-  Vie, Jill-Jênn, Fabrice Popineau, Françoise Tort, Benjamin Marteau, and Nathalie Denos (2017). “A Heuristic Method for Large-Scale Cognitive-Diagnostic Computerized Adaptive Testing”. In: *Proceedings of the Fourth (2017) ACM Conference on Learning @ Scale*. ACM, pp. 323–326.
-  Weidinger, Laura, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, et al. (2022). “Taxonomy of risks posed by language models”. In: *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 214–229.

 Yessad, Amel (Sept. 2022). “Personalizing the Sequencing of Learning Activities by using the Q-Learning and the Bayesian Knowledge Tracing”. In: *17th European Conference on Technology-Enhanced Learning*. Toulouse, France. URL: <https://inria.hal.science/hal-03710500>.