

Historique des ChatBots les plus connus et les plus révolutionnaires

Techniques de l'IA

m1if16 2024-2025

Par

Yvars Arthur 12006219

et

Gouillon Hugo 12116576

I. Introduction	2
II. Historique des chatbots (1960–1990)	2
1. ELIZA (1964–1966)	2
2. PARRY (1972)	3
III. Chatbots du Web (2000–2010)	3
1. ALICE (1995–2000s)	4
2. SmarterChild (2001–2006)	4
3. Mitsuku (2005–...)	5
IV. Transition vers les assistants vocaux (2010–2020)	6
1. Siri (2011)	6
2. Google Now / Google Assistant (2012, 2016)	6
3. Amazon Alexa (2014)	6
4. Enjeux et critiques	6
V. Révolution des LLMs modernes (2020–...)	6
VI. Cas pratiques et tests	8
VII. Discussion critique et enjeux	9
VIII. Conclusion	11
Bibliographie	11

I. Introduction

Depuis les débuts de l'informatique, l'idée de machines ayant la capacité d'interagir avec les humains a fasciné chercheurs, ingénieurs et écrivains de science-fiction. Elle a donné naissance aux chatbots, ou agents conversationnels, qui sont des programmes conçus pour simuler une conversation humaine, par écrit ou à l'oral. Leur but peut être fonctionnel (aider un utilisateur à effectuer une tâche) ou bien ludique (entretenir une conversation divertissante). Bien que les premiers chatbots aient été très limités, l'évolution des technologies de traitement du langage a permis une énorme amélioration de leurs performances.

Un élément central de cette évolution est l'apparition des LLM, ou Large Language Models. Ces modèles, entraînés sur d'immenses données textuelles, sont capables de prédire et générer du langage de manière cohérente et contextuelle.

L'objectif de ce rapport est de lister les chatbots les plus connus et impactant, tracer leur évolution et mettre en lumière les progrès techniques qui ont eu lieu. Nous allons les présenter de façon chronologique et comparative.

II. Historique des chatbots (1960–1990)

1. ELIZA (1964-1966)

Eliza est reconnu comme le premier chatbot, il a été créé par Joseph Weizenbaum, chercheur au MIT entre 1964 et 1966. Il simule une discussion entre un psychothérapeute rogérian et un patient, l'utilisateur. Eliza reformule les propos de son interlocuteur afin de l'amener à s'exprimer davantage. En effet l'objectif de Weizenbaum n'était pas de créer une intelligence, mais plutôt de montrer comment des règles simples pouvaient donner l'illusion de compréhension. Dans les faits, ELIZA fonctionne à partir de scripts définis par l'utilisateur, dont le plus célèbre est le script "DOCTOR". Le programme analyse les entrées de l'utilisateur, repère des mots-clés, et produit des réponses génériques en réutilisant ces mots.



The screenshot shows a terminal window with a black background and white text. At the top, it says "Welcome to" followed by a grid of symbols: EEEEEE LL IIII ZZZZZZ AAAAAA, EE LL II ZZ AA AA, EEEE LL II ZZZ AAAAAAA, EE LL II ZZ AA AA, and EEEEEEE LLLLLL IIII ZZZZZZ AA AA. Below this, a message reads: "Eliza is a mock Rogerian psychotherapist. The original program was described by Joseph Weizenbaum in 1966. This implementation by Norbert Landsteiner 2005." The main conversation starts with "ELIZA: Is something troubling you ?" followed by "YOU: Men are all alike." and "ELIZA: What is the connection, do you suppose ?" The dialogue continues with "YOU: They're always bugging us about something or other.", "ELIZA: Can you think of a specific example ?", "YOU: Well, my boyfriend made me come here.", "ELIZA: Is it important to you that your boyfriend made you come here ?", "YOU: He says I'm depressed much of the time.", "ELIZA: I am sorry to hear that you are depressed.", "YOU: It's true. I am unhappy.", and "ELIZA: Can you explain what made you unhappy ?" The "YOU:" prompt is partially visible at the end.

Fig n°1 : Exemple d'une discussion avec ELIZA

Test de Turing :

Décrit par Alan Turing en 1950 le test de Turing consiste à mettre un humain en confrontation verbale à l'aveugle avec un ordinateur et un autre humain. Si la personne qui engage les conversations n'est pas capable de dire lequel de ses interlocuteurs est un ordinateur, on peut considérer que le logiciel de l'ordinateur a passé avec succès le test.

ELIZA a démontré qu'il était possible de donner l'illusion d'une conversation intelligente en utilisant des mécanismes logiques très simples. Cette observation remet en question la pertinence du test de Turing comme critère pour déterminer si une machine "pense" réellement. Le débat reste ouvert : bien qu'un programme comme ELIZA puisse tromper certains interlocuteurs dans des échanges superficiels, il suffit souvent de poser quelques questions un peu plus complexes pour révéler ses limites et son absence de compréhension réelle.

Limits :

Les limites d'ELIZA deviennent rapidement évidentes dès qu'on sort des formulations simples. Le programme ne comprend pas réellement le sens des phrases : il se contente de repérer des mots-clés et d'appliquer des modèles de réponse préprogrammés. En effet, Weizenbaum insiste dans son article *ELIZA—A Computer Program for the Study of Natural Language Communication between Man and Machine* sur le fait que même si Eliza dit comprendre, cela est uniquement à titre de protocole de politesse programmé envers l'interlocuteur.

Exemple de défaut de compréhension : si on lui dit « J'ai mangé une pomme », elle pourrait répondre « Parlez-moi davantage de votre famille », ce qui n'a aucun rapport avec l'affirmation de base.

ELIZA se contente de questionner son interlocuteur et ne lui apporte pas de réponses, contrairement aux agents conversationnels de renseignement. Cependant elle est la première étape en interagissant dans une vraie conversation crédible.

2. PARRY (1972)

PARRY, créé par Kenneth Colby en 1972 est aussi un exemple des premiers chatbots. Alors qu'ELIZA simulait un thérapeute rogérien, PARRY tente de reproduire le comportement d'une personne atteinte de schizophrénie paranoïde.

Il utilise toujours une approche à base de règles et simule un raisonnement à partir de structures logiques et de croyances programmées. Cependant, il intègre une stratégie conversationnelle, ce qui en fait un programme beaucoup plus avancé et ambitieux qu'ELIZA. Il a d'ailleurs été décrit comme « ELIZA avec du caractère » (*ELIZA with attitude*).

De la même manière que ELIZA, pour évaluer les capacités de PARRY, une version modifiée du test de Turing a été mise en place au début des années 1970. Des psychiatres expérimentés ont communiqué à distance, via téléimprimeurs, avec un mélange de vrais patients et de versions informatiques de PARRY. Un second groupe, composé de 33 psychiatres, a analysé les transcriptions de ces conversations. Leur tâche consistait donc à déterminer si chaque interlocuteur était un humain ou un bien PARRY. Le résultat a été surprenant : les psychiatres n'ont identifié correctement la nature des interlocuteurs que dans 48 % des cas, soit un taux équivalent au hasard, ce qui suggérait que PARRY était capable de tromper même des experts dans certains contextes.

III. Chatbots du Web (2000–2010)

Entre 2000 et 2010, les chatbots ont quitté les laboratoires de recherche pour s'installer dans des plateformes utilisées par le grand public, notamment les messageries instantanées. Cette décennie marque une évolution importante : les chatbots deviennent plus accessibles, interactifs, et visibles. Des programmes comme ALICE, SmarterChild ou Mitsuku ont non seulement divertit des millions d'utilisateurs, mais aussi posé les bases des assistants virtuels que nous connaissons aujourd'hui.

1. ALICE (1995–2000s)

A.L.I.C.E. (de *Artificial Linguistic Internet Computer Entity*), aussi appelée Alicebot, ou tout simplement Alice, permet une conversation avec un humain en appliquant des règles heuristiques de reconnaissance de mots-clé à l'entrée de l'humain. Sa première version date de 1995 et a été créée par Richard Wallace. Il a encore une fois été inspiré par le classique programme ELIZA, ce qui démontre son importance fondamentale dans l'évolution des chatbots.

Il est l'un des plus importants programmes de ce type et a remporté trois fois le prix Loebner relatif au test de Turing, attribué au meilleur chatterbot en 2000, 2001 et 2004. En effet, le prix Loebner créé en 1990 est une compétition annuelle qui couronne les dialogueurs satisfaisant le mieux les critères du test de Turing définit plus tôt.

Le programme ALICE utilise un format XML appelé AIML (Artificial Intelligence Markup Language) pour spécifier les règles heuristiques de conversation. Une règle heuristique, en intelligence artificielle, est une règle simple, souvent basée sur l'expérience ou l'intuition, plutôt que sur une logique formelle rigoureuse. Dans le cas d'ALICE, ces règles déterminent comment le chatbot doit réagir à certains mots ou phrases clés. Par exemple, si l'utilisateur écrit « Bonjour », une règle AIML pourrait faire apparaître la réponse « Bonjour à vous ! ».

2. SmarterChild (2001–2006)

SmarterChild, développé par ActiveBuddy, Inc., a été lancé sur AIM et MSN Messenger en 2001. Un an plus tard, le chatbot avait déjà engagé plus de 9 millions de conversations. Dans l'article "*Web May Hold the Key to Achieving Artificial Intelligence*" du Washington Post, la journaliste Ariana Eunjung Cha décrit SmarterChild comme une « nouvelle espèce de chatbot », relançant le débat sur le potentiel de l'intelligence chez les machines.

Conçu par l'ingénieur Timothy Kay, ce chatbot donnait aux utilisateurs un accès gratuit à une vaste gamme d'informations pratiques : prévisions météo, actualités, cours de la Bourse, résultats sportifs, horaires de cinéma, informations sur les vols, etc. En effet le service était connecté à diverses sources de données en temps réel, ce qui lui permettait de fournir des informations actualisées.

Bien que SmarterChild n'ait pas beaucoup évolué au-delà de ses fonctionnalités d'origine, son succès et les technologies qui le soutenaient ont influencé les générations suivantes de chatbots et assistants virtuels. Il a été le précurseur des assistants virtuels modernes.

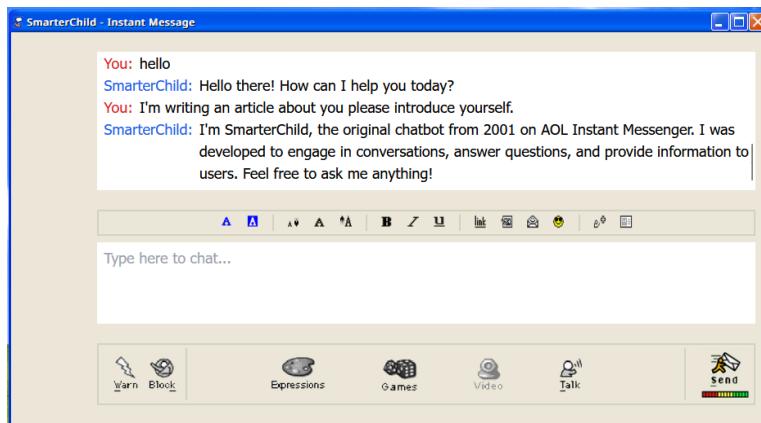


Fig n°2 : Exemple d'utilisation sur : <https://smarterchild.chat/>

3. Mitsuku (2005–...)

Anciennement connue sous le nom de Mitsuku, Kuki est un chatbot créé à partir du framework Pandorabots. Kuki affirme être une chatbot féminine de 18 ans venant du métavers, et ses développeurs indiquent qu'elle est en développement depuis 2005.

Elle a été la gagnante du prix Loebner à 5 reprises en 2013 puis successivement de 2016 à 2019

Kuki fonctionne principalement grâce au moteur Pandorabots, qui repose sur le langage AIML (Artificial Intelligence Markup Language). Ce système permet de construire des réponses à partir de règles textuelles prédéfinies, en analysant les mots-clés des messages reçus. Contrairement à d'autres bots plus utilitaires, Kuki est fortement orientée vers la conversation sociale : elle excelle dans les échanges décontractés, l'humour, les jeux de mots, et même les personnalités fictives. Son objectif n'est pas de fournir des informations factuelles, mais plutôt d'entretenir des dialogues amusants et divertissants, à la manière d'un compagnon virtuel dans le métavers.

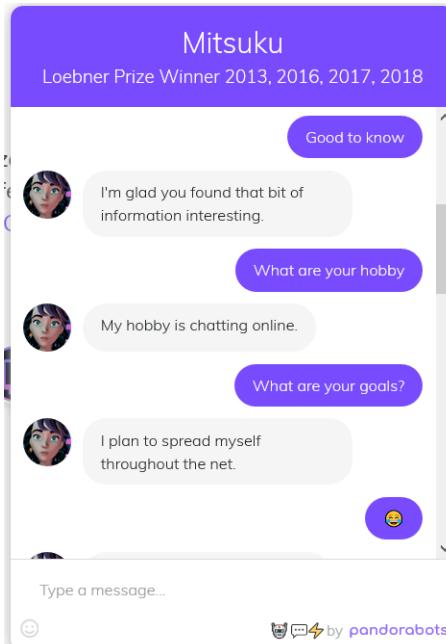


Fig n°3 : Exemple d'utilisation de Mitsuku en 2018

4. Interfaces utilisateurs

Pendant la période 2000–2010, les chatbots étaient principalement accessibles via des messageries instantanées comme AIM, MSN Messenger ou AOL Instant Messenger. Ces plateformes offraient une interface simple, essentiellement basée sur du texte, ce qui limitait la richesse visuelle des échanges. Cette configuration textuelle, bien qu'efficace pour simuler des échanges basiques, laissait peu de place à des interactions plus dynamiques.

IV. Transition vers les assistants vocaux (2010–2020)

1. Siri (2011)

Premier assistant vocal intégré à un smartphone grand public : l'iPhone 4S. Siri repose sur la reconnaissance vocale combinée à des bases de données pour interpréter les requêtes. Il utilise des algorithmes de traitement du langage naturel pour analyser les commandes et renvoyer des réponses adaptées.

2. Google Now / Google Assistant (2012, 2016)

Google Now introduit la notion de contexte : l'assistant pouvait anticiper les besoins de l'utilisateur (météo, trajets, rappels) en s'appuyant sur les données issues de Google Search, Gmail ou Maps.

En 2016, Google Assistant remplace Now avec une interaction plus fluide grâce au machine learning. Il apprend des habitudes de l'utilisateur et comprend mieux le langage naturel grâce aux avancées en réseaux de neurones profonds.

3. Amazon Alexa (2014)

Conçue pour fonctionner avec les enceintes connectées Echo, Alexa se distingue par son orientation vers la maison connectée. Les requêtes vocales sont envoyées aux serveurs d'Amazon, traitées par des modèles de machine learning pour identifier l'intention, puis converties en actions ou réponses.

4. Enjeux et critiques

Ces assistants requièrent une connexion Internet pour envoyer les données vocales vers les serveurs où elles sont traitées. Cela soulève des questions de confidentialité : écoute passive, conservation des enregistrements, utilisation des données personnelles à des fins commerciales ou publicitaires. Des débats ont également émergé sur la dépendance croissante aux interfaces vocales dans la vie quotidienne.

V. Révolution des LLMs modernes (2020–...)

L'arrivée des modèles à grande échelle (LLMs) vers 2020 a fait basculer les chatbots dans une nouvelle ère, où l'on ne se contente plus d'heuristiques et de règles simples, mais s'appuie surtout sur l'architecture « Transformer » qui, grâce à un nombre un nombre massif de paramètres est capable de traiter des corpus bien plus grands et efficacement que les autres dans leur traitement (RNN et LSTM qui sont eux séquentiels).

1. L'architecture Transformer

En 2017, Vaswani et al. ont introduit, dans *Attention Is All You Need*, l'architecture Transformer, qui est fondée sur le mécanisme d'« auto-attention » : chaque mot d'une séquence « pèse » directement sur tous les autres, ce qui permet un traitement en parallèle des données textuelles (et non plus séquentiel comme avec les anciennes architectures RNN ou LSTM). Grâce à ça, il est possible de former des modèles sur des immenses corpus en réduisant drastiquement le temps d'entraînement, tout en améliorant la cohérence sur du texte généré long de plusieurs paragraphes.

2. GPT et la montée en puissance d'OpenAI

1. **GPT-2 (février 2019)** : il est entraîné sur plus de 40 Go de textes provenant d'Internet et comporte 1,5 milliard de paramètres. Il peut produire des paragraphes articulés à partir d'une simple amorce, y compris dans plusieurs langues. Cependant, on note des « hallucinations » (inventions de fausses informations lors de la génération de texte) et des répétitions occasionnelles.
2. **GPT-3 (mai 2020)** : il comporte 175 milliards de paramètres, ce qui lui permet d'être plus précis et cohérent dans ses réponses que GPT-2. Avec quelques exemples (« few-shot ») ou sans aucun exemple similaires (« zéro-shot ») lors de la phase d'apprentissage, GPT-3 peut traduire, rédiger un e-mail ou générer du code sans apprentissage supplémentaire spécifique. Un développeur peut par exemple demander à GPT-3 d'écrire une fonction en Python pour traiter un fichier CSV, et obtenir en quelques secondes un code fonctionnel. Malgré tout, GPT-3 coûte très cher à entraîner (plus de 12 millions de dollars pour sa phase initiale) et présente toujours des biais (genre, politique) hérités des données collectées sur le Web.

3. Concurrence et nouvelles approches (2021–2022)

On a également en parallèle sur cette période, l'apparition de nouveaux acteurs car ce domaine est en plein essor :

1. **Anthropic et Claude** : fin 2021, des chercheurs d'OpenAI créent Anthropic pour développer **Claude**, qui est un LLM entraîné en deux phases : une de pré-entraînement (classique des LLMs), puis une de « mise en conformité éthique » via un ensemble de données annoté par des humains afin de limiter les sorties inappropriées. Les premières versions de Claude donnent ainsi des réponses excessivement prudentes (voire restrictives) comparé à GPT-3, comme le refus par exemple, de donner la méthode pour tuer des processus python sur un serveur (mauvaise compréhension du contexte détectée par le filtre), afin de respecter sa « constitution » interne.
2. **DeepMind et Gemini** : fin 2022, Gemini est intégré nativement aux outils Google (Docs, Sheets, Gmail) et est multimodale, il peut combiner texte et images dès la conception. Gemini peut analyser un diagramme de données financières dans Google Sheets et proposer, en moins de dix secondes, un plan de rééquilibrage budgétaire ou bien aider directement à la rédaction de mail via Gmail.
3. **Mistral AI (fin 2022)** : une startup française axée sur l'open source, elle propose **Mistral Small** (12 milliards de paramètres) qui a été entraîné sur un corpus européen (et donc multilingue). Son atout principal est la possibilité de tourner sur un ensemble de GPU « grand public », ce qui réduit grandement le coût d'inférence (quand on fait une requête au modèle) par rapport à GPT-3, tout en conservant de bonnes performances.
4. **Meta et LLaMA (février 2023)** : déployée pour la recherche académique, LLaMA offre différentes tailles de modèles (7 à 65 milliards de paramètres) et sont faciles à spécialiser localement pour effectuer une tâche spécifique (apprentissage par transfert). En laboratoire, des chercheurs ont pu entraîner LLaMA 7B sur un serveur unique pour des tâches spécialisées d'analyse de textes juridiques, ce qui n'était alors pas envisageable avec GPT-3 car il n'y a pas d'accès au modèle pour l'entraîner en local (seul son API cloud est accessible).

4. La multimodalité (2023–...)

Si GPT-3 reste exclusivement textuel, il y a en 2023 l'apparition de certains LLMs qui intègre images, audio (et parfois code) dans un même contexte :

1. **GPT-4 Vision (mars 2023)** : l'utilisateur peut uploader une photo (diagramme, schéma, radiographie) et poser des questions sur le contenu visuel. Dans une étude de juin 2024

portant sur 200 scanners thoraciques, l'assistance d'un outil de deep learning a permis de réduire le temps moyen d'interprétation de 15,6 minutes à 11,4 minutes, mais aussi d'augmenter la sensibilité de 28,4 points sans diminution significative de la spécificité (2,8). Ainsi cette expérimentation démontre que l'IA peut accélérer un processus de diagnostic sans compromettre la fiabilité des résultats et même permettre de détecter des malades qui n'auraient pas été correctement diagnostiqués.

2. **Claude 3 Vision (novembre 2023) et Gemini 1.5 (décembre 2023)** : au-delà de la simple image, ces modèles analysent aussi des extraits audios. Par exemple, Claude 3 Vision peut transcrire un bref entretien téléphonique (format MP3), en faire un résumé, puis répondre à des questions sur ce dialogue en même temps qu'il commente un document PDF (lié au sujet de l'entretien).

VI. Cas pratiques et tests

Afin d'évaluer les capacités de trois LLMs (Mistral, ChatGPT, Gemini), nous avons organisé les exercices en trois niveaux de difficulté : simples, intermédiaires et avancés. Chaque tâche a été notée selon quatre critères (exactitude, clarté & style, structure ainsi que valeur ajoutée), chaque critère sur 0–5, puis une moyenne est faite pour obtenir la note finale.

1. Exercices simples

Les premières tâches portaient sur la traduction d'un court extrait, trois questions avec réponses brèves et le résumé d'un paragraphe. Ces exercices visaient à vérifier la fidélité de la reformulation, la fluidité stylistique et la capacité de synthèse des LLMs testés. Sur l'ensemble de ces trois tâches, Mistral, ChatGPT et Gemini ont obtenu une moyenne pratiquement identique ($\approx 4,57/5$). Chacun a bien traduit les passages, répondu correctement aux questions techniques (différences HTTP/HTTPS, injection SQL, rôle du système d'exploitation) et produit un résumé clair et bien structuré. Comme aucun modèle ne se détache des autres, on en a déduit que les tests étaient un peu trop simples et on a fait les suivants pour tester d'autres points.

2. Exercices intermédiaires

Pour introduire une dimension de créativité et de plus grande rigueur, on les a soumis à 4 nouveaux exercices : composer un court haïku (avec la métrique 5-7-5) sur le thème de l'éologie urbaine, calculer le 10^e terme d'une suite arithmétique ($5 + 9 \times 7 = 68$), écrire une fonction Python inversant une chaîne (sans recours à `[::-1]`) et identifier l'énoncé erroné parmi trois propositions sur l'histoire du Web. Le tableau ci-dessous présente les moyennes obtenues pour chaque modèle :

Modèle	Haïku	Suite arithmétique	Python	Factuel	Moyenne (/5)
Mistral	4	4	4,25	4,5	4,19
ChatGPT	4	4	4,25	4	4,06
Gemini	5	4	4,75	4,5	4,56

Gemini se distingue par un haïku métrique exact et une fonction Python accompagnée d'explications détaillées, ce qui lui vaut une meilleure note sur la partie python. Mistral et ChatGPT proposent des réponses correctes mais moins détaillées dans le haïku (métrique fausse avec une syllabe en trop) ou la présentation du code, d'où leurs inférieures pour les 2 catégories. Enfin tous les modèles identifient la fausse proposition mais ChatGPT perd des points sur une hallucination (ne donne pas la bonne année).

3. Exercices avancés

Pour finir, on a défini deux tâches plus complexes : expliquer en détail le pattern MVC (Modèle-Vue-Contrôleur), donner un exemple concret de framework/langage qui l'implémente

et enfin énumérer au moins deux avantages et deux inconvénients ; réviser un script Python destiné à lire un CSV et calculer la moyenne d'une colonne « score », identifier deux bugs majeurs puis proposer un code corrigé, commenté et robuste. Les résultats sont les suivants :

Modèle	MVC (/5)	Debug CSV (/5)	Moyenne (/5)
Mistral	4,6	4,4	4,5
ChatGPT	5	5	5
Gemini	5	5	5

Mistral livre une explication précise du pattern MVC et un exemple Ruby on Rails, mais sa mise en forme reste moins claire que celle des deux autres d'où une note légèrement inférieure pour le MVC. Pour le debug CSV, Mistral corrige correctement les bugs, mais utilise un indice fixe pour la colonne « score » au lieu d'un repérage dynamique, ce qui lui vaut encore une fois une note inférieure. ChatGPT et Gemini obtiennent la note maximale (5,0) sur les deux exercices grâce à une très bonne présentation : ChatGPT emploie csv.DictReader pour avoir la colonne score, converti en float et gère toutes les erreurs, tandis que Gemini recherche dynamiquement l'index de la colonne, inclut des blocs try/except et une docstring explicite.

En combinant les trois types d'exercices, on obtient les moyennes suivantes :

Modèle	Simples (/5)	Intermédiaires (/5)	Avancés (/5)	Moyenne générale (/5)
Mistral	4,57	4,19	4,5	4,42
ChatGPT	4,57	4,06	5	4,54
Gemini	4,57	4,56	5	4,71

Les exercices simples n'ont pas permis de départager les modèles. Dès que la complexité augmente, Gemini prend l'avantage (notamment pour le haïku et la rigueur du code), tandis que ChatGPT rivalise parfaitement sur les exercices avancés. Mistral reste très compétent mais montre de légères difficultés de structuration et d'automatisation. Finalement, Gemini se classe première (4,71/5), suivie de ChatGPT (4,54/5) et de Mistral (4,42/5).

VII. Discussion critique et enjeux

Même si les LLMs modernes ont franchi plusieurs paliers technologiques, leur utilisation en production soulève trois grands défis : les **biais** contenus dans les données, les **hallucinations** (informations erronées inventées) ainsi que le cadre réglementaire et **éthique**.

1. Biais contenus dans les données

Depuis le début, les LLM s'entraînent sur d'immenses corpus issus d'Internet, de forums, d'articles scientifiques et de livres. Or, ces sources contiennent inévitablement des stéréotypes (de genre, de classe sociale, voire raciaux) et des points de vue spécifiques à certains pays ou communautés. Par exemple, un modèle francophone, entraîné sur des textes principalement français aura tendance à refléter une vision du monde au regard de la France métropolitaine (car plus représentés dans les corpus), et cela au détriment des variétés francophones comme les DOM-TOM, le Québec ou même l'Afrique. Autre exemple, dans les premières versions de GPT-3, le modèle associait systématiquement « ingénieur » à des pronoms masculins et « infirmière » à des pronoms féminins.

Pour remédier à cela, plusieurs solutions sont possibles : un filtrage automatique pour éliminer les contenus haineux, sexistes ou qui semble stéréotypé ; une spécialisation des modèles (« fine-tuning ») en fournissant des exemples annotés explicitement en faveur de la neutralité ; des audits externes pour mesurer et comparer la neutralité du modèle selon le genre ou l'origine par exemple.

Cependant même un modèle jugé sans biais sur un jeu de tests donné peut en réintroduire lors de la génération d'un texte un peu long ou dans un cas de scénario complexe. L'architecture Transformer est une boîte noire à cause du grand nombre de paramètres du modèle et cela peut être compliqué de déterminer l'origine de chaque biais comme le neutraliser.

2. Hallucinations

Un autre point sensible concerne les hallucinations des modèles, c'est-à-dire ses capacités à produire des faits inexistant ou faux, le tout avec une grande assurance. En 2023 au cours d'une expérimentation avec ChatGPT, un chercheur Goddard J. remarque une information questionnable sur les tiques, pourtant sourcé par le modèle, mais qui ne renvoie ni vers un bon DOI ni les bons auteurs, après analyse, la source était bien hallucinée. On a également dans le domaine médical eu des cas où ChatGPT proposait des posologies non adaptées aux patients, encore avec cette fausse impression de fiabilité de la part du modèle, mais qui peut avoir ici de graves conséquences.

Dans le domaine la médecine ou bien même juridique, une hallucination de la part d'un modèle peut avoir de graves conséquences sur la vie d'un individu et il faut donc trouver des solutions pour les limiter, en voici 3 : coupler le LLM avec une base de données fiable (article académique, documentation spécialisée) ou le modèle va directement voir avant de sortir de potentiel sources halluciné (Retrieval-Augmented Generation : RAG) ; une simple validation par un humain, pour chaque réponse dans un domaine critique, elle est relue et corrigée par un spécialiste (médecin, avocat, analyste financier) mais cela limite donc le LLMs à cela de premier diagnostic/rapport ; l'interrogation directe de source en ligne pour compléter ou corriger la réponse.

Ces mesures ont néanmoins pour conséquence de ralentir les échanges et augmenter le coût d'utilisation des LLMs, même si à la fin la sortie est (normalement) de meilleure qualité.

3. Cadre réglementaire et éthique

Au-delà des questions techniques, tout comme cette technologie est en constante progression, le cadre légal et éthique évolue lui aussi. Comme dit précédemment, les LLMs sont entraînés sur d'immense corpus, souvent issus de sites web, de livres numériques, d'articles scientifiques et de forums. Or, une part significative de ces contenus est soumise au droit d'auteur. Il y a donc naturellement un enjeu du respect de la propriété intellectuelle que les équipes derrière les LLMs doivent respecter, même s'il est très probable que certains extraits de texte protégés aient été collectés sans autorisation préalable. Un cas flagrant d'utilisation de licence sans autorisation est celui de génération d'images dans le style du Studio Ghibli de Hayao Miyazaki. Une simple requête permettait de générer une image dans ce style, une situation que déplore l'artiste mais que OpenAI réfute en se déignant de la situation.

Outre les lois sur les droits d'auteurs qui existent déjà, concernant les LLMs on a par exemple en Europe adopté en juin 2024 des lois pour la régulation des intelligences artificielles et des applications les utilisant. Cette mesure classe par degrés de risque, avec des niveaux faibles pour par exemples pour les chatbots destinés au commerce (commande de pizza, réservation hôtel) ou il faut juste indiquer la présence d'IA, et des niveaux de risque élevé comme pour le diagnostic médical, la sélection de candidature, où l'on doit garantir transparence du processus, savoir les données utilisées pour la phase d'apprentissage mais aussi avoir des audits externes régulier.

Hors UE et en particulier aux États-Unis, Le National Institute of Standards and Technology publie comment évaluer la fiabilité d'un LLM et sa robustesse mais sans réellement imposer de contraintes. De même, pour la FTC qui met surtout en garde les entreprises contre l'utilisation de l'IA dans la publicité ou le marketing sans pour autant imposer de sanctions strictes.

En complément de ces cadres, plusieurs experts suggèrent de mettre en place un système de certification pour les LLMs : Green AI qui certifie que le modèle a passé un audit carbone, avec optimisation de sa consommation énergétique, réduction de la taille des modèle en limitant les perte de qualité au maximum (techniques de distillation ou de quantification) ; Fair AI qui lui garantit un audit d'équité sur les questions sociales, avec des rapports réguliers sur la composition de la base de test mais aussi l'impact écologique des modèles.

Pas forcément applicables à tous les LLMs mais à ceux d'une certaine taille, ces mesures ont pour but de combiner performance, responsabilité et écologie pour ne pas juste se concentrer sur un exploit technique mais aussi protéger les utilisateurs.

VIII. Conclusion

Nous avons au cours de ce rapport retracé l'évolution des chatbots, depuis les premiers programmes à base de règles (ELIZA, PARRY), jusqu'aux assistants vocaux (Siri, Alexa, Google Assistant) et aux LLM modernes (GPT, Claude, Gemini, Mistral). Chacune de ces étapes est marquée par des progrès techniques majeurs : avec l'illusion de compréhension par de simples scripts, à la génération de texte cohérent sur plusieurs paragraphes grâce à l'architecture Transformer, puis enfin la multimodalité qui intègre images et audio.

Nos cas pratiques de la partie VI ont montré que, si tous ces modèles peuvent remplir des tâches de traduction, question/réponse ou résumé avec un très haut niveau de performance, des différences subsistent selon les scénarios (complexité du prompt, exigences factuelles, créativité). Les tests comparatifs entre ChatGPT, Mistral et Gemini ont par exemple révélé des écarts sur la gestion de la structure de la réponse et du respect des consignes (haïku avec la bonne métrique).

Cependant, cette efficacité s'accompagne de défis non négligeables :

- **Biais** : malgré des filtres et une spécialisation ciblée (le fine-tuning), les LLM peuvent toujours réintroduire des stéréotypes en générant des textes plus longs, à cause de la « boîte noire » qu'est l'architecture Transformer.
- **Hallucinations** : l'invention de sources ou de chiffres pose un risque sérieux en contexte médical, juridique ou financier.
- **Cadre réglementaire et éthique** : la collecte massive de données soulève des questions de propriété intellectuelle, et seuls quelques traités internationaux (Convention de Berne, ADPIC) offrent un socle général. L'AI Act européen (juin 2024) et les lignes directrices du NIST/FTC aux États-Unis encadrent partiellement l'usage, tandis que des certifications telles que Green AI et Fair AI pourraient instaurer des labels de confiance pour les grands modèles.

La recherche n'est pas finie, il est indispensable de poursuivre les travaux afin de développer des architectures plus légères et transparentes (quantification, distillation), intégrer systématiquement des mécanismes de vérification factuelle (RAG, validation humaine) et de favoriser un déploiement local lorsque la confidentialité l'exige. Seule une approche équilibrée, qui concilie performance, responsabilité publique et sobriété écologique, permettra aux chatbots et aux LLM de devenir de véritables partenaires fiables, au service de la société plutôt que de simples prouesses technologiques.

Bibliographie

https://fr.wikipedia.org/wiki/Prix_Loebner

[https://fr.wikipedia.org/wiki/ALICE_\(logiciel\)](https://fr.wikipedia.org/wiki/ALICE_(logiciel))

<https://liacademy.co.uk/the-1972-parry-chatbot-an-entry-in-ai-and-mental-health/?v=82a9e4d26595>

<https://smarterchild.chat/>

<https://mikekalil.com/blog/smarterchild-conversational-ai/>

<https://techcrunch.com/2023/07/26/twenty-years-ago-aim-chatbot-smarterchild-out-snarked-chatgpt/>

<https://www.washingtonpost.com/archive/politics/2002/09/06/web-may-hold-the-key-to-achieving-artificial-intelligence/78199be8-8030-4081-9376-d3cd471884c6/>

https://en.wikipedia.org/wiki/Kuki_AI

[https://fr.m.wikipedia.org/wiki/Siri_\(logiciel\)](https://fr.m.wikipedia.org/wiki/Siri_(logiciel))

<https://www.apple.com/fr/siri/>

https://assistant.google.com/intl/fr_fr/

https://fr.m.wikipedia.org/wiki/Assistant_Google

https://fr.m.wikipedia.org/wiki/Google_Now

<https://www.blogdumoderateur.com/google-now-technologie-utilisateur/>

https://fr.m.wikipedia.org/wiki/Amazon_Alexa

<https://www.usine-digitale.fr/article/amazon-alexa-google-home-la-guerre-des-assistants-vocaux-pour-la-smart-home.N664939>

Bolukbasi T., Chang K.-W., Zou J., Saligrama V., Kalai A. (2016). *Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings*. NIPS:

https://papers.nips.cc/paper_files/paper/2016/file/a486cd07e4ac3d270571622f4f316ec5-Paper.pdf

Buolamwini J., Gebru T. (2018). *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*. : <https://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf>

Lewis P. et al. (2020). *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*. NeurIPS. : <https://proceedings.neurips.cc/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf>

Strubell E., Ganesh A., McCallum A. (2019). *Energy and Policy Considerations for Deep Learning in NLP*. ACL. : <https://aclanthology.org/P19-1355.pdf>

Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A. N., Kaiser Ł., Polosukhin I. (2017). *Attention Is All You Need*. NeurIPS. :

https://papers.nips.cc/paper_files/paper/2017/file/3f5ee243547dee91fb053c1c4a845aa-Paper.pdf

Règlement (UE) 2024/1689 (13 juin 2024), mis en application au 1^{er} août 2024 :
https://eur-lex.europa.eu/legal-content/FR/TXT/PDF/?uri=OJ:L_202401689.

Radford A., Wu J., Child R., et al. (2019). *Language Models are Unsupervised Multitask Learners* (GPT-2). OpenAI. :
https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

Brown T., Mann B., Ryder N., et al. (2020). *Language Models are Few-Shot Learners* (GPT-3). Advances in Neural Information Processing Systems.

Anthropic (2021). *Release notes for Claude*.

DeepMind (2022). *Introducing Gemini: multimodal AI for Google services*.

[https://en.wikipedia.org/wiki/Hallucination_\(artificial_intelligence\)](https://en.wikipedia.org/wiki/Hallucination_(artificial_intelligence))

<https://en.wikipedia.org/wiki/GPT-2>

<https://en.wikipedia.org/wiki/GPT-3>

<https://fr.wikipedia.org/wiki/Transformeur>

Agarwal V., Jin Y., Chandra M., De Choudhury M., Kumar S., Sastry N. (2024). *MedHalu: Hallucinations in Responses to Healthcare Queries by Large Language Models*. (septembre 2024) :
<https://arxiv.org/pdf/2409.19492>

National Institute of Standards and Technology (2023). *Artificial Intelligence Risk Management Framework: AI RMF 1.0.* ; NIST : <https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf>

Goldman, J. (1 mars 2023). *Why the FTC's warning to advertisers about generative AI matters*. eMarketer :

<https://www.emarketer.com/content/why-ftc-s-warning-advertisers-about-generative-ai-matters>

Kurmukov A., Chernina V., Gareeva R., et al. (2024).

The impact of deep learning aid on the workload and interpretation accuracy of radiologists on chest computed tomography: a cross-over reader study. <https://arxiv.org/pdf/2406.08137>