UNIVERSITÉ LAVAL

IA générative : opportunités, défis et risques des Modèles massifs du langage (MMLs LLMs)

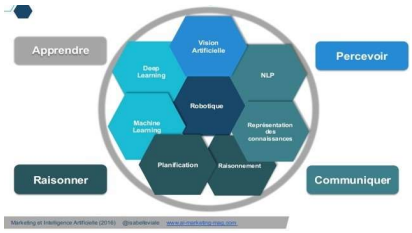
Pr. Brahim Chaib-draa

1

UNIVERSITÉ LAVAL

Cette présentation est soutendue par

Les MMLs peuvent-ils jouer un rôle important dans l'avancée de l'IA ?



2

UNIVERSITÉ LAVAL

Plan de la présentation

- Aperçu de IA générative et des MMDs
- Opportunités des MMDs
  Apports/possibilités/utilisations etc. (+ voire +++)
- Défis des MMDs
  Enjeux/difficultés/obstacles
- Risques des MMDs
  Dangers/abus/violation etc. (- voire ---)
- Futur et conclusion

3

IA vs ML vs IA Générative

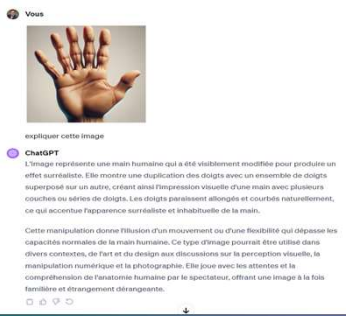
Caractéristiques	IA	AM	Générative IA
Objectif	Techniques informatiques visant à concevoir des systèmes visant à raisonner, planifier, apprendre, agir	Prédire et prendre des décisions une fois un apprentissage fait à partir des données	Génère un « nouveau » ensemble de données similaires à un ensemble de données à partir duquel il a appris
Données	G-IA⊆ AM ⊆ IA	L'AM apprend à partir de données pour prédictions/décisions sur de nouvelles données	G-IA produit de nouvelles données # des données d'apprentissage mais qui partagent certaines caractéristiques

4

Exemple d'image générée

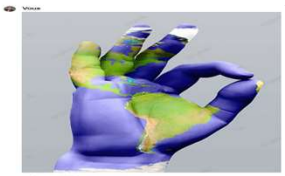
5

Génération d'explication à partir de l'image



6

Re-explications



explique cette image

ChatGPT

L'image téléchargée montre une main humaine qui a été peinte ou modifiée numériquement pour ressembler à une carte du monde. Les continents comme l'Amérique du Sud, l'Afrique et des parties de l'Amérique du Nord et de l'Europe sont représentés sur la main. Cette représentation peut avoir plusieurs interprétations symboliques :

1. **Responsabilité environnementale :** L'image peut symboliser que la santé et le futur de notre planète sont littéralement entre nos mains, ce qui souligne notre responsabilité individuelle et collective de prendre soin de l'environnement.
2. **Unité mondiale :** La main pourrait également représenter l'unité et l'interconnexion de toutes les nations du monde, avec une emphase sur la similarité globale.
3. **Art et créativité :** Cela pourrait aussi être une œuvre d'art visuelle et engageante, utilisant le corps comme toile pour une œuvre d'art visuelle et engageante.
4. **Géographie et humanité :** Elle pourrait aussi évoquer la manière dont l'humanité façonne et est façonnée par la géographie terrestre.

Cette image est souvent utilisée dans des contextes éducatifs, artistiques ou comme déclaration visuelle dans des campagnes de sensibilisation.

7

Génération de texte

Vous

explique l'IA générative et donne un exemple sous forme de texte

ChatGPT

L'intelligence artificielle (IA) générative désigne un type de technologie d'IA qui peut générer de nouvelles données qui ressemblent à des données fournies lors de l'entraînement. Ces données peuvent être des images, du texte, de la musique, ou même des simulations de données. Les IA génératives apprennent à partir d'un grand ensemble de données et utilisent ces informations pour créer du contenu nouveau et original.

Fonctionnement de Base

1. **Apprentissage:** L'IA générative étudie un grand ensemble de données (par exemple, des textes, des images, ou de la musique) et apprend les modèles, les structures, et les caractéristiques de ces données.
2. **Création:** Après l'apprentissage, l'IA peut générer de nouvelles données qui ressemblent à celles qu'elle a étudiées. Cette création se fait souvent à partir d'un point de départ aléatoire ou d'une instruction spécifique (comme un texte descriptif).
3. **Affinage:** En fonction du type d'IA générative utilisée, elle peut affiner ses créations pour les rendre plus précises ou réalistes.

Exemple en Texte

Contexte: Supposons que nous ayons une IA générative formée sur des œuvres littéraires classiques.

Tâche: Créer un début de conte de fées.

Sortie de l'IA:

"Par une nuit d'étoiles scintillantes, dans un royaume lointain et oublié, vivait un vieux sage nommé Alaric. Dans sa tour d'ivoire, entourée de forêts anciennes et de mystères, Alaric passait ses jours à étudier les secrets de l'univers. Les villageois le craignaient pour sa connaissance des arts anciens, mais le respectaient pour sa sagesse. Un jour, une jeune fille nommée Elara vint frapper à sa porte, cherchant la clé d'une énigme qui avait tourmenté son village pendant des générations..."

Dans cet exemple, l'IA utilise son apprentissage des structures narratives, des archétypes de personnages et des styles littéraires pour générer un nouveau texte qui ressemble à un conte de fées classique. Il est important de noter que le contenu généré est original, bien qu'inspiré des données d'entraînement.

8

IA générative pour le texte

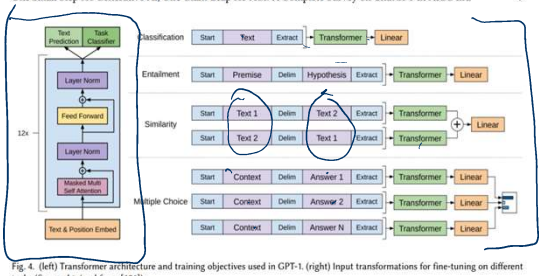


Fig. 4. (left) Transformer architecture and training objectives used in GPT-1. (right) Input transformations for fine-tuning on different tasks (figure obtained from [136]).

9

UNIVERSITÉ LAVAL

Processus d'évolution des LLMs (MMLs)

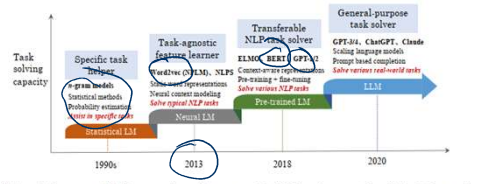


Fig. 2: An evolution process of the four generations of language models (LM) from the perspective of task solving capacity. Note that the time period for each stage may not be very accurate, and we set the time mainly according to the publish date of the most representative studies at each stage. For neural language models, we abbreviate the paper titles of two representative studies to name the two approaches: NPLM [1] ("*A neural probabilistic language model*") and NLPS [2] ("*Natural language processing (almost) from scratch*"). Due to the space limitation, we don't list all representative studies in this figure.

A Survey of Large Language Models

10

UNIVERSITÉ LAVAL

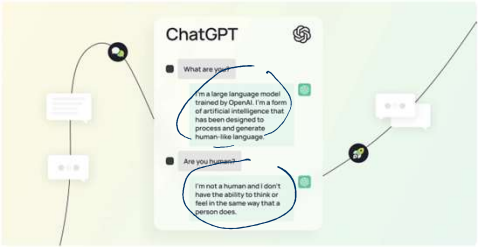
Autres Modèles Massifs du Langage

- BERT, LamdDA et PaLM développés par Google
- Llama développé par Meta)
- BLOOM modèle multilingues du langage (BigScience)
- GLM-130B développé par Tsinghua University (Chine)
- GPT 3.5 et GPT4 : Successeur de GPT3 par OpenAI
- Bien d'autres
- DALL-E 2 : Open AI pour générer des images basées sur des descriptions "prompts"

11

UNIVERSITÉ LAVAL

De GPT à ChatGPT



From Ultimate.ai

https://chat.openai.com

12

Opportunités (O) des MMLs: applis

Language	Visual	Auditory
Generate and complete text	Image generation	Music generation
Code documentation	Video generation	Voice synthesis
Answer questions and support research	3D models	Voice cloning
Design proteins and drug descriptions	Optimize imagery for healthcare diagnostics	
Supplement customer support experiences	Create immersive storytelling and video game experiences	
Generate synthetic data	Synthetic media	
Code generation	Procedural generation	

13

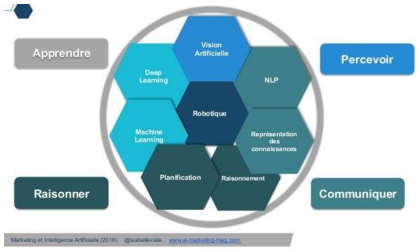
Opportunités (O) des MMLs

- **Productivité accrue :** Les MMLs peuvent améliorer la productivité dans de nombreux secteurs. Par exemple, ils peuvent rédiger des courriels, générer du code, créer du contenu ou aider à la recherche, ce qui rend les flux de travail plus efficaces.
- **Éducation et formation** : les MMLs peuvent être utilisés comme tuteurs, fournissant des explications sur une variété de sujets et aidant les élèves à apprendre à leur propre rythme.
- **Créativité améliorée** : De la rédaction d'histoires à la génération de paroles de musique, les MMLs peuvent être utilisés pour stimuler la créativité et fournir de l'inspiration.
- **Capacités multilingues :** Les MMLs peuvent comprendre et générer du contenu dans plusieurs langues, combler les lacunes de communication et permettre une communication globale plus inclusive.
- **Accessibilité** : Les MMLs peuvent aider les personnes ayant des capacités différentes, par exemple en convertissant le texte en parole ou vice versa.
- **Recherche :** Les MMLs peuvent faciliter l'analyse des données, la revue de la littérature ou même la génération d'hypothèses.

14

Opportunités (O) : Couvrir tout l'IA ?

Les MMLs peuvent-ils jouer un rôle important dans l'avancée de l'IA ?



15

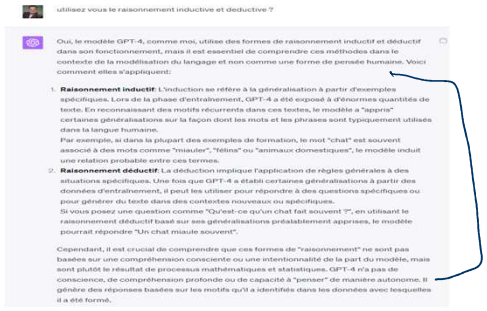
O : MMLs peuvent traiter le langage naturel

- **La catégorisation de contenu**. Un résumé linguistique du document, comprenant la recherche et l'indexation, des alertes de contenu, etc.
- **Modélisation et découverte automatique de thèmes**. Capture précise du sens et des thèmes cachés dans une collection de textes,.
- **Extraction contextuelle**. Extraction automatique d'informations structurées prenant en compte le contexte de la phrase à partir de sources textuelles brutes.
- **Analyse de sentiment**. Identification de l'humeur ou des opinions subjectives dans de larges volumes de textes, et identification automatique des sentiments.
- **Conversion parole-texte et texte-parole**. Transformation des commandes vocales en texte écrit, et inversement.
- **Synthèse de documents**. Génération automatique de résumés.
- **Traduction automatique**. Traduction automatique de textes ou discours d'une langue source vers une langue cible.

https://www.sas.com/fr_ca/insights/analytics/what-is-natural-language-processing-nlp.html

16

O : Raisonnement par induction et déduction



utilisez vous le raisonnement inductive et deductive ?

Oui, le modèle GPT-4, comme moi, utilise des formes de raisonnement inductif et déductif dans son fonctionnement, mais il est essentiel de comprendre ces méthodes dans le contexte de la modélisation du langage et non comme une forme de pensée humaine. Voici comment elles s'appliquent:

1. **Raisonnement inductif**: L'induction se réfère à la généralisation à partir d'exemples spécifiques. Lors de la phase d'entraînement, GPT-4 a été exposé à d'énormes quantités de texte. En reconnaissant des motifs récurrents dans ces textes, le modèle a "appris" certaines généralisations sur la façon dont les mots et les phrases sont typiquement utilisés dans la langue humaine.
   Par exemple, si dans la plupart des exemples de formation, le mot "chat" est souvent associé à des mots comme "miauler", "félins" ou "animaux domestiques", le modèle induit une relation probable entre ces termes.
2. **Raisonnement déductif**: La déduction implique l'application de règles générales à des situations spécifiques. Une fois que GPT-4 a établi certaines généralisations à partir des données d'entraînement, il peut les utiliser pour répondre à des questions spécifiques ou pour générer du texte dans des contextes nouveaux ou spécifiques.
   Si vous posez une question comme "Qu'est-ce qu'un chat fait souvent ?", en utilisant le raisonnement déductif basé sur ses généralisations préalablement apprises, le modèle pourrait répondre "Un chat miaule souvent".

Cependant, il est crucial de comprendre que ces formes de "raisonnement" ne sont pas basées sur une compréhension consciente ou une intentionnalité de la part du modèle, mais sont plutôt le résultat de processus mathématiques et statistiques. GPT-4 n'a pas de conscience, de compréhension profonde ou de capacité à "penser" de manière autonome. Il génère des réponses basées sur les motifs qu'il a identifiés dans les données avec lesquelles il a été formé.

17

O : Formuler et raisonner

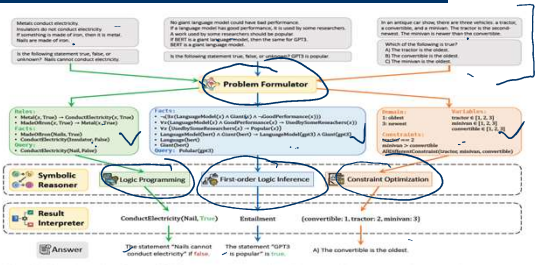


Figure 2: Overview of our LOGIC-LM model, which consists of three modules: (1) *Problem Formulator* generates a symbolic representation for the input problem with LLMs via in-context learning (2) *Symbolic Reasoner* performs logical inference on the formulated problem, and (3) *Result Interpreter* interprets the symbolic answer.

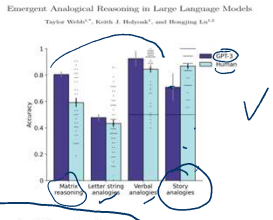
LOGIC-LM: Empowering Large Language Models with Symbolic Solvers for Faithful Logical Reasoning by L. Pan et al.

18

UNIVERSITÉ LAVAL

O : Raisonnement par analogie

- P est similaire à Q à certains égards
- P englobe certaines caractéristiques – R
- Donc, probablement Q a aussi R.



Analogical prompting

Specifically, given a target problem to solve x, our prompt augments it with instructions like:

```
# Problem: [x]
# Relevant problems: Recall three relevant and distinct problems. For each
problem, describe it and explain the solution.
# Solve the initial problem:
```


19

UNIVERSITÉ LAVAL

O : Raisonnement type « Système Expert »

Si on vous entraîne sur une base de connaissances type règles, pouvez mimer un système expert?

Oui, GPT-4 peut être utilisé pour mimer le comportement d'un système expert dans une certaine mesure, en particulier si les règles ou les directives sont clairement définies. Cependant, il y a des nuances et des limitations à prendre en compte.

1. **Base de connaissances**: Si GPT-4 est "informé" (c'est-à-dire si on lui présente des informations pendant une session ou si ces informations font partie de son entraînement initial) de certaines règles ou logiques spécifiques, il peut générer des réponses en fonction de ces règles. Par exemple, si vous fournissez des règles pour diagnostiquer une maladie à partir de symptômes donnés, GPT-4 peut utiliser ces règles pour faire des recommandations basées sur les symptômes que vous mentionnez ensuite.
2. **Mémoire à court terme**: Un point clé à comprendre est que GPT-4 n'a pas de mémoire persistante entre les sessions. Tout ce que vous "enseignez" à GPT-4 pendant une session est oublié une fois cette session terminée. Cela signifie que pour chaque nouvelle session, si vous souhaitez que le modèle opère en fonction d'une base de connaissances spécifique, vous devrez réintroduire ces informations.
3. **Limitations par rapport aux véritables systèmes experts**: Les systèmes experts traditionnels sont conçus pour des domaines spécifiques avec une logique et des règles explicites bien définies. Ces systèmes ont souvent des mécanismes pour gérer l'incertitude, les contradictions et la priorisation des règles. GPT-4, bien qu'il soit capable de générer des réponses basées sur des règles, n'a pas ces mécanismes sophistiqués intégrés. C'est un modèle de langage généralisé.
4. **Fiabilité**: Même si GPT-4 peut mimer un système expert, cela ne signifie pas qu'il soit aussi fiable qu'un véritable système expert conçu pour une application spécifique. L'utilisation de GPT-4 pour des applications critiques, en particulier dans des domaines tels que la médecine ou la sécurité, doit être abordée avec prudence.

En conclusion, bien que GPT-4 puisse simuler le comportement d'un système expert dans une certaine mesure, il ne remplace pas un véritable système expert conçu spécifiquement pour un domaine d'application. Cependant, pour des tâches moins critiques ou pour des applications éducatives ou expérimentales, GPT-4 pourrait offrir une alternative intéressante ou un complément aux systèmes experts traditionnels.

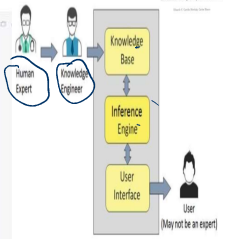


Fig. 1. Visualization of the modules and information flux of an expert system.

20

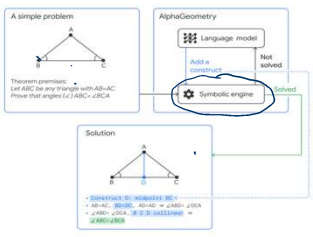
UNIVERSITÉ LAVAL

O : Raisonnement symbolique

Google DeepMind @GoogleDeepMind

AlphaGeometry is a system made up of 2 parts:
🔵 A neural language model, which can predict useful geometry constructions to solve problems
🔵 A symbolic deduction engine, which uses logical rules to deduce conclusions

Both work together to find proofs for complex geometry theorems.

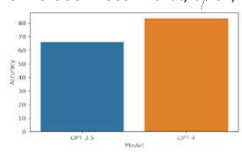


21

O : Raisonnement sens commun

On pourrait utiliser CommonsenseQA, une base de données genre QR, consistant en 12.247 exemples pour évaluer le raisonnement sens commun. Mind Meets Machine, by Sifatkaur et al.

EX: Le matin, dans ma rue déneigée, la poubelle est renversée: PossR: Chat, Chien, Raton, Camion de déneigement



GPT-3.5 66%
GPT-4 83.2%
Humain 89%

Figure 2: Accuracy of GPT-3.5 and GPT4 on CommonSenseQA

22

O : Explication du raisonnement

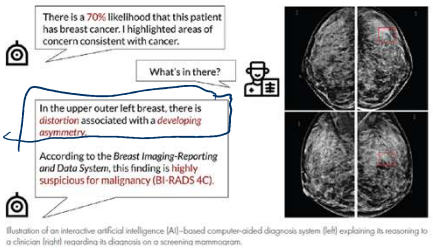


Illustration of an interactive artificial intelligence (AI)–based computer-aided diagnosis system (left) explaining its reasoning to a clinician (right) regarding its diagnosis on a screening mammogram.

ChatGPT and other Large Language Models....By Yiqin Chen et al.

23

O : Apport à la Vision

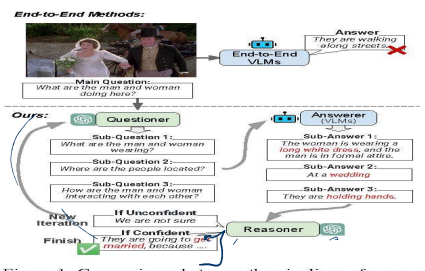
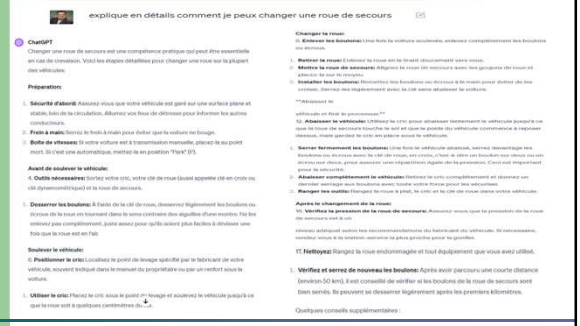


Figure 1: Comparisons between the pipelines of prevalent end-to-end VLMs (Upper) and our proposed IdealGPT (Below) for VL reasoning tasks.

IdealGPT by H. You et al.

24

O : Planification



25

O : Planification

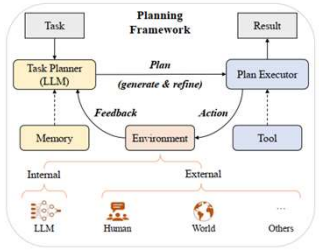


Fig. 16: An illustration of the formulation for prompt based planning by LLMs for solving complex tasks.

La planification via MMLs est à ses débuts
Benchmark < 10%

A Survey of Large Language Models

26

O : Apport à la Robotique

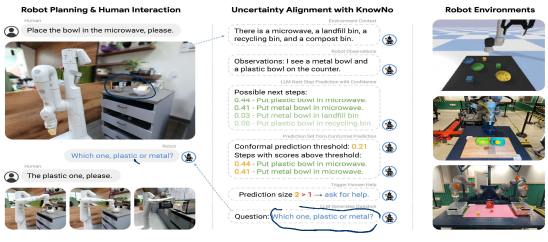


Figure 1: KnowNo uses Conformal Prediction (CP) to align the uncertainty of LLM planners. Given a language instruction, an LLM generates possible next steps and its confidences (scores) in these options. CP then provides a prediction set that includes the options with scores above a certain quantile. If there is more than one option in the set, the robot asks for help. Experiments across multiple embodiments and a variety of ambiguous situations show that KnowNo significantly improves efficiency and autonomy compared to baselines.

27

O: Analyse de sentiments (Émotions)

Analyse de sentiment. Identification de l'humeur ou des opinions subjectives dans de larges volumes de textes, et identification automatique des sentiments.

Prompt (avec émotions). Peut booster des réponses voire les mieux contextualisées

GPT3.5

Admiration	Concern	Disappointment	Excitement	Hopeless	Outrage	Resignation	Support
Amusement	Confusion	Disapproval	Fear	Hostility	Panic	Resilience	Surprise
Anger	Contempt	Disbelief	Frustration	Inspiration	Pessimism	Resolve	Sympathy
Anticipation	Contentment	Disdain	Gratitude	**Joy**	**Positive**	Respect	Trust
Anxiety	Courage	Disgust	Grief	Love	Pride	**Sadness**	Uncertainty
Appreciation	Despair	Empathy	Happiness	**Negative**	Regret	Shock	Unity
Caring	Desperation	Encouragement	Hope	**Neutral**	Relief	Solidarity	Urgency
Compassion	Determination	Encouraging	Hopeful	**Optimism**	Resentment	Sorrow	Worry

28

Défis (D) des MMLs

- **Biais de données :** Les MMLs apprennent de grandes quantités de données, qui peuvent contenir des biais. Cela signifie qu'ils peuvent parfois produire des résultats qui reflètent ces biais.
- **Coûts et Préoccupations environnementales :** La formation des MMLs nécessite des coûts faramineux et une puissance de calcul importante, entraînant une consommation d'énergie élevée et des préoccupations environnementales potentielles.
- **Dépendance excessive :** Il y a un risque que les gens deviennent trop dépendants des MMLs, conduisant à un déclin potentiel des compétences humaines dans cértains domaines.
- **Manque d'interprétabilité** : La nature « boîte noire » des MMLs rend difficile la compréhension ou l'explication de leur processus décisionnel, ce qui peut être problématique dans les applications à enjeux élevés.
- **Taille et efficacité du modèle** : À mesure que les MMLs se développent, leurs besoins de stockage et de calcul augmentent également. Le déploiement de ces modèles dans des environnements aux ressources limitées, comme les appareils mobiles, devient difficile.

29

Défis (D) des MMLs

- **Généralisation et mémorisation :** Bien que les MMLs soient conçus pour généraliser à partir des données, ils peuvent parfois mémoriser des informations, ce qui pourrait entraîner des problèmes de confidentialité.
- **Sur-ajustement (Overfitting)** : Même avec des quantités massives de données, les MMLs peuvent sur-ajuster, via un modèle, l'ensemble d'entraînement. Cela peut les rendre moins fiables dans des environnements réels et diversifiés.
- **Attaques adverses** : Les MMLs peuvent être vulnérables aux entrées adverses conçues pour les tromper ou les induire en erreur, ce qui entraîne des sorties incorrectes.
- **Fiabilité dans divers domaines :** Bien que les MMLs soient polyvalents, ils ne fournissent pas toujours des informations précises ou fiables dans tous les domaines de connaissances, en particulier dans les domaines hautement spécialisés.
- **Dépendance aux données d'entrainement:** La performance d'un MML dépend fortement de la qualité et de la diversité de ses données d'entrainement. Si les données ne sont pas représentatives ou contiennent des erreurs, le modèle héritera de ces défauts.

30

Défis (D) des MMLs

- **Réglementation et conformité :** À mesure que l'importance et l'impact des MMDs augmentent, les appels à la réglementation se multiplient. Ça peut entraver l'innovation, voir l'exemple européen

- **Mises à jour continues :** Il est difficile de s'assurer que les modèles restent à jour avec des informations nouvelles, exactes et impartiales.

- **Mesures d'évaluation :** Les mesures traditionnelles peuvent ne pas saisir l'ensemble du spectre de performance des MMDs. De nouvelles mesures et de nouveaux repères sont nécessaires pour évaluer ces modèles de façon exhaustive.

- **Oubli catastrophique** : Le réglage fin des MMDs sur les nouvelles tâches peut parfois les amener à "oublier" les informations précédemment apprises, ce qui réduit leur efficacité sur les anciennes tâches.

- **Manque d'interprétabilité :** La nature « boîte noire » des MMLs rend difficile la compréhension ou l'explication de leur processus décisionnel, ce qui peut être problématique dans les applications à enjeux élevés.

31

Défis : Hallucinations

- Les MMLs génèrent des réponses sous forme d'extrapolation à partir du prompt fourni.
- Le résultat de l'extrapolation n'est pas nécessairement étayé par les données d'entrainement, mais est plus corrélé au prompt fourni.
- Ceci peut mener à des réponses incorrectes, absurdes voire carrément inexistantes, appelées hallucinations

.

32

Défis : Hallucinations

give me 4 articles on hallucinations in LLM and rank them by relevance

Ranking by Relevance:

1. Hallucinations in Large Language Models: Understanding and Mitigating Unwanted Outputs
2. The Limits of LLMs: Why Do Models Like GPT-4 Hallucinate?
3. User Experience and LLM Hallucinations: How Do End-Users Perceive Unexpected Outputs?
4. Comparative Analysis: Hallucinations in GPT-3 vs. GPT-4

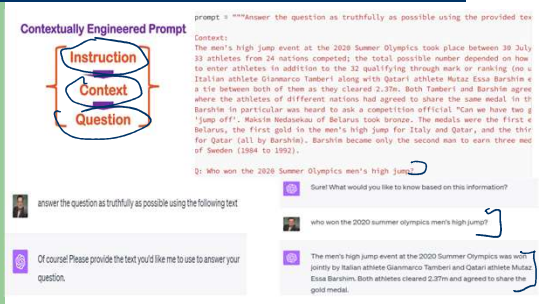
Article 4: "Comparative Analysis: Hallucinations in GPT-3 vs. GPT-4"

Summary: Through a series of tests and examinations, this article compares the frequency and nature of hallucinations between two versions of OpenAI's models, GPT-3 and GPT-4. It aims to understand if iterative versions exacerbate, mitigate, or merely change the nature of these hallucinations.

Relevance: Slightly relevant. The article focuses on a comparison but still touches upon the subject of hallucinations in LLMs.

33

Défis : Éviter les hallucinations via le contexte



34

Défis : Alignement avec les humains

Ce que dit ChatGPT : Le besoin d'alignement avec les êtres humains lorsqu'il s'agit de modèles d'apprentissage automatique comme ChatGPT découle de plusieurs raisons

- **Sécurité** : L'IA non alignée pourrait ne pas agir dans l'intérêt des humains. À mesure que les systèmes d'IA deviennent plus performants, il devient encore plus crucial de s'assurer qu'ils comprennent et respectent les valeurs et les objectifs humains.
- **Utilité** : Le but principal des systèmes d'IA est d'être utiles aux humains. Si une IA n'est pas alignée sur les valeurs humaines, elle peut fournir des réponses ou prendre des mesures qui ne sont pas bénéfiques ou même contre-productives.
- **Confiance** : Pour que les humains fassent confiance aux systèmes d'IA et les utilisent, ils doivent sentir que l'IA comprend et respecte leurs valeurs et leurs intentions. L'IA mal alignée peut éroder la confiance et entraver l'adoption.
- **Éviter les conséquences involontaires :** Un modèle d'apprentissage automatique peut, s'il n'est pas correctement aligné, amplifier les biais présents dans les données d'entraînement ou afficher des comportements qui n'étaient pas prévus par ses concepteurs. L'harmonisation aide à atténuer ces risques.

35

D: Alignement avec les humains

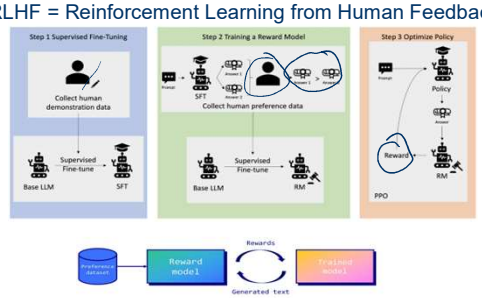
Ce que dit ChatGPT : Le besoin d'alignement avec les êtres humains lorsqu'il s'agit de modèles d'apprentissage automatique comme ChatGPT découle de plusieurs raisons

- **Interactivité** : Pour de nombreuses applications, l'IA doit travailler aux côtés des humains, qu'il s'agisse de travail collaboratif, d'aide à la décision ou de divertissement. Un alignement approprié garantit que les interactions sont fluides et bénéfiques.
- **Considérations éthiques** : À mesure que nous construisons et déployons de puissants systèmes d'IA, il est essentiel de s'assurer qu'ils fonctionnent de manière éthique et qu'ils ne causent pas de tort par inadvertance ou ne perpétuent pas d'injustices.
- **Apprentissage et adaptabilité :** Les modèles d'IA, en particulier ceux conçus pour l'apprentissage continu, ont besoin d'un alignement pour s'assurer qu'ils évoluent de manière cohérente avec les intérêts et les valeurs humaines.

36

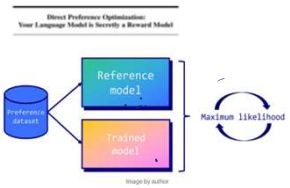
D: Alignement via l'apprentissage par renforcement

RLHF = Reinforcement Learning from Human Feedback



37

D: Alignement via DPO



Le DPO simplifie le contrôle en traitant la tâche comme un problème de classification. Concrètement, il utilise deux modèles : le modèle entraîné (ou modèle de politique) et une copie de celui-ci appelé le modèle de référence. Pendant l'entraînement, l'objectif est de s'assurer que le modèle entraîné produit des probabilités plus élevées pour les réponses préférées que le modèle de référence. À l'inverse, il devra produire des probabilités plus faibles pour les réponses pas préférées. Cela signifie que nous pénalisons le LLM pour les mauvaises réponses et le récompensons pour les bonnes.

https://towardsdatascience.com/fine-tune-a-mistral-7b-model-with-direc...

DPO > RLHF

- Stable
- Performant
- Computation moins exigeante

Défi pour l'alignement : Une approche multidisciplinaire qui tient compte des considérations techniques, éthiques et sociales est nécessaire pour développer des systèmes d'IA sûrs et bénéfiques.

38

Risques (R) des MMLs

- **Désinformation** : Si elles sont utilisées de façon malveillante, les MMLS peuvent générer des fausses nouvelles, des histoires ou des informations erronées convaincantes.
- **Problèmes de sécurité** : Les MMLs pourraient être utilisées dans des cyberattaques, pour des stratagèmes de phishing ou exploiter des vulnérabilités dans des systèmes en générant du code malveillant.
- **Préoccupations éthiques** : Décider comment utiliser ou non les résultats des LLM dans des domaines sensibles (comme les décisions juridiques ou les diagnostics médicaux) est une préoccupation importante.
- **Inégalités économiques** : Comme les entreprises ayant accès à de grandes ressources informatiques développent principalement des MMLs, il existe un risque de centralisation du pouvoir et de création de disparités technologiques.
- **Perte de vie privée** : Si les MMLs ne sont pas utilisées de façon responsable, il y a un risque potentiel de les utiliser de pour générer des informations basées sur des données privées.
- **Dévalorisation de l'expertise humaine** : Il y a un risque pour la société de sous-évaluer les experts humains dans les domaines où les MMLs deviennent importantes.

39

Risques des MMLs

- **Homogénéisation de l'information** : Si de nombreux systèmes et plateformes reposent sur quelques MMLs largement utilisées, cela peut conduire à une homogénéisation de l'information, réduisant ainsi la diversité de la pensée et de la créativité.
- **Manipulation :** Les MMLs peuvent être affinées ou adaptées à des fins nuisibles, en créant des modèles qui produisent spécifiquement du contenu dangereux ou malveillant.
- **Problèmes de sécurité :** Les acteurs malveillants peuvent utiliser des MMLs pour automatiser les tentatives d'hameçonnage, créer des attaques d'ingénierie sociale sophistiquées ou produire des spams à grande échelle.
- **Confidentialité des données** : Il est possible que les MMLs produisent par inadvertance des informations qu'ils ont vues pendant la formation, soulevant des préoccupations concernant la fuite de données ou les atteintes à la vie privée.
- **Plagiat et inconduites**
- **Hallucinations et autres limitations techniques** (réponses incorrectes, illogiques, inconstitantes)
- **Évolution vers des Agents Autonomes**

40

R: Empoisonner les MMDs

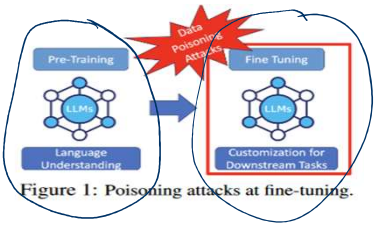


Figure 1: Poisoning attacks at fine-tuning.

Forcing Generative Models to Degenerate Ones: The Power of Data Poisoning Attacks

41

R : Plagiat et inconduites

Nouvelles séances offertes

Évaluer les apprentissages à l'ère de l'intelligence artificielle générative

La présence de plus en plus grande d'outils d'intelligence artificielle générative, notamment d'agents conversationnels tels que ChatGPT, bouscule le monde de l'enseignement supérieur.

La validité du processus d'évaluation des apprentissages de vos étudiantes et étudiants vous préoccupe ?

Au terme de cette formation, vous serez en mesure de:

- Identifier les impacts possibles de ChatGPT sur vos évaluations
- Adapter vos modalités d'évaluation à l'aide de différentes stratégies

Mercredi 13 septembre
11h30 à 12h15 (45 minutes)
- En ligne sur Zoom
- Présentation entièrement magistrale

Pour s'inscrire

Vendredi 22 septembre
9h30 à 11h30 (120 minutes)
- En présence, LAU-0415
- Présentation magistrale, discussions et activité d'exploration de ChatGPT

Pour s'inscrire

42

UNIVERSITÉ LAVAL

R: des MMLs aux Agents Autonomes

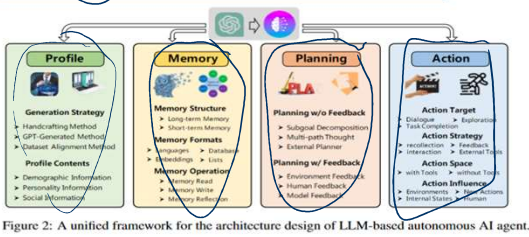


Figure 2: A unified framework for the architecture design of LLM-based autonomous AI agent.

A Survey on Large Language Model based Autonomous Agents Lei Wang et al.

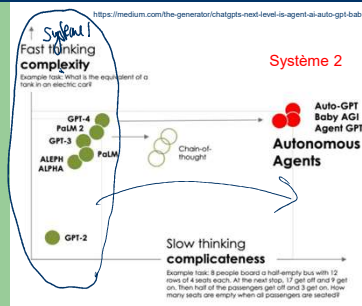
Les agents autonomes exécutent généralement des tâches en assumant des rôles spécifiques, tels que les codeurs, les enseignants et les experts de domaine Le module de profilage vise à indiquer les profils de rôle des agents, qui sont généralement écrits dans l'invite pour influencer les comportements de LLM. Dans les travaux existants, il existe trois stratégies couramment utilisées pour générer des profils d'agents

43

UNIVERSITÉ LAVAL

R : Des MMDs aux Autonomes Agents

https://medium.com/the-generator/chatgpts-next-level-is-agent-ai-auto-gpt-babyagi-agentgpt-microsoft-jarvis-friends-d354aa18f21



Système 2

Complexe désigne le niveau de composants d'un système. Si un problème est complexe, cela signifie qu'il comporte de nombreux composants. La complexité n'évoque pas la difficulté.

Compliqué se réfère à un niveau élevé de difficulté. Si un problème est compliqué, il peut y avoir ou pas beaucoup de parties, mais il faudra certainement beaucoup de travail pour le résoudre.

44

UNIVERSITÉ LAVAL

Kahneman's dual-system : System 1 et System 2

Table 1. Characteristics of system 1 and system 2

System 1	System 2
Does not require working memory	Requires working memory
Autonomous	Cognitive decoupling; mental simulation
Fast	Slow
High capacity	Capacity limited
Parallel	Serial
Non-conscious	Conscious
Biased responses	Normative responses
Contextualised	Abstract
Automatic	Controlled
Associative	Rule-based
Experience-based decision making	Consequential decision making
Independent of cognitive ability	Correlated with cognitive ability

Design thinking, fast and slow: A framework for Kahneman's dual-system theory in design by Udo Kannengiesser1 and John S. Gero2

45

MMLs → AAs : les risques

- Faible risque si MMLs → AAs se fait graduellement selon un soft takeoff des AAs
- Fort risque dû à un « Hard takeoff » des AAs
  - On est pris de court;
  - On n'a pas le temps pour des corrections
  - On n'a pa as le temps d'élaborer un certain contrôle

46

Conclusion

- IA-Générative : évolution ou révolution
- LLMs = avancées significatives/spectaculaires de l'IA
- À l'avenir il convient de faire de l'IA autre (ou en plus) que l'apprentissage automatique
- Le système 1 de Kahneman's est « presque réalisé ». Beaucoup de partants pour le système 2
- If You Think AI Is Hot, Wait Until It Meets Quantum Computing

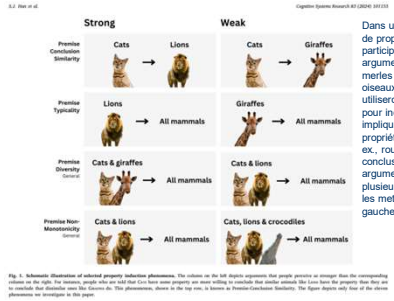
47

MERCI

48

UNIVERSITÉ LAVAL

O : Raisonnement inductif



Dans une tâche typique d'induction de propriété, on demande aux participants d'évaluer la force des arguments inductifs comme « les merles ont la propriété P, donc les oiseaux ont la propriété P ». Nous utiliserons la notation robin→ bird pour indiquer un argument qui implique de généraliser une propriété à partir d'une prémisse (p. ex., rouge-gorge) jusqu'à la conclusion (p. ex., oiseaux). Les arguments peuvent également avoir plusieurs prémisses, indiquées en les mettant entre parenthèses à gauche.