L'intelligence Artificielle

Hes·so VALAIS WALLIS

Comment utiliser IA générative de manière efficace ?

Institut Entrepreneuriat & Management
Institut Unternehmertum & Management

2025

© Dr Serge Imboden, HES-SO

1

Nos enfants
ne comprendront
jamais le lien

2

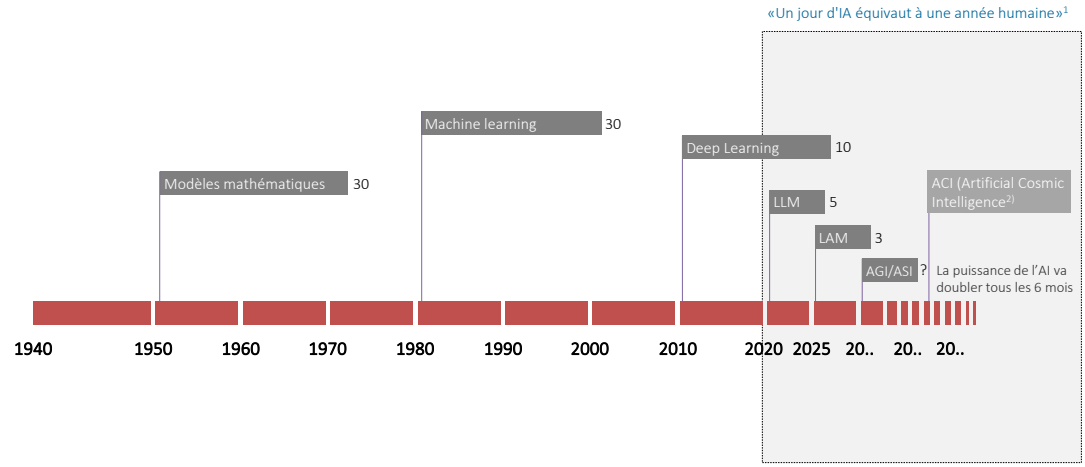
3

Agenda

1. L’accélération à l’ère de l’IA
2. Comment fonctionnent l’intelligence artificielle et les LLMs ?
3. Huit limitations et les conséquences sur l’utilisation
4. Conclusion

4

L'intelligence artificielle accélère de manière inédite le développement dans tous les domaines de notre vie

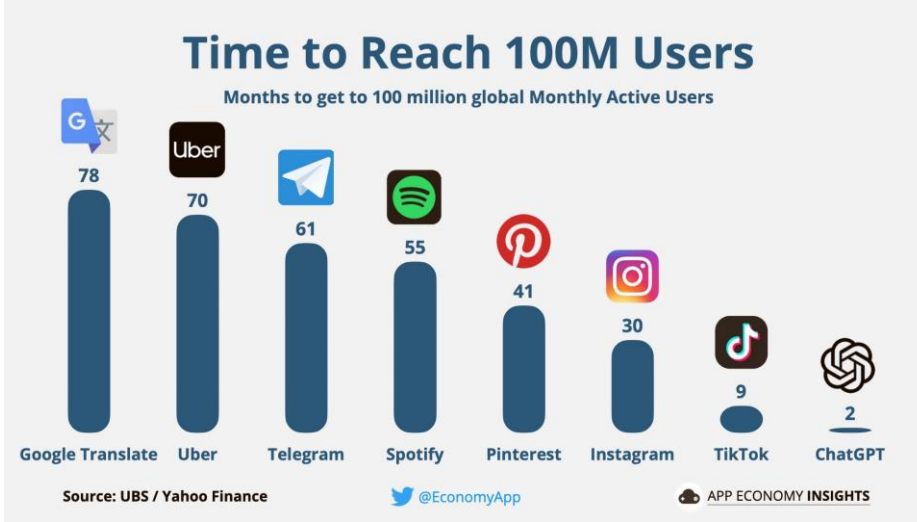


[1] Pinar Seyhan Demirdag, AI Director at Cuebric
[2] Ce terme n'est pas couramment utilisé, il pourrait être exploré dans des contextes spéculatifs, philosophiques, ou liés à la science-fiction

5

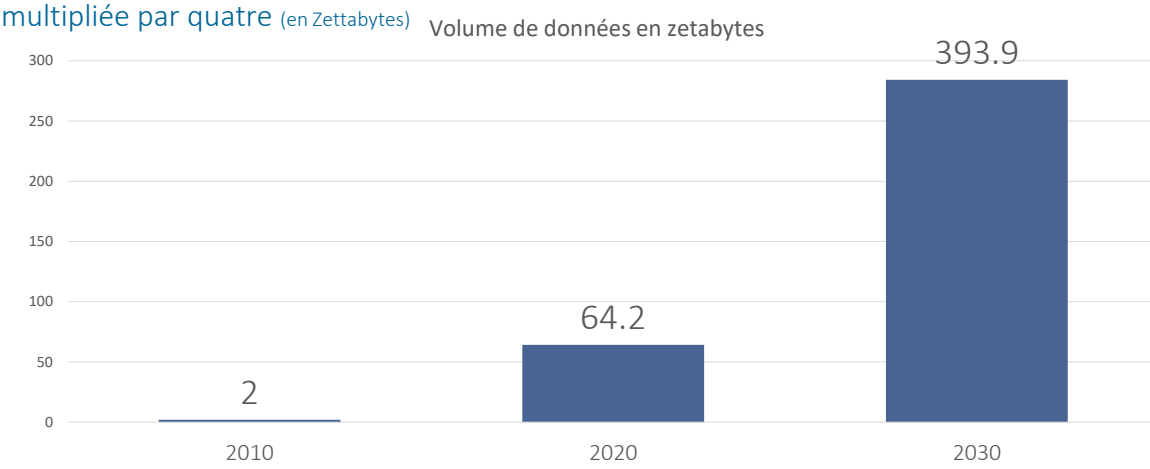
6

Combien de mois pour atteindre 100 Mio. d'utilisateurs?



7

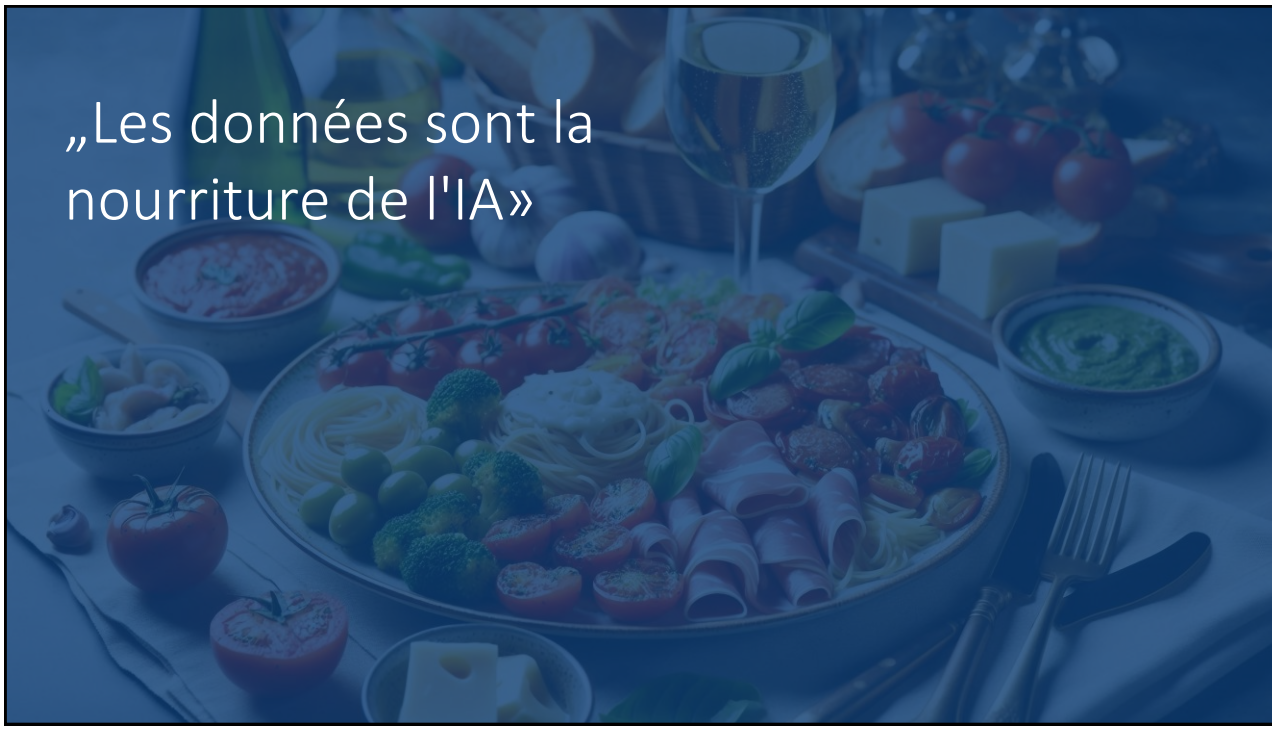
D'ici 2029, la quantité de données numériques générées chaque année dans le monde sera multipliée par quatre (en Zettabytes)



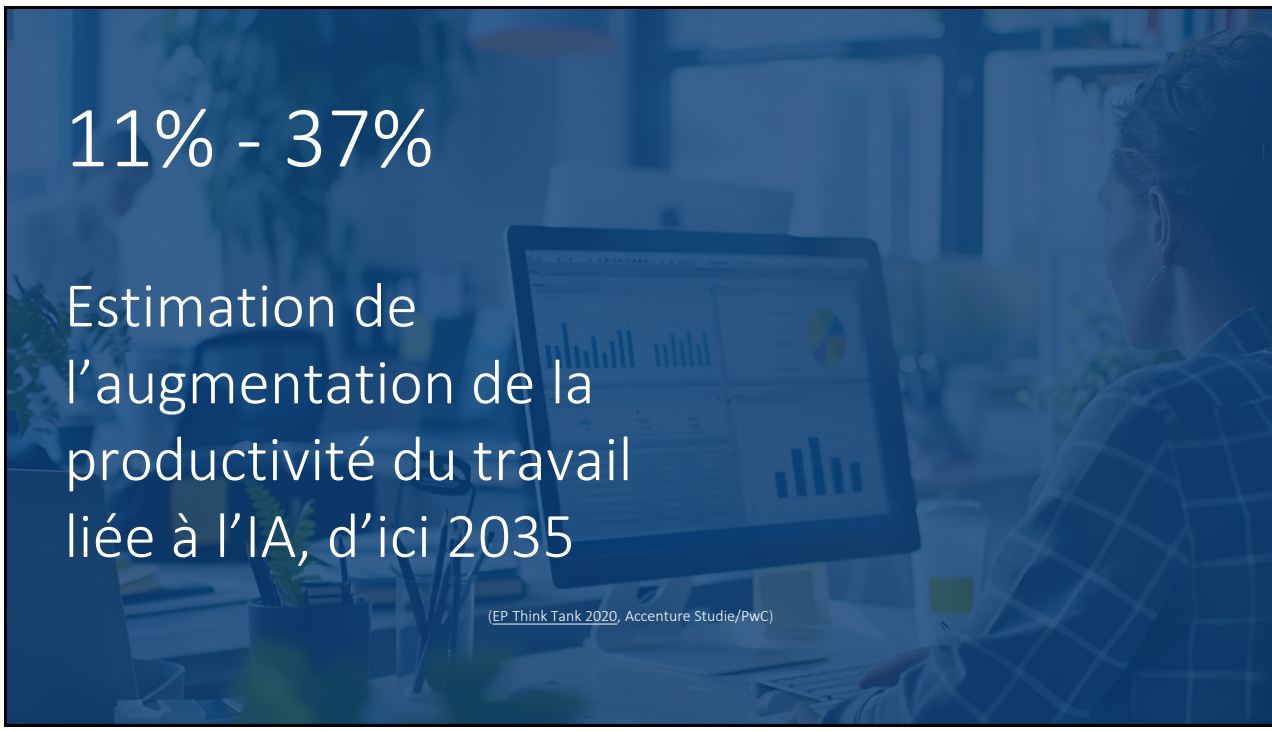
3'939'000'000'000'000'000'000'000'000 Quadrilliards d'octets

(www.statista.com, 2023)

8

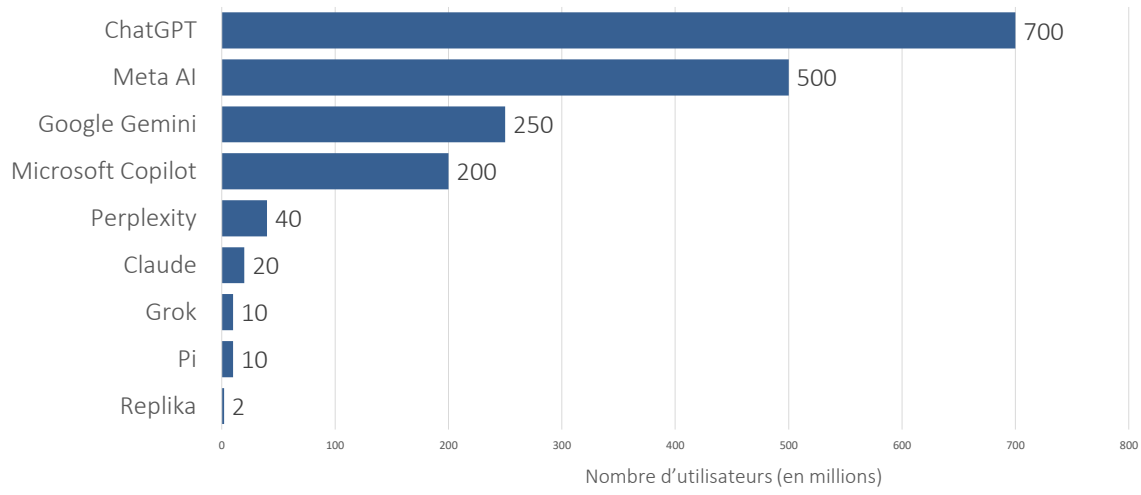

„Les données sont la nourriture de l'IA»

9


11% - 37%
Estimation de l’augmentation de la productivité du travail liée à l’IA, d’ici 2035
(EP Think Tank 2020, Accenture Studie/PwC)

10

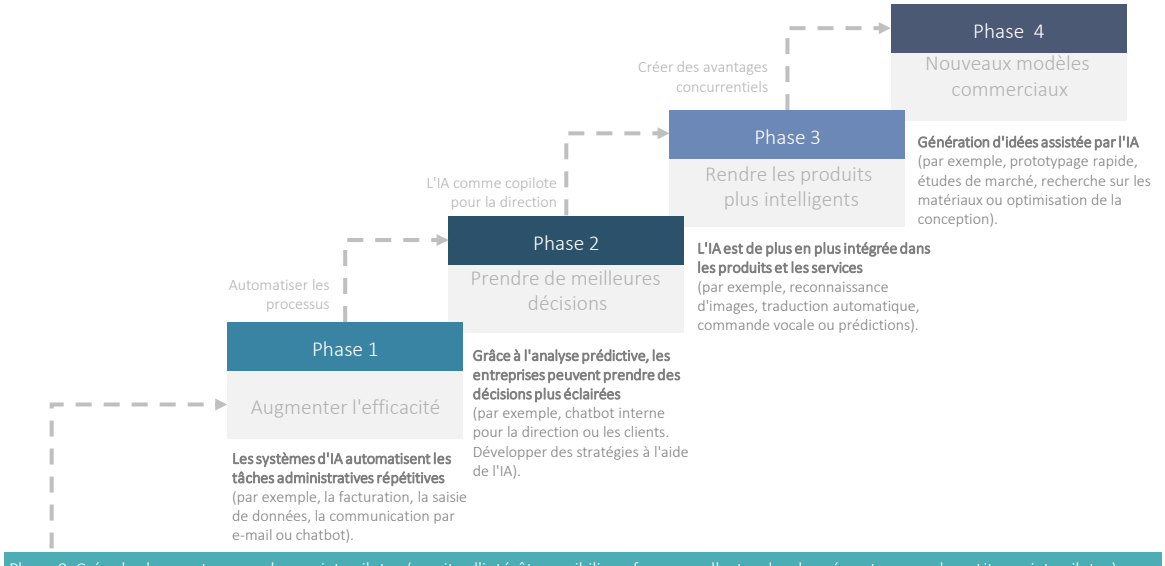
Nombre estimé d'utilisateurs des principaux chatbots d'IA générative (août 2025)



Sources: Windows Central (2025, 4. August); Exploding Topics (2025, März); FirstPageSage (2025, Juli); Lunden, I. (2025, 12. Februar), *TechCrunch*.

11

Possibilités d'utilisation de l'IA dans l'entreprise



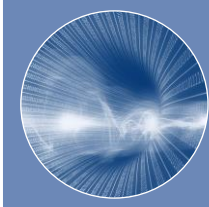
12

Agenda

1. L'accélération à l'ère de l'IA
2. Comment fonctionnent l'intelligence artificielle et les LLMs ?
3. Huit limitations et les conséquences sur l'utilisation
4. Conclusion

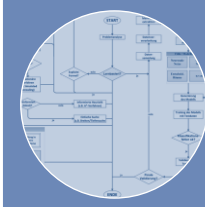
13

Il faut au moins 3 éléments pour que l'IA fonctionne ?



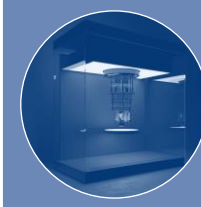
Données

- **De grandes quantités :** Les modèles d'IA ont besoin d'énormes quantités de données pour reconnaître et apprendre des modèles.
- **Diversité :** les données doivent être aussi variées que possible afin de couvrir un large éventail de situations et de scénarios.
- **Qualité :** les données doivent être propres, correctes et pertinentes pour la tâche en question.



Algorithmes (Machine learning)

- **Les algorithmes d'apprentissage :** Ces algorithmes permettent à l'IA d'apprendre et de s'adapter à partir des données.
- **Réseaux neuronaux :** les réseaux neuronaux profonds, en particulier, sont très efficaces pour les tâches complexes.
- **Apprentissage automatique :** il s'agit de développer des algorithmes qui permettent à l'IA d'apprendre sans programmation explicite.



Puissance de calcul

- **Matériel :** des processeurs puissants (GPU, TPU) et de grandes capacités de stockage sont nécessaires pour effectuer les calculs complexes.
- **Cloud computing :** l'utisation d'infrastructures en nuage permet d'accéder à d'énormes ressources de calcul.

Pour le moment...

Personnes

- **Scientifiques des données :** ils préparent les données, développent les modèles et les entraînent.
- **Experts de domaine :** ils apportent leur expertise pour adapter les applications d'IA aux besoins spécifiques.
- **Formateurs IA :** ce terme est de plus en plus utilisé pour désigner les personnes qui se concentrent spécifiquement sur l'entraînement des modèles d'IA.

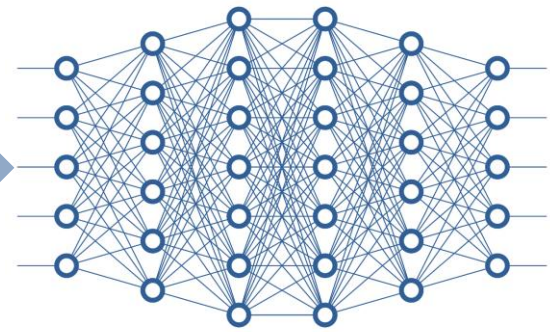
Souhaitable...

Sécurité & éthique

- **Conséquences involontaires :** les systèmes d'IA peuvent prendre des décisions complexes dont les conséquences ne sont pas toujours prévisibles. En l'absence de directives éthiques, ces systèmes pourraient prendre des décisions entraînant des dommages pour les personnes ou l'environnement.
- **Sécurité :** des systèmes d'IA puissants pourraient tomber entre de mauvaises mains et être utilisés à des fins malveillantes.

14

Des réseaux neuronaux artificiels multicouches imitent notre cerveau

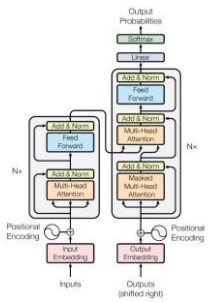


Deep Neural Networks (DNNs)

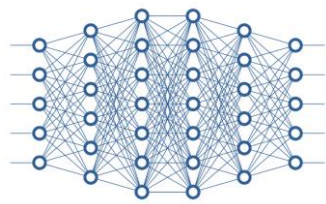
15

Les « Transformer » sont constitués de plusieurs couches de réseaux neuronaux

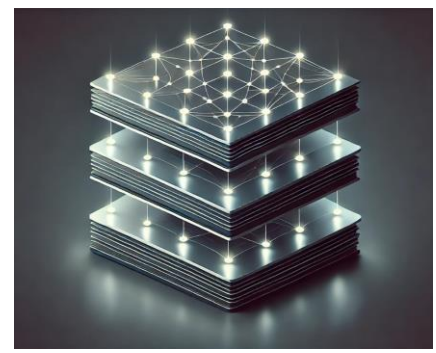
The Transformer - model architecture



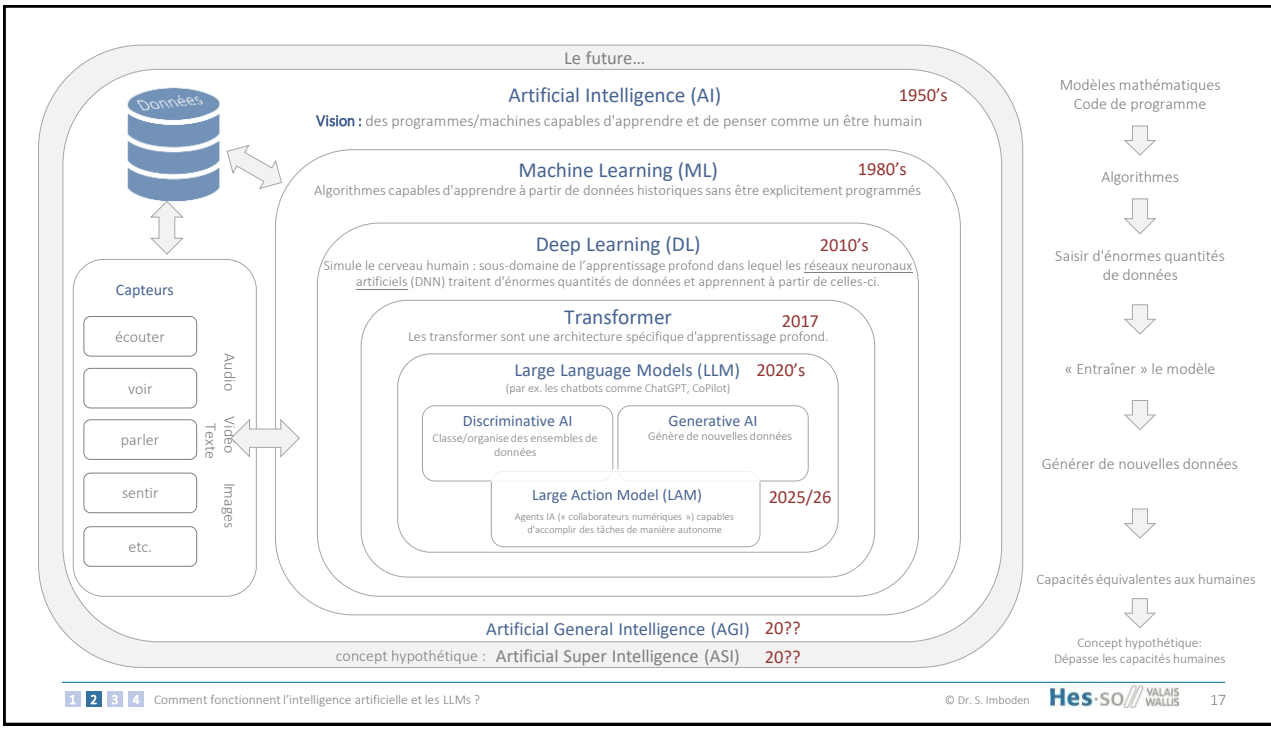
Deep Neural Networks (DNNs)



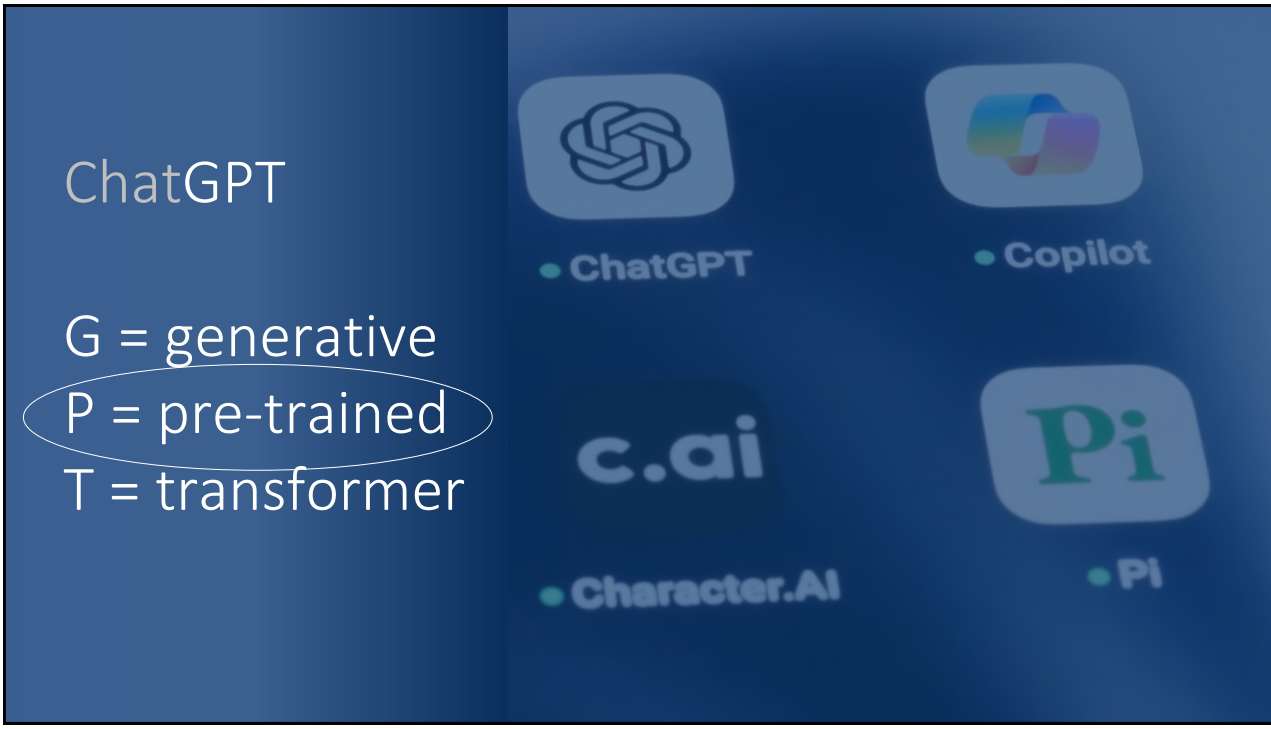
Transformer multicouche



16

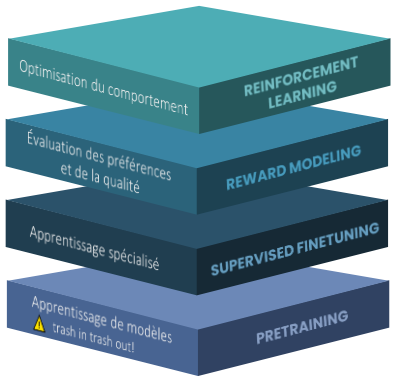

Le future...
Artificial Intelligence (AI) 1950's
Vision : des programmes/machines capables d'apprendre et de penser comme un être humain
Machine Learning (ML) 1980's
Algorithmes capables d'apprendre à partir de données historiques sans être explicitement programmés
Deep Learning (DL) 2010's
Simule le cerveau humain : sous-domaine de l'apprentissage profond dans lequel les réseaux neuronaux artificiels (DNN) traitent d'énormes quantités de données et apprennent à partir de celles-ci.
Transformer 2017
Les transformer sont une architecture spécifique d'apprentissage profond.
Large Language Models (LLM) 2020's
(par ex. les chatbots comme ChatGPT, CoPilot)
Discriminative AI
Classe/organise des ensembles de données
Generative AI
Génère de nouvelles données
Large Action Model (LAM) 2025/26
Agents IA (« collaborateurs numériques ») capables d'accomplir des tâches de manière autonome
Données
Capteurs
écouter
voir
parler
sentir
etc.
Audio
Video
Texte
Images
Artificial General Intelligence (AGI) 20??
concept hypothétique : Artificial Super Intelligence (ASI) 20??
Modèles mathématiques
Code de programme
Algorithmes
Saisir d'énormes quantités de données
« Entraîner » le modèle
Générer de nouvelles données
Capacités équivalentes aux humaines
Concept hypothétique:
Dépasse les capacités humaines
1 2 3 4 Comment fonctionnent l'intelligence artificielle et les LLMs ?
© Dr. S. Imboden
Hes·so VALAIS WALLIS
17

17


ChatGPT
G = generative
P = pre-trained
T = transformer
ChatGPT
Copilot
c.ai
Character.AI
Pi
Pi

18

Le processus d'apprentissage des LLM passe par au moins quatre phases



4. Apprentissage renforcé-> Reinforcement Learning with Human Feedback

À la dernière étape, le modèle est affiné à l'aide de récompenses et de punitions. Il doit être capable d'évaluer dans quelle mesure une réponse du modèle correspond aux attentes humaines, par exemple en termes de politesse, d'utilité et de précision. Cette phase est très exigeante en calcul, car des milliards de paramètres sont à nouveau ajustés.

3. Modèle de récompense-> Reward Model

Au lieu d'apprendre un nouveau comportement linguistique, cette phase consiste à entraîner un modèle de récompense. Celui-ci doit être capable d'évaluer dans quelle mesure une réponse du modèle correspond aux attentes humaines. Pour cela, des annotateurs humains comparent plusieurs réponses d'un LLM et indiquent celle qu'ils préfèrent. Cette évaluation manuelle demande un effort humain considérable.

2. Ajustement fin supervisé-> Supervised Fine-Tuned Model

Dans cette phase, le modèle est adapté à l'aide de jeux de données supervisés et soigneusement sélectionnés par des humains pour des tâches ou domaines spécifiques – par exemple répondre à des questions médicales, analyser des textes juridiques ou apprendre un style d'écriture particulier (transfer learning). La charge de calcul est nettement inférieure à celle du pré-entraînement, car le volume de données est plus réduit.

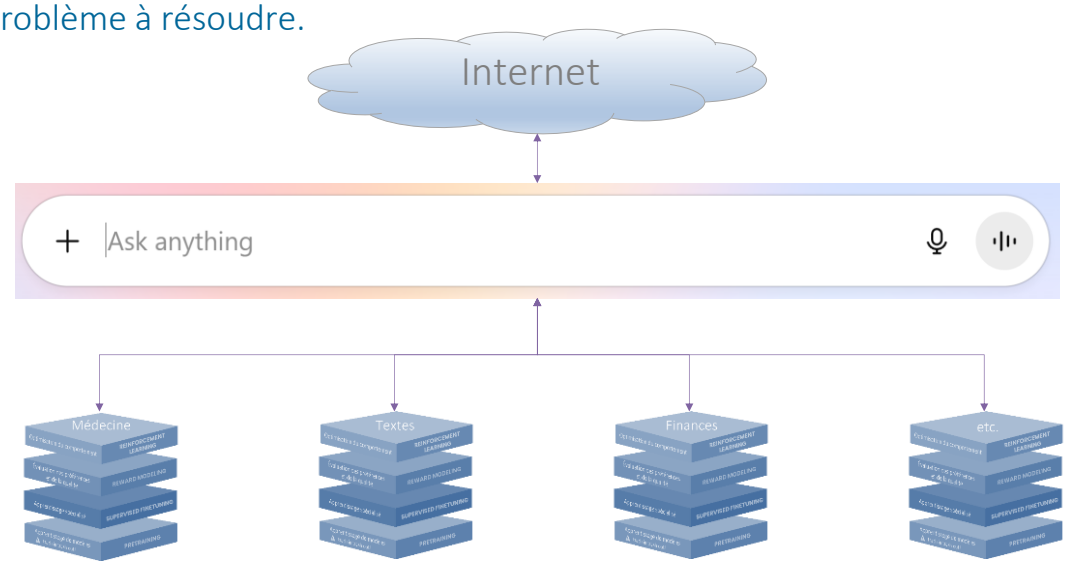
1. Pré-entraînement -> Modèle de base

Lors du pré-entraînement, le modèle apprend de manière autonome à partir d'énormes volumes de texte (corpus), provenant par exemple de contenus web, de Wikipédia, de livres ou de conversations en ligne, afin de capter les schémas linguistiques et les connaissances générales. Cette phase constitue la base de tous les apprentissages ultérieurs. GPT-3, par exemple, a été entraîné sur des milliards de phrases issues d'Internet, ce qui lui a permis d'apprendre à combler des lacunes textuelles. Cette étape est particulièrement exigeante en calcul et consommatrice de ressources.

© Dr. S. Imboden

19

Il en résulte des systèmes experts que l'on peut solliciter en fonction du problème à résoudre.



© Dr. S. Imboden

20

Agenda

1. L’accélération à l’ère de l’IA
2. Comment fonctionnent l’intelligence artificielle et les LLMs ?
3. Huit limitations et les conséquences sur l’utilisation
4. Conclusion

21

22

Limitation No. 1: TITO

23

La règle de base en informatique = TITO

Trash in → Trash out

24

⚠ TITO - Conséquences sur l'utilisation

- **Un mauvais prompt produit un mauvais résultat.**
  → Des questions imprécises ou ambiguës fournissent des réponses vagues, superficielles ou erronées.
- **Des données d'entrée erronées ou obsolètes produisent des sorties incorrectes.**
  → Si, par exemple, un document erroné est téléchargé ou un tableau incorrect est analysé, la sortie se fonde sur ces erreurs – parfois sans avertissement.
- **Une utilisation non réfléchie de l'IA renforce les biais existants.**
  → Lorsque des données biaisées ou unilatérales sont fournies, l'IA peut traiter et reproduire ces biais sans esprit critique.

25

Limitation No. 2: Les token

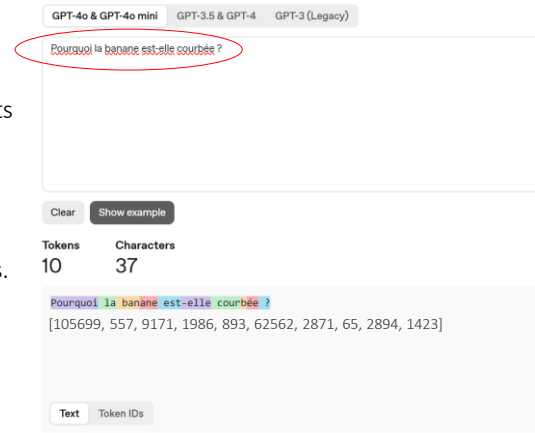
26

La langue des LLM, ce sont les « tokens ».

Pour l'IA, les mots ne sont pas du texte – ce sont des nombres. Et ces nombres forment des motifs géométriques dans l'espace.

Les tokens sont les éléments constitutifs des mots en nombres :

- *Pomme* = 1 token = [11756, 27849]
- *Je mange une pomme = 5 tokens*
- Un mot long peut comporter plusieurs tokens.



https://platform.openai.com/tokenizer

27

Les LLM ont un accès contextuel limité : ils « voient » seulement une partie

Modèle	Contexte max. (tokens)	Correspond approximativement à
GPT-3.5	env. 4'096 token	~ 3–4 pages texte
GPT-4 (standard)	env. 8'192–32'768 token	~ 6–25 pages texte
GPT-5 (128k)	Jusqu'à 128'000 token	env. 100 pages (mais techniquement complexe)

⚠ Le modèle « oublie » les parties plus anciennes de la conversation (elles sont coupées). Même pour les documents longs (p. ex. PDF), seule une partie est analysée.

28

⚠ Limitation des tokens- Conséquences sur l'utilisation

- **Les textes longs ne sont traités que partiellement.**
  → En cas de trop grand nombre de tokens, l'IA n'analyse qu'un extrait – des passages importants peuvent être ignorés.
- **Les informations antérieures dans l'historique du chat se perdent.**
  → Dans des conversations longues, l'IA « oublie » les messages plus anciens, ce qui entraîne des répétitions ou des contradictions.
- **Les résumés et les analyses peuvent être incomplets.**
  → Si le volume du texte dépasse la limite de tokens, la sortie n'est plus globale.
- **Soyez précis dans vos demandes (promptes)**
  → Plus votre demande est précise, plus la réponse sera pertinente

29

Limitation No. 3: La mémoire

30

Les LLM n’ont (encore) aucune mémoire.

Humain

Je me souviens

VS

Intelligence artificielle

Je consulte

À chaque prompt, ChatGPT doit relire ce qui a été dit auparavant

ChatGPT 5 peut actuellement relire au maximum env. 128 000 tokens (env. 100 pages). Le reste est coupé !

31

En revanche, ChatGPT, par exemple, enregistre des données te concernant. Cela vise à rendre les réponses plus cohérentes.

Cependant, ces données personnelles peuvent être consultées et supprimées !

32

⚠ Problème de mémoire - Conséquences sur l'utilisation

- **Chaque session commence à zéro.**
  → Sans fonction de mémoire, l'IA ne connaît ni les sujets précédents, ni les préférences, ni les corrections – tu dois tout réexpliquer.
- **Pas de véritable planification à long terme possible.**
  → L'IA ne peut pas poursuivre des objectifs ou des stratégies transversales sur plusieurs conversations (p. ex. planification d'un projet sur plusieurs semaines).
- **La cohérence du fil doit être assurée manuellement.**
  → Dans les dialogues longs, il faut réintroduire les informations précédentes, sinon elles se perdent et des incohérences apparaissent.
- **L'utilisateur doit donc segmenter.**
  → Diviser le document en sections logiquement structurées (p. ex. par chapitre : « Introduction », « Méthodologie », « Résultats quantitatifs », etc.).

33

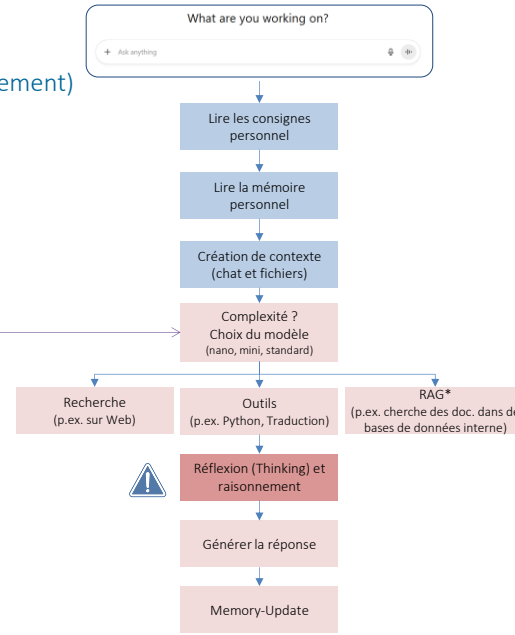
Limitation No. 4: Coûts - Performance

34

Coûts - Avantages - Performances

(Le principal facteur de coût est la profondeur du raisonnement)

- **Nano :** très rapide, bon marché, pour des tâches simples, bon marché
- **Mini :** équilibré, complexité moyenne
- **Standard/Thinking :** performances maximales,, pour des analyses complexes, cher

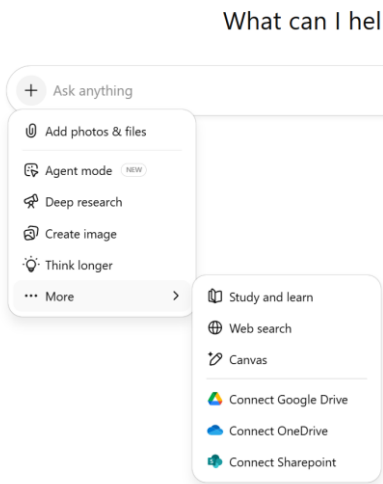


*Retrieval-Augmented Generation

© Dr. S. Imboden

35

Vous pouvez manuellement augmenter la performance



⚠ Plus de puissance ne signifie pas automatiquement plus de vérité !

(Une personne disposant d'un bureau plus grand peut étaler plus de dossiers à la fois, mais cela ne signifie pas pour autant qu'elle tire automatiquement de meilleures conclusions.)

© Dr. S. Imboden

36

⚠ Coûts - Avantages - Performances : Conséquences sur l'utilisation

- **Conception efficace des invites (prompts)**
  → [RÔLE] + [CONTEXTE] + [OBJECTIF] + [FORME/STYLE] + [LIMITES]
- **Erreurs courantes des invites**
  → Trop vagues, trop longues, sans structure ou mélangeant plusieurs tâches
- **Techniques avancées**
  → Chain-of-Thought-instructions: « Réfléchis étape par étape. » => Oblige à réfléchir entre chaque étape.
  → Few-Shot-exemples: Fournir des exemples de réponses pour contrôler le format et le ton.
  → Intégrer l'autocontrôle : « Donne-moi d'abord tes réflexions brutes, puis vérifie-les pour détecter les erreurs logiques, et ensuite seulement donne ta réponse finale. »
- **Une bonne invite**
  → permet non seulement d'économiser des ressources informatiques, mais réduit également la probabilité que le routeur doive passer à un modèle plus coûteux, car un modèle plus petit peut déjà fonctionner avec précision.

37

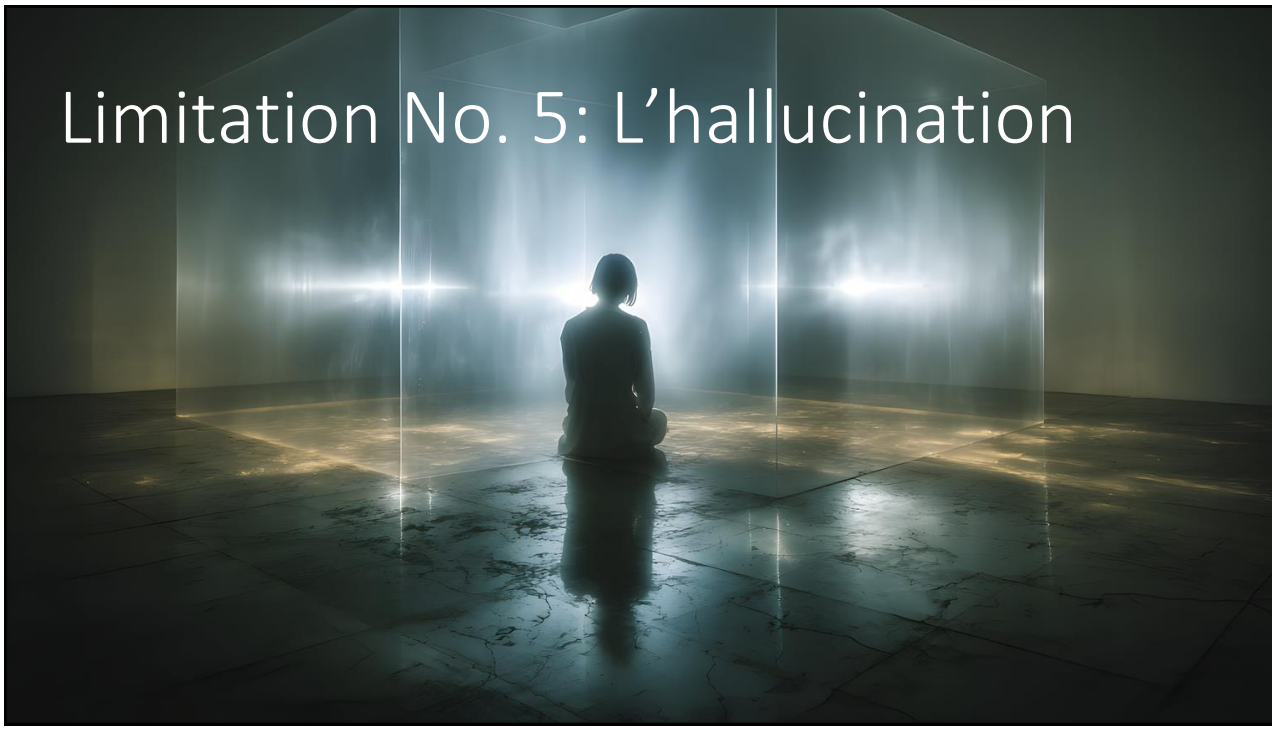
Comment puis-je être plus prompt ?

1	**Donne un rôle...** (p. ex. agis en tant qu'écrivain, agis en tant qu'expert, agis en tant qu'apiculteur, agis en tant que Youtubeur)	Rôle
2	**Explique la tâche ...** (p. ex. rédige un essai universitaire, donne-moi une recette pour corriger ce texte, analyse le texte suivant)	Tâche
3	**Fixe les restrictions / le contexte...** (p. ex. donne la priorité à l'exactitude, écris avec des phrases courtes, utilises un style d'écriture poétique, soit concis, ajoutes beaucoup de détails, mets en évidence ce qui fait de chaque méthode un choix exceptionnel, voici un example.)	Règles
4	**Définis le format de l'output...** (p .ex. réponds en points de liste, inclus un tableau, inclus des emojis, inclus des sous-titres, inclus un graphique, fais-moi un PowerPoint, sors-moi un document Word)	Format
5	**Détermine à quel moment la tâche est accomplie ...** (p .ex. la tâche est terminée lorsque trois méthodes vérifiées sont renvoyées dans le format spécifique, Tu peux t'arrêter dès qu'au moins cinq sources d'études scientifiques fondées)	Stopp

- → **Sois spécifique :** Plus ta demande est précise, plus la réponse sera pertinente
- → **Ajoute des exemples :** Fournis un exemple du style ou du format attendu peut aider l'IA à mieux comprendre
- → **Testes et itère :** Si la première réponse n'est pas parfaite, précises ce qui doit être ajusté dans ton prochain prompt

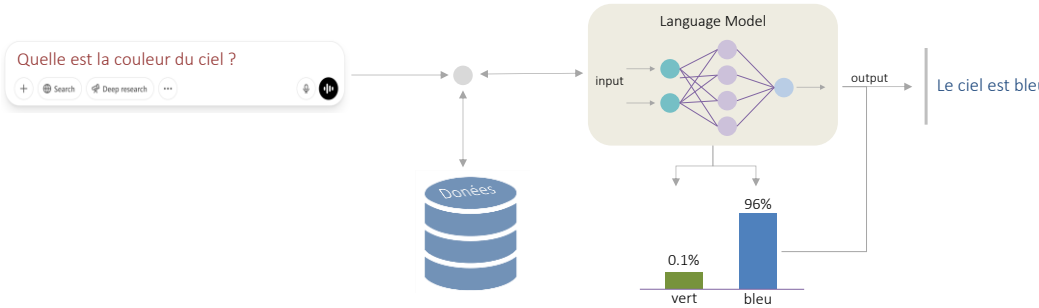
38

Limitation No. 5: L'hallucination



39

Les LLM ne sont pas des machines à savoir – ce sont des machines à probabilités.



Les LLM sont probabilistes, et non déterministes.

« Probabiliste » = le modèle donne le mot suivant le plus probable, pas nécessairement le vrai (génération contrefactuelle).

40

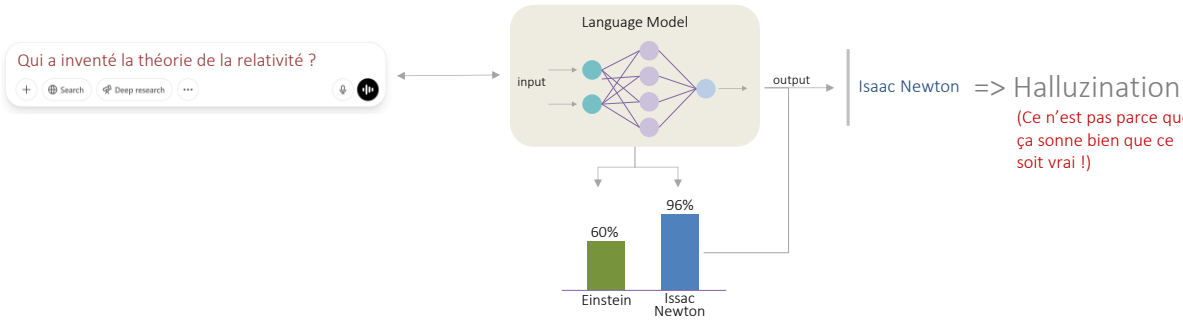
Aucun texte issu d'un LLM n'est, par nature, digne de confiance – seulement probable.

Les LLM peuvent halluciner !

41

Les LLM hallucinent et ne disent pas toujours la vérité.

Hallucinations = invention de contenu – le modèle génère des informations qui semblent plausibles, mais qui sont factuellement fausses ou entièrement inventées.



→ Les LLM n'apprennent pas des « connaissances », mais les probabilités de suites de mots.
→ Ils inventent des liens lorsque le contexte est incomplet ou ambigu.
→ Il n'existe actuellement aucune mémoire factuelle intégrée – seulement une reconnaissance de motifs.

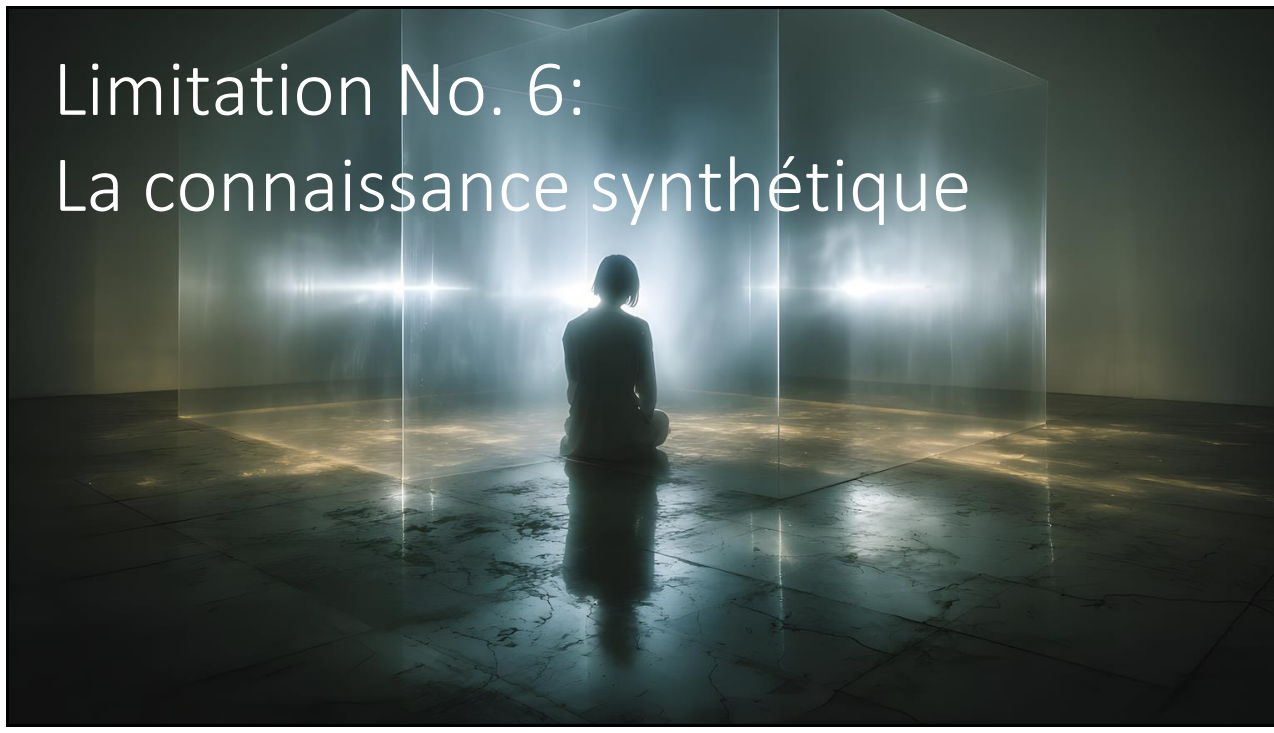
42

⚠ Hallucinations : Conséquences sur l'utilisation

- **La fiabilité doit être vérifiée activement.**
  → Les réponses de l'IA ne doivent pas être reprises sans vérification – la vérification des faits est obligatoire.
- **La plausibilité ne remplace pas la vérité**
  → Les LLM formulent souvent de manière convaincante – même si le contenu est inventé.
- **Références et citations erronées**
  → L'IA peut générer des sources ou des études qui semblent réalistes, mais qui n'existent pas.
- **Risque accru dans les applications sensibles**
  → En médecine, en droit ou dans l'éducation, les hallucinations peuvent conduire à des erreurs dangereuses.

43

Limitation No. 6: La connaissance synthétique



44

La connaissance réelle est comme une photo – la connaissance synthétique, comme une peinture.

Les fake news trompent volontairement. La connaissance synthétique trompe involontairement – mais souvent avec plus de crédibilité !

<u>Connaissance réelle</u>

– Repose sur une observation réelle
– Correspond à la réalité (vérifiable)
– Possède une source (primaire) et une date
– Peut être vérifiée

<u>Connaissance synthétique</u>

– Basé sur l'interprétation et la reconstruction.
– Imagine la réalité (de manière créative et plausible).
– A du style et de l'impact, mais n'est pas une source primaire.
– Semble crédible, mais n'est pas mesurable.

Image générée avec Sora

45

Le cercle vicieux de la pollution web générée par l'IA

Saturation d'Internet avec des textes synthétiques générés par l'IA qui supplantent les contenus réels créés par des humains => Model Collapse

Les humains ou l'IA publient des hallucinations, des connaissances approximatives ou des informations erronées.

Par exemple, un article de blog, une contribution sur un forum, un texte généré automatiquement.

Ces contenus sont publiés en ligne

Par exemple, dans les forums, les sites d'actualités, les PDF, Quora, les publications LinkedIn

Lors de la prochaine session de formation, les contenus suivants seront abordés

Le modèle ne peut plus faire la distinction entre « vrai » et « synthétique ». (Epistemic Feedback Loop)

Le LLM répète (ou renforce) ces informations dans les réponses futures

Il en résulte un cercle vicieux de plausibilité sans contrôle de la véracité. (Model Collapse)

AI-Generated Pollution of the Web

46

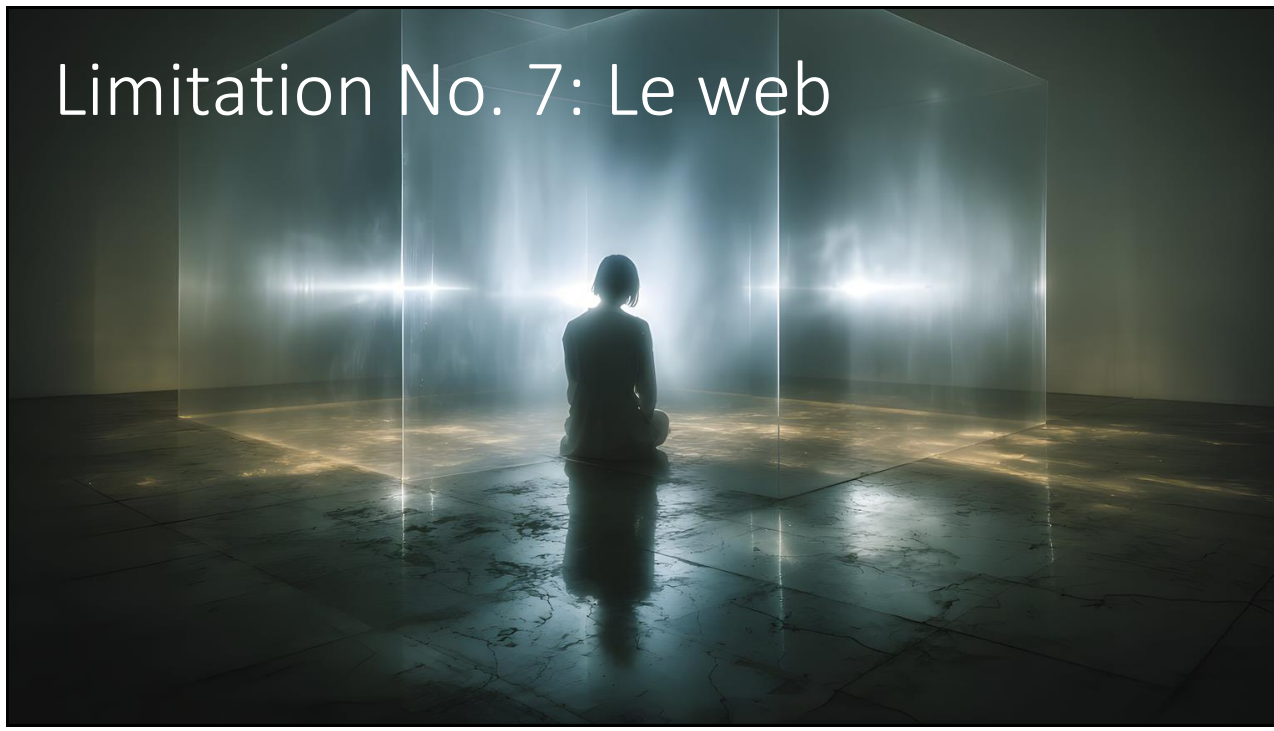
⚠ Connaissances synthétiques : Conséquences sur l'utilisation

- **Les sources d'information deviennent moins fiables**
  → De plus en plus de sites web, de forums, de critiques ou d'articles de blog contiennent du contenu généré par l'IA sans vérification des faits, ce qui entraîne une perte de confiance.
- **Les contenus humains disparaissent ou sont noyés dans la masse**
  → Les textes générés par l'IA supplantent les voix authentiques (par exemple dans les médias sociaux, les évaluations, les articles journalistiques).
- **Appauvrissement autoréférentiel des modèles d'IA (« consanguinité »)**
  → Lorsque les modèles d'IA apprennent de plus en plus à partir de textes générés par l'IA, ils perdent leur diversité, leur créativité et leur fidélité aux faits.
- **Désactiver l'accès au Web**
  → Pour certaines requêtes, il est judicieux de désactiver l'accès à Internet.

© Dr. S. Imboden

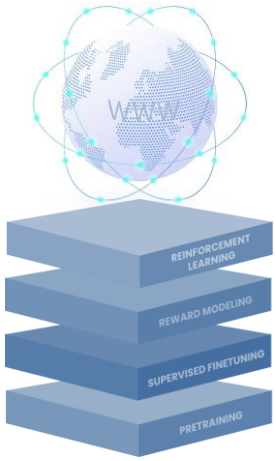
47

Limitation No. 7: Le web



48

Actualité vs fiabilité = Naviguer ou ne pas naviguer ?



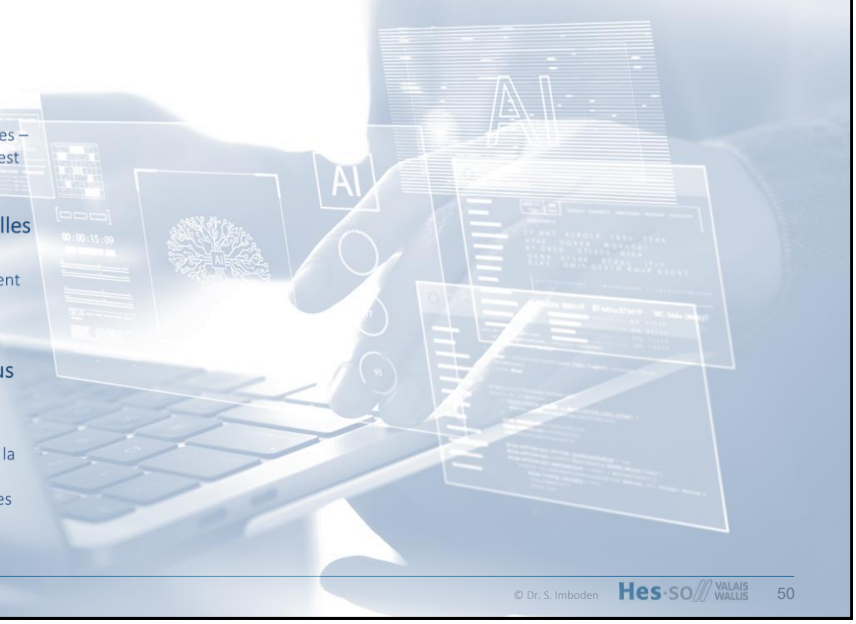
Web : actuel mais pas toujours fiable

Model Checkpoint : fiable mais pas toujours actuel

49

⚠ Accès au web : Conséquences sur l'utilisation

- **Les utilisateurs doivent décider eux-mêmes si l'accès au Web est utile.**
  → Activer pour les questions d'actualité (par exemple, actualités, disponibilité, tendances). Désactiver pour les sujets théoriques ou sensibles – un contrôle conscient de la part de l'utilisateur est nécessaire.
- **On ne sait jamais exactement sur quelles sources repose la réponse.**
  → Lorsqu'on accède à Internet, il manque souvent des indications sur la qualité, la partialité ou la crédibilité de la source.
- **Les connaissances organisées sont plus fiables, mais elles deviennent rapidement obsolètes.**
  → Par exemple, les connaissances de GPT-5 sur la législation, les versions logicielles ou les développements politiques deviennent obsolètes en quelques mois.



50

Limitation No. 8: Les fake news

51

Une nouvelle exclusive de ce matin…

52

⚠ Fake News et Deep Fakes : Conséquences sur l'utilisation

- **Les contenus manipulateurs sont plus faciles à créer, mais plus difficiles à détecter.**
  → Les outils d'IA peuvent générer des images, des vidéos et des textes d'un réalisme trompeur, sans respecter les normes journalistiques ni vérifier les faits.
- **La confiance dans les contenus authentiques est systématiquement sapée.**
  → Si tout peut « paraître réel », tout devient potentiellement discutable, même les contenus véridiques.
- **La désinformation peut être diffusée plus rapidement et de manière plus ciblée.**
  → L'IA permet une diffusion massive de la propagande, par exemple via des robots automatisés sur les réseaux sociaux ou des informations erronées personnalisées.

53

Ne faites confiance à aucune réponse fournie par l'IA que vous n'avez pas vérifiée vous-même !

54

Agenda

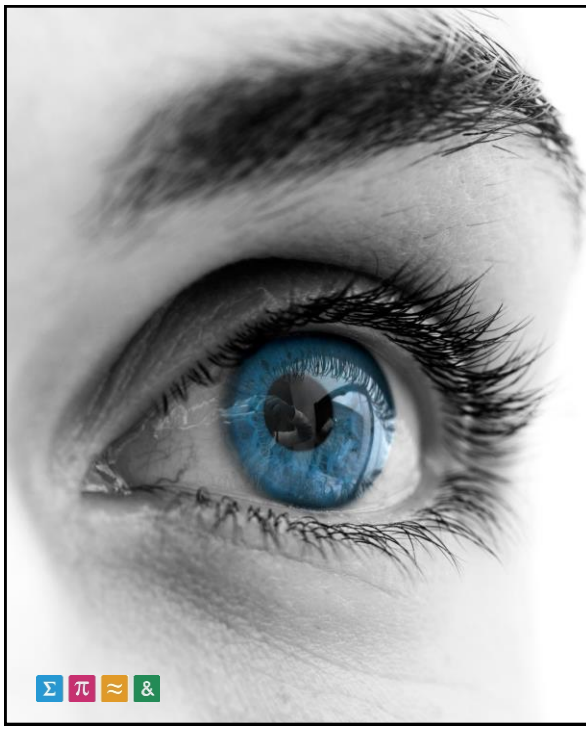
1. L'accélération à l'ère de l'IA
2. Comment fonctionnent l'intelligence artificielle et les LLMs ?
3. Huit limitations et les conséquences sur l'utilisation
4. Conclusion

55

Conclusion

1. **Croissance exponentielle :** l'IA générative progresse au rythme de mois, pas d'années – un potentiel immense : ceux qui tardent à agir risquent d'être dépassés.
2. **Comment fonctionnent les LLM** : ils prédisent le mot le plus probable, sans comprendre ni vérifier – les réponses semblent crédibles, mais ne sont pas garanties exactes.
3. **Déploiement progressif :** commencer par des pilotes et l'automatisation des processus, utiliser ensuite l'IA comme copilote décisionnel, puis rendre les produits plus intelligents pour – à terme – créer de nouveaux modèles d'affaires.
4. **Limites à connaître :** l'IA générative peut halluciner (inventer du contenu plausible mais faux) – ne jamais lui faire confiance aveuglément : toujours vérifier.
5. **Principe TITO & compétence de prompt :** « Trash in – Trash out » : des entrées de qualité et des prompts bien structurés sont indispensables pour obtenir des résultats utiles.

56



Merci de votre attention

Haute école de gestion
Dr. Serge Imboden
Techno-Pôle 3
3960 Sierre
+41 27 606 90 72
+41 79 217 06 08
serge.imboden@hevs.ch
www.2iManagement.ch

57