Centres for European Policy Network
FREIBURG | BERLIN | PARIS | ROMA

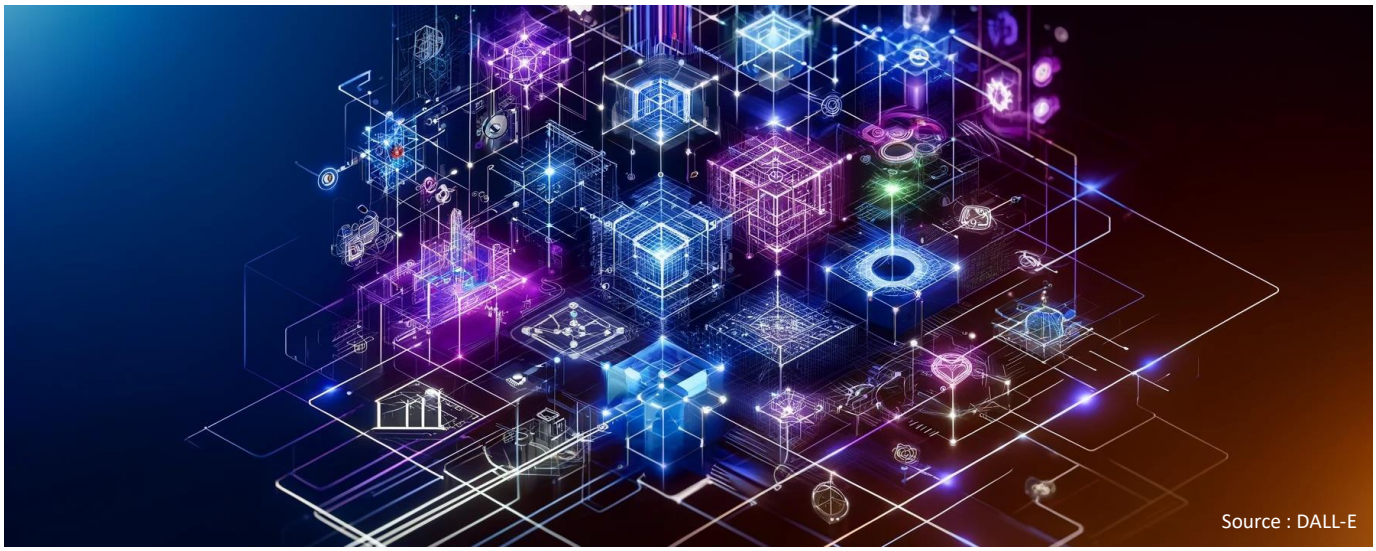
Input du cep

No 7 | 2024 16 avril 2024

Les technologies de modèles de langage, un avantage concurrentiel pour l'UE

Dix points pour un emploi optimal de l'IA générative dans les PME

Anselm Küsters



Source : DALL-E

Les modèles de langage tels que ChatGPT représentent un défi majeur, mais aussi une opportunité pour l'Europe. Plutôt que d'y répondre par le protectionnisme, les projets phares ou l'aversion au risque, il est nécessaire d'adopter une approche pragmatique de l'utilisation à grande échelle des technologies de modèles de langage de l'IA dans l'économie pour maintenir la compétitivité et exploiter le potentiel d'innovation. Cet input du cep décrit dix facteurs que les petites et moyennes entreprises devraient prendre en compte lors de la mise en œuvre afin d'exploiter les avantages concurrentiels existants.

- Les PME devraient comprendre conceptuellement comment l'IA peut améliorer leurs processus grâce à une analyse des besoins et à une planification stratégique. Le choix de services basés sur le cloud ainsi que l'utilisation de modèles ouverts influencent la capacité d'adaptation des outils d'IA et la dépendance ultérieure vis-à-vis d'entreprises externes.
- Grâce au « fine tuning » et à l'utilisation de la « génération augmentée de récupération » (RAG), les PME peuvent spécialiser leurs applications. Le risque d'erreur des modèles d'IA doit être aligné sur la tolérance d'erreur interne. En outre, il convient de développer des compétences internes dans le domaine du « prompt design » et du « online design ».
- Pour une utilisation durable et socialement responsable, il faut tenir compte du cadre légal, effectuer des tests internes en continu et mesurer l'efficacité énergétique. Les mécanismes de feedback doivent être utilisés pour maintenir la technologie à jour.

Table des matières

1 Introduction : employer les technologies linguistiques au lieu de les reproduire

Un grand modèle de langage est utilisé pour chaque requête adressée à ChatGPT. Face à l'évolution fulgurante de ce domaine de l'intelligence artificielle (IA) en pleine croissance, l'Europe est confrontée à un défi de taille, qui a été relevé jusqu'à présent en mettant trop l'accent sur des projets phares[1] et pas assez sur une application à grande échelle dans l'économie. Actuellement, la politique européenne et allemande vise surtout à s'assurer à long terme la plus grande part possible de la chaîne de création de valeur de l'IA afin d'éviter toute dépendance stratégique ultérieure. Cela comprend la construction d'espaces de données[2], le subventionnement massif d'usines de puces[3] et, plus récemment, l'exploitation de superordinateurs spécialisés dans l'IA, la construction de ce que l'on appelle les « *AI Factories* »[4] et une « alliance européenne pour les technologies linguistiques » afin de construire ses propres grands modèles linguistiques[5]. L'espoir à peine dissimulé de créer leurs propres champions nationaux a guidé l'Allemagne et la France dans les négociations finales de la loi européenne sur l'IA.[6] Dans l'ensemble, ces initiatives de construction de technologies linguistiques propres doivent être comprises comme faisant partie d'une tendance plus large de « *home-shoring* »[7] qui prend de plus en plus d'importance dans le contexte des tensions géopolitiques de l'Ukraine à Taiwan.

Même si une telle réflexion stratégique - longtemps négligée au niveau de l'UE - est utile à long terme, la dynamique de l'IA ne permet pas de différer davantage son application concrète. Le développement exponentiel de la technologie, qui s'est jusqu'à présent concentré sur les États-Unis[8], oblige à trouver des « solutions de second ordre » qui diffèrent de l'approche politique adoptée jusqu'à présent en Europe. Au lieu de mettre en place exclusivement des plans de développement d'infrastructures et de modèles propres, accompagnés de procédures d'adjudication publique lentes et d'une réglementation européenne détaillée, il faudrait maintenant parler davantage de la mise en œuvre concrète de cette technologie. En raison de la situation économique difficile et de la lutte pour la compétitivité mondiale, les petites et moyennes entreprises (PME) européennes en particulier ne peuvent pas attendre que les fournisseurs nationaux développent des modèles compétitifs. Les modèles commerciaux américains et les modèles libres à source ouverte sont d'une valeur inestimable pour l'application rapide et généralisée des technologies linguistiques de l'IA.

Bien que l'IA générative ait le potentiel de créer une valeur ajoutée de 2,6 à 4,4 billions de dollars dans tous les secteurs grâce à ses nombreux cas d'application[9], l'intégration des modèles existants dans la pratique entrepreneuriale a jusqu'à présent fait défaut en Allemagne[10]. Bien que l'on estime que des

---

1 Voir à propos de cette critique des « phares » : Friesike et Sprondel (2022), Träg Transformation. Welche Denkfehler den digitalen Wandel blockieren, Stuttgart : Reclam.

2 Espaces européens communs de données | Shaping Europe's digital future (europa.eu).

3 Küsters et Kullas (2023), Le Chips Act peut-il favoriser la résilience de l'Europe ?, Audit Committee Quarterly II/2023.

4 La Commission lance le paquet d'innovation sur l'IA (europa.eu).

5 LEAM étude de faisabilité 2023 - KI-Verband ; Launching an 'AI moonshot' to develop a European large language model is the game changer that Europe needs - CEPS.

6 Règles de l'UE pour ChatGPT et Aleph Alpha : l'Allemagne et la France contre (faz.net).

7 Foroohar (2022), Homecoming : The Path to Prosperity in a Post-Global World, Penguin Random House.

8 The Economist (2024), How San Francisco staged a surprising comeback (Feb 12th 2024).

9 Voir les statistiques chez : Potentiel économique de l'IA générative | McKinsey.

10 Voir par exemple : Un sondage montre que l'IA est en mauvaise posture dans les entreprises allemandes - Tagesspiegel Background. En général, sur le retard : Gutachten zu Forschung, Innovation und Technologischer Leistungsfähigkeit Deutschlands 2024 (e-fi.de), p. 116ff.

outils d'IA comme ChatGPT pourraient automatiser 60 à 70 % du temps de travail des travailleurs à haut niveau connaissances, les cadres sont jusqu'à présent plutôt réticents - notamment parce que la technologie n'est pas considérée comme mûre ou trop imprécise[11]. Malgré toutes les inquiétudes concernant les erreurs et même les risques existentiels liés à l'IA, il existe également un risque de ne pas profiter des avantages des technologies d'IA par excès de prudence, comme l'ont même souligné récemment les Nations unies[12]. En effet, on néglige souvent les progrès méthodologiques qui ont été réalisés en très peu de temps dans le domaine de l'IA générative. La crainte, également souvent exprimée, de tomber dans une dépendance ultérieure lors de l'intégration de technologies étrangères semble moins urgente pour le moment, compte tenu de la forte pression concurrentielle. De plus, ce risque stratégique est moins important qu'on ne le pense généralement : La disponibilité de modèles de langage modernes permet justement aux petites équipes de données non spécialisées d'obtenir des applications flexibles en utilisant uniquement le langage naturel, sans avoir besoin de code ou de modules spécifiques[13]. Cela réduit considérablement les coûts d'apprentissage et de changement (« switching costs »), diminue le potentiel de dépendance ultérieure et offre un fort contraste avec les défis précédents dans l'environnement du marché numérique[14].

Pour les PME européennes, il est donc grand temps d'intégrer des modèles linguistiques d'IA de qualité dans leurs processus internes et externes. Cet input du cep sert d'aperçu conceptuel des facteurs à prendre en compte à cet égard et esquisse dix éléments centraux d'une stratégie en matière de technologies linguistiques pour les PME, qui peuvent être utilisés comme base pour l'élaboration d'une politique interne concernant l'utilisation d'outils d'IA générative[15]. Ces éléments vont de la conception des requêtes (« prompts ») aux préoccupations en matière de protection des données. De manière générale, cette publication ne vise pas à fournir des conseils juridiques, mais à informer sur les possibilités et les applications de ces nouveaux outils et à faire comprendre où se situent (aujourd'hui encore) leurs problèmes. Bien entendu, il convient d'examiner au cas par cas si l'utilisation de modèles linguistiques est judicieuse ou non du point de vue de l'entreprise, mais en tant que « technologies à usage général », ils sont sur le point d'être utilisés à grande échelle et dans tous les secteurs, ce qui, selon les estimations de la littérature actuelle, pourrait avoir une influence directe sur 10 à 30 % de tous les travailleurs en Europe[16]. Dans ce contexte, l'intégration des technologies linguistiques offre non seulement l'opportunité d'accroître l'efficacité et la capacité d'innovation, mais aussi de renforcer l'Europe dans la concurrence mondiale. Les PME devraient agir dès maintenant de manière proactive afin de profiter des avantages de la technologie IA - tout en évaluant soigneusement les risques.

2 Facteurs d'utilisation de l'IA générative dans les PME

[11] Voir les statistiques chez : Top 30 Must Know Generative AI Stats in 2024 (aimultiple.com).
[12] UN AI Advisory Body, Interim Report : Governing AI for Humanity, December 2023, interim_report.pdf (un.org), p. 12.
[13] Voir également à ce sujet l'argumentation présentée ci-dessous dans la section 2.1.
[14] Commission (2024), Communication de la Commission sur la définition du marché en cause aux fins du droit de la concurrence de l'Union, Bruxelles, le 8.2.2024, C(2023) 6789 final, point 98.
[15] Il ne s'agit pas de conseils juridiques, par exemple en ce qui concerne les questions de protection des données encore en suspens. Pour un exemple de politique interne en matière d'IA, voir par exemple : BBC (2024), Guidance : The use of Artificial Intelligence (bbc.co.uk).
[16] Mauro Cazzaniga et al. (2024), Gen-AI : Artificial Intelligence and the Future of Work (imf.org). Voir également : Albanesi, Stefania et Dias da Silva, Antonio et Jimeno, Juan F. et Lamo, Ana et Wabitsch, Alena, Nouvelles technologies et emplois en Europe (2023). Document de travail du NBER n° w31357.

Qu'est-ce que l'IA générative et que recouvre la notion de modèles linguistiques ? Dans le contexte de ChatGPT et autres, l'abréviation IA est aujourd'hui le plus souvent utilisée pour désigner une classe de modèles avancés qui sont entraînés à produire des contenus qui ne se distinguent guère du travail humain - qu'il s'agisse de textes, d'images, de code ou même de courtes vidéos. Au cœur de cette technologie se trouvent les grands modèles linguistiques (« *Large Language Models* », LLM)[17], qui, en traitant de grandes quantités de texte, apprennent à saisir les nuances du langage humain et à les utiliser pour l'exploration créative. Les modèles linguistiques définissent une distribution de probabilité pour les séquences de mots et peuvent donc être utilisés à des fins génératives en prédisant les prochains mots les plus probables au début d'un texte. Ces dernières années, ces modèles sont devenus de plus en plus grands (c'est-à-dire qu'ils se basent sur plus de données d'apprentissage et utilisent plus de paramètres dans le modèle), ce qui a permis d'améliorer de manière significative leurs capacités de prédiction de texte (« *scaling law* »)[18]. Lors de la mise à l'échelle de ces modèles linguistiques, de nouvelles capacités peuvent apparaître soudainement et de manière imprévisible, comme le calcul, la réponse à des questions et la synthèse de textes, qui ne sont pas directement entraînées, mais uniquement apprises par l'observation du langage naturel (« capacités émergentes »)[19]. Les sauts énormes et imprévisibles dans les capacités de ces modèles ont récemment conduit à un véritable « boom de l'IA ».

Grâce aux progrès décrits, les modèles les plus récents peuvent non seulement restituer des informations existantes de manière cohérente et contextuelle, mais aussi générer des contenus originaux sur la base des modèles appris. Cela ouvre de nombreuses possibilités d'application pour les entreprises, de la rédaction automatisée de textes à la génération d'œuvres créatives en passant par le développement de chatbots. Entre-temps, les experts ont rassemblé plus d'une centaine de cas d'application généraux et spécifiques à un secteur de l'IA générative, dont beaucoup devraient également concerner les PME[20]. Les processus qui impliquent un travail important avec des mots, des images, des chiffres et des sons (ce que l'on appelle le travail WINS, abréviation de *Words, Images, Numbers, Sounds*) sont ceux qui devraient le plus bénéficier de la nouvelle technologie[21]. Les outils d'IA générative peuvent par exemple faciliter la rédaction de textes de marketing et de vente, soutenir le développement d'idées de marketing créatives, créer automatiquement des modèles de documents ou reconnaître les mises à jour réglementaires[22]. Un exemple particulièrement impressionnant est fourni par le prestataire de services de paiement Klarna, dont l'assistant IA basé sur ChatGPT a pris en charge en un mois le travail de 700 collaborateurs à temps plein, résolvant ainsi de manière plus précise les problèmes rencontrés dans les discussions avec les clients, ce qui a ensuite entraîné une baisse de 25 % des demandes[23].

L'analyse suivante ne s'arrête pas à une pesée générale des avantages et des inconvénients de l'IA générative, mais tente d'identifier les éléments clés pour une intégration rapide des modèles

---

17 Comme GPT (« Generative Pre-trained Transformer »). Pour une vue d'ensemble, voir : [2402.06196] Large Language Models : A Survey (arxiv.org).

18 Sardana et asl. (2023), [2401.00448] Beyond Chinchilla-Optimal : Accounting for Inference in Language Model Scaling Laws (arxiv.org).

19 Wei et al. (2022), [2206.07682] Emergent Abilities of Large Language Models (arxiv.org).

20 Top 100+ Generative AI Applications / Use Cases in 2024 (aimultiple.com).

21 Où votre entreprise doit-elle commencer avec GenAI ? (hbr.org).

22 Pour les exemples suivants, voir : Potentiel économique de l'IA générative | McKinsey.

23 L'assistant Klarna AI gère deux tiers des conversations du service client au cours de son premier mois.

linguistiques dans les processus des PME européennes et allemandes. Une stratégie efficace en matière de technologies linguistiques devrait prendre en compte les dix éléments suivants.

2.1 Analyse des besoins : comprendre l'IA générative d'un point de vue conceptuel

Les PME devraient d'abord effectuer une analyse approfondie de leurs besoins afin de comprendre quels processus internes et externes peuvent être améliorés par l'intégration de modèles linguistiques. Actuellement, de nombreuses entreprises échouent dans l'intégration parce qu'elles considèrent à tort, d'un point de vue méthodologique, l'IA générative comme une forme traditionnelle d'automatisation et non comme un agent de soutien qui devient plus intelligent au fil du temps[24]. Le développement de l'apprentissage automatique (AA) et son déploiement progressif dans les entreprises au cours de la dernière décennie offrent un parallèle intéressant qui montre à quel point il peut être difficile de passer de la simple fascination pour une nouvelle technologie à une compréhension stratégique de ses applications concrètes[25]. Bien que les fonctions AA avancées telles que la reconnaissance d'image et la reconnaissance vocale soient devenues de plus en plus connues dans les années 2010, de nombreuses entreprises ne savaient pas au départ comment utiliser de telles techniques, surtout si elles n'étaient pas directement liées au cœur de leur activité. Ce n'est qu'avec le temps que la perception de l'AA en tant qu'outil pour des tâches spécifiques a changé pour devenir un système de reconnaissance de formes très sophistiqué. En quels problèmes existants pouvaient être reformulés en problèmes de reconnaissance des formes, les entreprises et les start-ups ont créé de nouvelles opportunités commerciales. La valeur ajoutée des nouvelles technologies ne consiste donc pas seulement à les intégrer dans les processus commerciaux existants, mais aussi à redessiner ces processus.

Pour l'IA générative, comme par exemple l'utilisation de grands modèles de langage, il faut maintenant procéder à un changement conceptuel similaire à celui de l'intégration progressive de l'AA. À quels problèmes fondamentaux d'un domaine d'activité donné cette technologie peut-elle être appliquée de manière pertinente ? Contrairement à l'automatisation classique par des robots, la fonctionnalité de l'IA générative se comprend mieux par sa fonction de dialogue, qui permet à la technique et à l'homme de partager des responsabilités de manière dynamique[26]. Pour ce faire, il est utile d'imaginer l'IA générative comme l'équivalent de millions de stagiaires, c'est-à-dire imaginatifs et énergiques, imprécis et quelque peu imprévisibles, mais moins coûteux et bien plus évolutifs que de vrais stagiaires[27]. Par exemple, la littérature sur le *design thinking* visant à promouvoir l'intelligence institutionnelle a montré que les résultats des grands modèles de langage (LLM) doivent être considérés comme des idées et non comme des réponses définitives, et que ces systèmes doivent donc être positionnés en interne comme un outil d'aide à la perception humaine[28]. En d'autres termes : Alors que l'apprentissage automatique classique a pu être intégré dans les processus commerciaux au cours des dix dernières années grâce à sa capacité de reconnaissance des formes, l'IA générative sera à la disposition des entreprises comme un partenaire de dialogue itératif, comparable à un pool de stagiaires. Pour les

[24] Votre organisation n'est pas conçue pour travailler avec GenAI (hbr.org).
[25] Voir à ce sujet l'essai instructif dans : Benedikt Evans, « Abstracting Ai », in : Benedict's Newsletter : No. 528 (20. Feb. 2024).
[26] Pour cette approche « Designing for Dialogue », voir : Your Organization Isn't Designed to Work with GenAI (hbr.org).
[27] Voir : Giacomelli, G. (2024), Au-delà de « l'humain dans la boucle » : l'IA fiable dans les flux de travail d'entreprise (linkedin.com).
[28] Rick et al. (2023), Supermind Ideator : Exploring generative AI to support creative problem-solving (arxiv.org).

PME, la première question conceptuelle qui se pose est donc de savoir comment elles peuvent se transformer en interne de manière à pouvoir utiliser au mieux cette fonction de dialogue. Dans cette mesure, la valeur du LLM réside dans son intégration au sein de systèmes plus vastes et non dans son utilisation individuelle.

La définition des objectifs doit se baser sur cette analyse et fixer des objectifs clairs et mesurables pour la mise en œuvre des technologies linguistiques. Un bon exemple est le secteur de la publicité et du marketing, où l'IA générative est déjà utilisée pour de nombreuses applications, telles que la création de contenus écrits et de textes publicitaires (58 %), la recherche de mots-clés SEO (43 %) et les résumés d'e-mails, de réunions et de campagnes (38 %)[29]. Un exemple impressionnant est celui de Coca-Cola, qui a récemment fait état de l'utilisation de l'IA générative pour créer automatiquement des milliers de contenus marketing. L'entreprise a délibérément déplacé ses dépenses médiatiques des publicités télévisées, dont la production prenait souvent des mois et qui ne pouvaient plus être modifiées par la suite, vers des canaux numériques pour lesquels la technologie de modèles de langage a permis de produire quelque « 1 000 contenus contextuellement pertinents », dont les résultats étaient en outre mesurables en temps réel[30]. L'impact de ces dépenses marketing a été clairement visible dans les résultats financiers de l'entreprise. Dans l'ensemble, cette phase d'analyse des besoins et de réflexion conceptuelle et stratégique sur l'IA générative est cruciale pour la réussite de l'introduction des technologies linguistiques, car elle constitue la base de toutes les étapes ultérieures et garantit que l'introduction de la technologie est adaptée aux besoins et aux objectifs spécifiques de l'entreprise. Dans ce qui suit, nous partons du principe que la direction a déjà procédé à une telle analyse des besoins et fixé des objectifs.

2.2 Les dépendances stratégiques : réflexion sur le pouvoir de marché

Une fois qu'une entreprise a décidé d'utiliser l'IA générative, elle doit choisir entre l'utilisation de l'IA générative en tant que produit « cloud » d'un fournisseur tiers (« *Artificial Intelligence as a Service* », AIaaS) et la création et l'installation sur place (« *on premise* ») - un choix crucial pour les entreprises, car il a un impact direct sur l'évolutivité et la flexibilité des solutions d'IA[31]. L'AIaaS permet aux entreprises d'accéder à des algorithmes avancés et à des ressources de calcul (par exemple via Microsoft Azure) sans avoir à maintenir leur propre infrastructure, comme des serveurs GPU spécialisés. Inversement, opter pour des installations d'IA en interne peut permettre un alignement plus étroit avec les besoins spécifiques de l'entreprise, ce qui se traduit généralement par une efficacité et une précision accrues des applications d'IA. Outre le choix stratégique discuté dans cette section, les PME devraient également tenir compte des coûts énergétiques et des exigences de maintenance (section 2.9) ainsi que de la protection des données (section 2.7) lors de ce choix.

Comme elles ne disposent pas de grandes fermes de serveurs et de puces spécialisées pour entraîner leurs propres modèles linguistiques à grande échelle, à l'instar des grandes entreprises technologiques mondiales (GAFAM) et des start-ups qu'elles financent[32], les PME doivent choisir des technologies qui peuvent être facilement adaptées et qui offrent la possibilité d'ajouter des fonctionnalités ou des

[29] The GPT Store is not ChatGPT's 'app store' - but it's still significant for marketers (econsultancy.com).
[30] Le PDG de Coca-Cola : l'innovation est au service d'un 'avantage concurrentiel' (marketingweek.com).
[31] Bitkom (2024), L'IA générative dans l'entreprise. Questions juridiques relatives à l'utilisation de l'intelligence artificielle générative dans l'entreprise, p. 16, 41.
[32] Von Thun (2024), Euractiv - EU does not need to wait for the AI Act to act - Open Markets Institute ; Küsters et Kullas (2024), cep - Centre de politique européenne.

capacités supplémentaires si nécessaire. En outre, les technologies linguistiques choisies doivent être évolutives et flexibles afin de pouvoir suivre la croissance de l'entreprise et l'évolution de ses besoins.

Ces deux éléments plaident en faveur du recours à des LLM préentraînés. Ceux-ci peuvent être soit propriétaires et payants (par exemple les derniers modèles GPT d'OpenAI), soit gratuits et open-source (comme le modèle Llama de Meta). Il est important de reconnaître que tous les modèles gratuits ne sont pas automatiquement « véritablement » open-source et ne peuvent pas présenter de restrictions d'utilisation ; ce sont plutôt les modalités selon lesquelles l'accès est accordé qui sont déterminantes : Les modèles peuvent être entièrement fermés (non disponibles pour quiconque en dehors de l'organisation de développement), être mis à disposition via une interface web (par exemple l'API de GPT-4), offrir un accès uniquement fondé sur le cloud, donner un accès au réglage fin (GPT-3 d'OpenAI), divulguer leurs poids (comme Stable Diffusion de Stability AI et Llama 2 de Meta), ou être disponibles avec tous les poids, codes et données[33]. Pour les PME, les réflexions ci-dessus montrent que ces dernières catégories, souvent regroupées sous l'appellation « modèles de base ouverts », sont particulièrement appropriées, car les modèles sont publiés de manière transparente et avec des poids largement disponibles.

Un autre critère important est la performance des modèles. Comme celle-ci est déterminante pour un produit ou un service compétitif sur les marchés en aval sur lesquels la PME est active, même de petites différences jouent un rôle important. À cet égard, les différentes métriques et enquêtes sectorielles développées par les chercheurs montrent une image claire : le modèle GPT-4 développé par OpenAI est (actuellement) le leader du secteur et bat tous les autres LLM, tant dans les benchmarks classiques que dans les tests conçus pour être évalués par des humains[34]. Récemment, les modèles Claude 3 (Opus) et Gemini Ultra (de Google) ont toutefois réussi à rattraper leur retard[35], ce qui donne aux entreprises une certaine marge de manœuvre. Viennent ensuite des modèles open source comme Llama de Meta, et plus récemment Mistral AI de France. Plutôt que d'essayer de construire elles-mêmes des modèles coûteux ou d'attendre que des fournisseurs nationaux exclusifs lancent des modèles équivalents, les PME européennes devraient donc se tourner le plus rapidement possible vers ces modèles déjà établis. Si l'on opte pour un modèle propriétaire (comme GPT-4) plutôt que pour un modèle open source en raison des différences de performance, il convient toutefois d'exclure l'utilisation des données saisies par le fournisseur d'IA (comme OpenAI) par un accord contractuel ou le choix d'une licence spécifique.

Cela ne crée-t-il pas une dépendance à long terme et, en raison de la répartition inégale du pouvoir sur le marché, des dépendances défavorables et des effets de verrouillage ? En ce qui concerne l'utilisation de l'IA générative dans les entreprises, Bitkom met en garde contre « l'apparition d'une dépendance vis-à-vis de prestataires de services externes, par exemple en raison de la divulgation du savoir-faire ou des données, ainsi que des coûts consécutifs (en raison par exemple de mises à jour, de maintenance, de changement de prestataire de services externe à une date ultérieure) »[36]. Toutefois,

[33] Bommasani (2023), Considerations for Governing Open Foundation Models, Governing-Open-Foundation-Models.pdf (stanford.edu).

[34] State of AI Report 2023, Welcome to State of AI Report 2023. Voir aussi les évaluations actuelles dans : Chatbot Arena : Benchmarking LLMs in the Wild with Elo Ratings | LMSYS Org.

[35] Voir l'analyse comparative de : Warren (2024), Putting GPT-4's new rivals to the test (exponentialview.co).

[36] Bitkom (2024), L'IA générative dans l'entreprise. Questions juridiques relatives à l'utilisation de l'intelligence artificielle générative dans l'entreprise, p. 16.

contrairement aux développements précédents du marché numérique, les dépendances stratégiques dans le domaine de la technologie de modèles de langage devraient être beaucoup moins importantes, car les modèles sont relativement faciles à remplacer en raison de l'accès par le langage naturel, ce qui réduit le pouvoir de marché des principaux développeurs. Comme l'a récemment fait remarquer un expert en IA : « *It's remarkable that we can control a multi-trillion parameter bit of software sitting on hundreds of gigabytes of input data with ordinary English* »[37]. Cette accessibilité par le biais d'un langage (relativement) simple signifie qu'un petit nombre de collaborateurs disposant d'un bagage technique rudimentaire peut déjà interagir avec des systèmes logiciels sophistiqués et influencer leur fonctionnement. En revanche, il fallait auparavant une compréhension approfondie de langages de programmation tels que l'assembleur ou la manipulation de mémoire pour obtenir un niveau de contrôle comparable.

Pour les PME, ce changement dans l'écosystème de l'IA a des répercussions importantes sur les coûts de transition potentiels en cas d'augmentation des prix ou de modèles obsolètes. La démocratisation du contrôle des logiciels par le langage naturel réduit le besoin de compétences informatiques spécialisées et rend l'adoption de nouvelles technologies plus facile et moins coûteuse pour les PME. Non seulement cela augmente leur flexibilité pour intégrer des solutions innovantes, mais cela permet également d'équilibrer les conditions de concurrence avec des concurrents plus importants, ce qui peut accélérer la transformation numérique et favoriser un environnement de marché plus compétitif. Cela suggère que les PME devraient évaluer les modèles de base ouverts déjà établis en fonction de leurs résultats de test et des ressources nécessaires à leur mise en œuvre, sans trop s'inquiéter des dépendances ultérieures. Il est important de choisir le bon modèle individuel pour s'assurer qu'il répond aux besoins spécifiques de l'entreprise et qu'il peut contribuer efficacement à la réalisation des objectifs fixés. Mais la rapidité de la sélection est encore plus importante - plus l'introduction de l'IA générative commence tôt, plus il y a de temps et d'espace pour les expérimentations nécessaires.

2.3 Fine-Tuning et RAG : personnaliser l'IA avec ses propres sources de données

Les LLM préentraînés sont désormais largement répandus sur le web (par exemple via la plateforme populaire HuggingFace) et sont considérés comme d'excellentes technologies « à usage général » qui ne nécessitent pas ou peu d'exemples spécifiques à une tâche pour effectuer des tâches complexes, de la rédaction de communiqués de presse et de rapports d'activité plus courts à la création de graphiques, de présentations ou même d'applications. Des modèles pré-entraînés célèbres, faciles à télécharger ou à utiliser via une interface de programmation (API), ont été créés par Meta, Google et Mistral. Mais pour intégrer avec succès de tels modèles linguistiques préentraînés dans les processus commerciaux existants des PME, il faut les adapter au cas d'application concret. Deux techniques sont notamment disponibles à cet effet : le Fine-Tuning et la génération augmentée de récupération (RAG). Toutes deux permettent d'optimiser les LLM en fonction du contexte en ajoutant d'autres sources de données, mais elles possèdent chacune leurs propres avantages et inconvénients[38].

Le réglage fin des grands modèles linguistiques consiste à adapter les LLM déjà entraînés sur un ensemble de données général en les entraînant sur un ensemble de données plus restreint et spécifique à une tâche, afin qu'ils soient mieux adaptés à un domaine particulier. Un bon exemple est « LEGAL-

[37] Citation tirée de : The brilliant, complicated simplicity of ChatGPT (exponentialview.co).
[38] Pour une comparaison, voir : Retrieval augmented generation : Keeping LLMs relevant and current - Stack Overflow.

BERT », qui a optimisé le modèle linguistique BERT bien connu pour le domaine juridique et l'application de la technologie juridique en l'entraînant en plus sur différents textes juridiques (par exemple la législation, les procédures judiciaires, les contrats)[39]. Cette technique permet donc aux modèles d'adapter leurs vastes connaissances générales aux exigences différenciées de certaines applications, ce qui améliore leur performance pour des tâches spécifiques et leur précision. Pour les PME, le réglage fin des LLM offre la possibilité d'adapter la technologie linguistique de pointe de l'IA à leurs besoins commerciaux individuels, sans avoir à supporter les coûts élevés de développement et d'entraînement de modèles de base entièrement nouveaux. En ajustant finement les LLM sur des ensembles de données et de documents internes qui reflètent leur contexte commercial spécifique, les PME peuvent obtenir des résultats plus précis et plus efficaces dans des domaines tels que l'automatisation du service à la clientèle, le marketing personnalisé et la création de contenu, ce qui leur permet de se démarquer dans le secteur (puisque personne d'autre n'a accès à leurs données internes).

La génération augmentée de récupération (RAG) améliore l'efficacité des LLM en intégrant des sources d'information externes après l'apprentissage (dans la « phase de récupération »). Concrètement, l'algorithme recherche activement des bribes d'informations pertinentes en réponse aux requêtes des utilisateurs et les récupère de manière à ce qu'elles puissent ensuite être synthétisées par des modèles de langage génératifs, tels que des modèles basés sur des transformateurs tels que GPT, afin de produire des réponses cohérentes et pertinentes en termes de contexte[40]. L'accès aux faits les plus récents et les plus pertinents devrait permettre d'améliorer considérablement les réponses, par exemple lorsqu'elles sont utilisées dans des chatbots de questions-réponses pour les clients. Le RAG favorise également la transparence et la fiabilité en permettant aux utilisateurs de vérifier les sources d'information utilisées par l'IA. Malgré son potentiel, le concept de la RAG est encore inconnu dans de nombreuses entreprises, car il ne fait l'objet de discussions que depuis relativement peu de temps[41]. Alors que cette approche améliore l'adaptabilité et la précision des LLM en surmontant les limites de connaissances statiques inhérentes à ces modèles, elle implique également des exigences de calcul accrues, des temps de latence plus longs et des invites plus complexes. Il n'est donc recommandé (pour l'instant) que pour les cas d'application où la vitesse d'inférence et la consommation de ressources ne sont pas trop importantes. Bien que la construction d'un modèle de RAG soit relativement simple, selon un récent travail de synthèse, des adaptations importantes et une compréhension relativement approfondie du domaine d'application sont nécessaires pour aboutir à une application robuste et fiable[42]. Enfin, il convient de souligner que les modèles RAG ne sont pas une panacée et souffrent encore de problèmes méthodologiques[43]. Une évaluation récente des modèles RAG dans différents domaines cliniques a montré que l'inclusion des RAG réduisait significativement le nombre d'erreurs, mais que même dans le meilleur modèle (GPT-4 RAG), jusqu'à 30 % des affirmations n'étaient pas étayées par l'une des sources indiquées[44].

---

[39] Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, et Ion Androutsopoulos. 2020. LEGAL-BERT : Les Muppets tout droit sortis de l'école de droit. In Findings of the Association for Computational Linguistics : EMNLP 2020, pages 2898-2904, en ligne. Association for Computational Linguistics.

[40] Voir : 12 Retrieval Augmented Generation (RAG) Tools / Software dans '23 (aimultiple.com).

[41] Voir par exemple Andrew Ng : My Daily Note Taking Device : reMarkable 2 (2023) (youtube.com).

[42] Fatehkia et al. (2024), [2402.07483] T-RAG : Lessons from the LLM Trenches (arxiv.org).

[43] Voir à ce sujet la position sceptique de Gary Marcus : Non, RAG is probably not going to rescue the current situation (substack.com).

[44] Wu et al. (2024), [2402.02008] How well do LLMs cite relevant medical references ? An evaluation framework and analyses (arxiv.org).

Ensemble, Fine-Tuning et RAG permettent le développement d'applications LLM sur mesure pour les PME, spécialisées dans des domaines spécifiques et utilisant des connaissances contextuelles. À l'avenir, de tels assistants d'intelligence artificielle d'entreprise joueront un rôle de plus en plus important pour rendre les processus de travail plus efficaces en utilisant mieux les connaissances internes. Un bon exemple en est GitHub Copilot, qui utilise l'environnement de programmation existant comme base de connaissances pour contextualiser les demandes des programmeurs internes et mieux y répondre. On peut s'attendre à ce que des « copilotes » similaires soient désormais formés par de nombreuses entreprises.

2.4 Développer les compétences internes en PNL : (Prompt) Design

Les PME doivent développer une compréhension de base des modèles linguistiques disponibles et du traitement du langage naturel (NLP). Cela implique non seulement une évaluation des différents modèles en termes de capacités, de limites et de coûts de suivi éventuels (voir ci-dessus), mais aussi et surtout le développement en interne de compétences en matière de prompting. Les invites dans les modèles d'IA génératifs sont des entrées textuelles qui contrôlent la sortie du modèle et vont de simples questions à des tâches détaillées[45]. Dans les modèles générateurs d'images comme DALL-E, les invites sont souvent descriptives, tandis que dans les modèles linguistiques comme GPT-3, elles peuvent aller de simples questions à des problèmes complexes. En raison de l'évolution actuelle vers ce que l'on appelle les LLM multimodaux, les instructions textuelles peuvent aujourd'hui généralement être complétées par des images, des textes complémentaires ou des données audio, qui sont ensuite pris en compte par le système d'IA lors de la création des formats correspondants. Comme le développement d'invites appropriées nécessite un investissement considérable en temps et en personnel, celles-ci constituent probablement même un secret commercial digne de protection au sens juridique du terme[46].

La stratégie de prompting, sans doute la plus connue, est celle qui consiste à décomposer les tâches complexes en leurs éléments constitutifs. Jusqu'à présent, les preuves scientifiques soutiennent cette stratégie « pas à pas » - il semble que la qualité du travail s'améliore lorsque le modèle est invité à décomposer une tâche en ses éléments. Une étude empirique a pu démontrer que cette amorce de « chaîne de pensée » (CoT) peut guider avec succès les modèles linguistiques à travers des processus de pensée à plusieurs niveaux afin d'atteindre des performances élevées dans des tâches complexes telles que l'arithmétique et le raisonnement symbolique[47]. Les modèles linguistiques sont donc capables d'obtenir de bonnes performances dans des tâches sans exemples codés coûteux, en ajoutant une invitation à réfléchir aux problèmes étape par étape (dans l'original anglais, par exemple : « *Let's think step by step* »). Des études ultérieures ont montré que les LLM peuvent améliorer de manière significative la productivité et la qualité du processus de génération d'idées lorsqu'un tel CoT-Prompting est utilisé[48]. Les PME peuvent utiliser de telles techniques d'incitation pour mieux libérer les capacités cognitives des LLM à leurs fins spécifiques.

---

[45] Pour une vue d'ensemble, voir : [2401.14423] Prompt Design and Engineering : Introduction and Advanced Methods (arxiv.org).

[46] Bitkom (2024), L'IA générative dans l'entreprise. Questions juridiques relatives à l'utilisation de l'intelligence artificielle générative dans l'entreprise, p. 41.

[47] Kojima et al. (2022), [2205.11916] Large Language Models are Zero-Shot Reasoners (arxiv.org).

[48] Meincke et al. (2024), Prompting Diverse Ideas : Increasing AI Idea Variance, SSRN.

Certaines des techniques d'invite les plus efficaces semblent parfois contre-intuitives et sont le résultat d'expériences surprenantes, ce qui montre l'importance d'un expert du domaine qui connaît bien les développements actuels. Des études ont par exemple montré que les appels aux émotions dans l'invite (par exemple « C'est personnellement très important pour moi ») conduisent à des résultats significativement meilleurs. Dans l'article en question, les chercheurs ont testé des invitations à des modèles de langage avec et sans émotions supplémentaires et ont constaté que ces dernières entraînaient une amélioration moyenne de 10,9 % dans les domaines de la performance, de la véracité et du sens des responsabilités[49]. On ne comprend pas encore le mécanisme sous-jacent, mais cela fonctionne. L'étude a testé des modèles proéminents tels que ChatGPT, Llama 2 et d'autres LLM ; il est donc probable que cela s'applique également aux modèles linguistiques internes des PME basés sur ces modèles préentraînés.

Outre la position exceptionnelle des designers prompts dans l'intégration des LLM, un deuxième groupe de professionnels, plus inattendu, joue un rôle particulier dans la pratique entrepreneuriale - les designers en ligne classiques. À l'ère de l'IA générative, la compétitivité et l'attractivité pour les clients ne dépendront pas seulement des capacités technologiques, mais aussi de la qualité et des capacités de conception des entreprises[50]. L'écosystème numérique qui est en train d'émerger, dans lequel les modèles linguistiques deviennent de plus en plus la nouvelle plateforme et infrastructure d'entrée sur Internet[51], récompense ceux qui offrent des interfaces utilisateur supérieures et une intégration sans faille - des attributs qui sont la marque de fabrique des designers qualifiés. Pour rester compétitives, les PME doivent donc donner la priorité au recrutement et à la promotion des talents en matière de design. Les start-ups, en particulier, peuvent acquérir un avantage concurrentiel en se concentrant sur un design innovant afin de créer de nouvelles expériences utilisateur transformatrices et de se démarquer ainsi d'un marché établi où les concurrents sont technologiquement compétents, mais peu doués pour le design.

2.5 Cas des hallucinations : adapter le taux d'erreur de l'IA à sa propre tolérance aux erreurs

L'intégration des LLM dans les processus commerciaux des PME n'est pas sans poser des défis, notamment en ce qui concerne la tendance de ces modèles à ce que l'on appelle des hallucinations - la génération d'informations plausibles, mais en fait fausses ou absurdes[52]. D'où proviennent ces erreurs ? Les grands modèles linguistiques ne sont pas conçus pour la recherche d'informations externes et leur performance est encore plus limitée par le volume et l'actualité des données avec lesquelles ils ont été entraînés. Lorsque les LLM ne disposent pas de suffisamment d'informations pour fournir une réponse fondée, ils fabriquent des réponses sur la base d'entrées antérieures. Ce phénomène représente un risque important pour les PME, car des imprécisions dans la communication avec les clients, la génération de contenu ou l'analyse des données peuvent conduire à la diffusion d'informations trompeuses, qui sapent la confiance des clients et peuvent éventuellement nuire à la réputation de la marque. D'un point de vue juridique, il convient de garder à l'esprit que de telles erreurs d'IA intégrées

[49] Li et al. (2023), Large Language Models Understand and Can Be Enhanced by Emotional Stimuli, 2307.11760.pdf (arxiv.org).
[50] C'est l'argument de : Belsky (2024), The Era of Abstraction & New Creative Tensions (implications.com).
[51] De manière générale sur l'influence future des modèles linguistiques, voir : [2305.07961] Leveraging Large Language Models in Conversational Recommender Systems (arxiv.org). Sur l'impact d'un monde numérique de plus en plus abstrait, voir : The Era of Abstraction & New Creative Tensions (implications.com).
[52] Explorer les grands modèles linguistiques (LLM) : AI et hallucinations | ZS.

dans des produits ou des services pourraient entraîner une rupture de contrat, une responsabilité et des amendes[53].

Outre la place laissée aux erreurs « involontaires », l'intégration de bots vocaux ouvre également un potentiel d'abus non négligeable, par exemple par des attaquants externes. Étant donné que les fonctionnalités des LLM peuvent être modulées de manière flexible par des invites en langage naturel (plutôt que par du code), elles sont vulnérables aux invites ciblées qui permettent aux attaquants de contourner les instructions et les contrôles initiaux. Des chercheurs ont décrit des vecteurs d'attaque qui permettent d'abuser à distance des applications intégrées aux LLM en insérant des invites ciblées dans des données susceptibles d'être consultées (*indirect prompt injection attacks*)[54]. Un groupe de pirates informatiques proches de la Russie a revendiqué des attaques qui ont temporairement mis ChatGPT hors service fin 2023[55]. Cela a entraîné des pannes partielles et, dit-on, des taux d'erreur plus élevés chez les utilisateurs de ChatGPT. Si une PME dépend d'un accès continu aux modèles GPT d'OpenAI, de telles attaques peuvent interrompre les processus commerciaux internes. Enfin, les chercheurs ont pu montrer qu'il était possible d'intégrer des « portes dérobées » dans les LLM, c'est-à-dire de les entraîner de manière à ce qu'ils adoptent un comportement trompeur et exécutent par exemple un code malveillant à une date ultérieure[56]. Ce comportement frauduleux, appelé « agents dormants » par les chercheurs, a persisté même après les procédures de formation standard en matière de sécurité. Ces résultats de recherche et ces expériences mettent en évidence la nécessité de contre-mesures efficaces pour sécuriser les systèmes exploités par LLM, tout en suggérant que les mesures existantes ne sont pas encore suffisantes.

Néanmoins, au cours des deux dernières années, les programmeurs et les utilisateurs ont découvert de nombreuses mesures qui peuvent aider à atténuer le niveau d'hallucinations et d'autres erreurs, ainsi que le potentiel d'abus externe[57]. Ainsi, le simple fait de fournir des exemples concrets (texte, code ou données) dans la requête adressée à un LLM peut déjà améliorer considérablement la pertinence et la qualité de la sortie. En outre, la technique RAG évoquée plus haut réduit le risque de résultats imprécis ou erronés, car elle recherche des informations pertinentes dans des sources crédibles et les utilise pour compléter les réponses du LLM[58]. Dans la littérature sur la conception à l'invite (voir la section 2.4), le risque d'erreur des modèles linguistiques est encore amélioré par l'intégration de certains composants techniques (« outils », « connecteurs » ou « compétences »)[59]. Ces extensions des LLM ordinaires permettent à l'outil linguistique d'accéder à des sources de données externes et d'interagir avec elles, ainsi que d'effectuer des tâches qui vont au-delà des capacités intégrées. Le potentiel d'utilisation ainsi élargi rend la technologie de modèles de langage plus attrayante pour un large éventail de PME européennes, car les tâches ainsi rendues possibles vont de la simple récupération de données à des interactions complexes avec des bases de données ou des API, en passant par des résumés de texte ou des traductions vocales sensibles au contexte. À l'avenir, les LLM pourraient même apprendre d'eux-mêmes quand et comment appeler et utiliser des outils externes via des API

[53] Bitkom (2024), L'IA générative dans l'entreprise. Questions juridiques relatives à l'utilisation de l'intelligence artificielle générative dans l'entreprise, p. 44.

[54] [2302.12173] Pas ce que vous avez signé pour : Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection (arxiv.org).

[55] Les hackers liés à la Russie réclament le crédit pour la fuite d'OpenAI cette semaine - BNN Bloomberg - OECD.AI.

[56] [2401.05566] Sleeper Agents : Training Deceptive LLMs that Persist Through Safety Training (arxiv.org).

[57] Pour un bon aperçu, voir : How to reduce hallucination in a Large Language Model (LLM) ? (linkedin.com).

[58] 12 Retrieval Augmented Generation (RAG) Tools / Software en '23 (aimultiple.com).

[59] Voir : [2401.14423] Prompt Design and Engineering : Introduction and Advanced Methods (arxiv.org).

simples[60]. Ils se transforment ainsi en agents d'intelligence artificielle qui peuvent en outre être intégrés dans des processus et produits physiques (voir également la section ci-dessous).

En résumé, il n'existe pas encore de cadre de vérification et de validation solide que les PME pourraient mettre en œuvre rapidement pour éliminer complètement les effets négatifs des hallucinations LLM et d'autres catégories d'erreurs. Cela peut avoir des conséquences juridiques : Un tribunal canadien a récemment décidé qu'Air Canada devait verser des dommages et intérêts à un passager parce que le chatbot alimenté par IA du service clientèle l'avait conseillé de manière trompeuse et que le passager avait donc dû payer presque le double de son billet d'avion[61]. Dans une certaine mesure, de telles erreurs resteront toujours possibles malgré les progrès de l'IA générative, car les modèles fonctionnent en fin de compte de manière probabiliste, c'est-à-dire qu'ils prédisent simplement une séquence de jetons donnée sans disposer d'un modèle sous-jacent du monde (ce qui explique que la sortie peut parfois être différente malgré un prompt identique). Il est donc essentiel pour les PME d'adapter le taux d'erreur de l'IA (et le potentiel d'abus) à leur propre tolérance aux erreurs internes, de sorte qu'en cas d'hallucination, l'intégrité et la fiabilité de l'entreprise ne soient pas compromises. En d'autres termes, avant toute utilisation, il convient de se demander si une hallucination dans ce domaine concret aurait des conséquences importantes ou pourrait être facilement corrigée. En outre, il est possible de se protéger à l'aide de dispositions contractuelles en matière de garantie, de responsabilité et d'exonération[62].

2.6 Les agents embarqués : Intégration dans les processus, produits et services

Dans la littérature sur l'IA, le terme « agent » décrit un système capable d'effectuer certaines tâches de manière autonome[63]. L'un des aspects les plus remarquables des modèles linguistiques de l'IA est leur capacité à utiliser des outils logiciels externes pour atteindre des objectifs prédéfinis. Tout comme les humains écrivent du code ou utilisent des logiciels qui dépassent leurs capacités immédiates, les LLM peuvent très bien imiter ce processus pour accomplir certaines tâches. Ils peuvent par exemple être formés à reconnaître quand il est utile d'avoir recours à une interface de programmation, à traiter les données reçues et à adapter leurs actions en conséquence[64]. Cela permet de développer des agents IA avancés qui utilisent différents logiciels pour améliorer leurs capacités ou combler leurs lacunes. Ces agents IA sont conçus pour interagir à la fois avec les utilisateurs et leur environnement, et pour prendre des décisions éclairées en fonction des entrées reçues et de leurs objectifs prédéfinis[65]. Ils sont destinés à des tâches qui requièrent un certain degré d'autonomie dans la prise de décision et la résolution de problèmes, au-delà de la simple génération de réponses.

Même si la mise en œuvre de ce modèle n'est pas encore prête pour la pratique, les experts prédisent que les développements ultérieurs d'agents fondés sur des LLM deviendront de plus en plus pertinents

---

60 Voir : [2302.04761] Toolformer : Language Models Can Teach Themselves to Use Tools (arxiv.org).

61 Air Canada ordonne de payer les frais de passagers après que le chatbot ait parlé de réductions pour le transport (gizmodo.com).

62 Bitkom (2024), L'IA générative dans l'entreprise. Questions juridiques relatives à l'utilisation de l'intelligence artificielle générative dans l'entreprise, p. 49.

63 Pour une vue d'ensemble, voir : [2401.14423] Prompt Design and Engineering : Introduction and Advanced Methods (arxiv.org).

64 [2302.04761] Toolformer : Les modèles de langage peuvent vous apprendre à utiliser des outils (arxiv.org).

65 [2401.14423] Conception et ingénierie de l'invite : Introduction et méthodes avancées (arxiv.org).

sur le plan commercial[66]. OpenAI travaille par exemple sur un agent IA qui prend le contrôle de l'appareil de l'utilisateur et permet au logiciel d'effectuer des clics, des saisies et d'autres actions[67]. De la même manière, Apple travaille actuellement à la mise en place de l'IA générative sur les appareils mobiles[68]. Google a déjà intégré ses vastes modèles linguistiques Gemini dans un grand nombre de ses services, notamment Android, l'application Google pour iOS et Gmail[69]. Ici, l'introduction de « personnalités » basées sur le LLM a contribué au développement de nombreux bots qui interagissent de manière autonome avec les utilisateurs et peuvent simuler des amitiés ou des intérêts mieux que jamais auparavant[70]. Ces assistants plus performants pourraient être combinés à l'avenir avec la synthèse vocale de l'IA. D'un point de vue juridique, il est intéressant de noter qu'un agent d'IA générative n'a pas de capacité juridique et commerciale propre, mais qu'il peut aider en tant qu'auxiliaire d'exécution dans certains processus, comme justement l'automatisation de tâches de routine et l'analyse de données (la responsabilité restant celle de l'opérateur)[71].

Que signifie cette évolution pour les entreprises européennes qui fabriquent des produits physiques ? Les PME ne devraient pas considérer les nouvelles technologies linguistiques comme des technologies purement textuelles qui restent sur des écrans d'ordinateur, mais réfléchir à temps à la manière dont elles peuvent être de plus en plus intégrées dans les produits et services physiques de leur secteur. On peut par exemple imaginer un agent IA basé sur le LLM qui aurait accès à une API d'achat, obtiendrait des informations de sources externes (par exemple un comparateur de prix) et effectuerait ensuite certains achats de manière autonome via l'API sur la base de ces informations (par exemple, se faire livrer le produit le moins cher du moment). De la même manière, les PME pourraient utiliser des agents IA pour optimiser leurs systèmes de gestion de la chaîne d'approvisionnement ou proposer des assistants qui guident les clients tout au long du processus d'achat et permettent des configurations de produits personnalisées. Ce type d'interaction homme-machine peut non seulement renforcer la fidélisation de la clientèle, mais aussi fournir un aperçu des préférences des clients, qui peuvent à leur tour être utilisées comme données d'entraînement pour le développement futur des produits et l'amélioration des modèles.

Toutefois, une confiance excessive dans les agents fondés sur des LLM pour les processus décisionnels critiques, sans surveillance adéquate, peut conduire à des erreurs stratégiques, raison pour laquelle ils devraient être soigneusement testés et ne pas être utilisés pour des fonctions critiques. Un exemple dramatique souligne ce risque : des scientifiques ont étudié l'utilisation d'agents fondés sur des LLM dans des jeux stratégiques de type militaire et ont trouvé « des formes et des schémas d'escalade difficilement prévisibles »[72]. Ils ont constaté que les modèles développaient une dynamique négative en s'appuyant sur des « justifications inquiétantes » telles que des tactiques de première frappe. En général, la recherche a montré qu'il est possible de modifier l'orientation morale et éthique d'un

---

66 Cette évaluation se base principalement sur des entretiens avec des experts en marge du 8e Open European Dialogue à Helsinki, voir : openeuropeandialogue.org/download-file/2296/. Voir également la discussion optimiste de : Lazar (2024), Can philosophy help us get a grip on the consequences of AI ? | Essais d'Aeon.

67 OpenAI change le terrain de bataille de l'IA en un logiciel qui opère des appareils, automatise des tâches - The Information.

68 Apple booste ses plans pour apporter l'IA générative aux iPhones (ft.com).

69 Le Gemini de Google est désormais présent dans tout. Voici comment vous pouvez l'essayer. | MIT Technology Review.

70 [2303.06135] Rewarding Chatbots for Real-World Engagement with Millions of Users (arxiv.org). Voir aussi : My AI lover | Psyche Films.

71 Bitkom (2024), L'IA générative dans l'entreprise. Questions juridiques relatives à l'utilisation de l'intelligence artificielle générative dans l'entreprise, p. 59.

72 [2401.03408] Escalation Risks from Language Models in Military and Diplomatic Decision-Making (arxiv.org). Voir aussi : Could GPT-5 revolutionize military strategy ? (substack.com).

modèle d'IA, même pour les modèles les plus puissants comme GPT-4[73]. Compte tenu de tels risques, les agents autonomes de modèles linguistiques ne devraient donc pas être utilisés dans un premier temps pour prendre des décisions stratégiques importantes.

2.7 Les conditions juridiques : Protéger les données et les connaissances, exploiter la loi sur l'IA

Malgré l'euphorie, les PME doivent tenir compte de certains cadres juridiques lors de l'intégration de LLM dans leurs processus commerciaux. Tant l'utilisation des données injectées dans les modèles d'IA que l'utilisation des résultats de l'IA touchent à un certain nombre de problématiques juridiques, notamment le droit d'auteur, la protection des données, les questions de responsabilité et le droit du travail[74]. La protection de la vie privée, en particulier, fait l'objet d'un débat animé visant à mettre à jour les règles existantes afin de suivre le rythme juridique d'un monde de plus en plus centré sur les données[75]. Les PME établies en Europe doivent tenir compte du règlement général de l'UE sur la protection des données (RGPD), de la loi européenne sur l'IA dont les négociations finales se sont terminées fin décembre 2023, ainsi que d'autres réglementations nationales et internationales (en Allemagne, par exemple, la loi fédérale sur la protection des données et la loi sur la protection des secrets commerciaux). Elles devraient garantir que les données personnelles et les secrets commerciaux sont protégés et que l'utilisation de l'IA générative, par exemple sous forme de chatbots, se fait de manière transparente. Trois domaines problématiques sont abordés plus en détail ci-dessous : Premièrement, des données sensibles peuvent s'échapper des systèmes et, le cas échéant, devenir visibles pour les attaquants ou même les personnes non concernées (« fuite de données »). Deuxièmement, il existe toujours des problèmes de droits d'auteur concernant les sources d'où provient la technologie de modèles de langage. Troisièmement, de nouvelles obligations - mais aussi des droits intéressants - découlent de la législation européenne sur l'IA.

Les LLM peuvent être instrumentalisés pour espionner des données privées de personnes ou d'entreprises[76]. En utilisant des méthodes relativement simples, comme demander de répéter indéfiniment un mot comme « poème », des chercheurs ont réussi à faire en sorte que ChatGPT révèle involontairement une grande partie de ses données d'entraînement[77]. Pour les PME, la question est donc de savoir comment découvrir en toute sécurité de nouveaux cas d'application pour les LLM sur la base de données internes ; notamment parce que tout ce qui est téléchargé dans les services LLM commerciaux pourrait potentiellement être enregistré comme futures données d'entraînement[78]. OpenAI a entretemps réagi à certaines vulnérabilités connues et a pris des mesures pour empêcher les attaquants d'envoyer à leur insu les données des utilisateurs à des serveurs externes[79]. Malgré ces améliorations, des inquiétudes subsistent, car des fuites de données sont toujours possibles avec certaines méthodes

---

73 [2311.05553] Suppression des protections RLHF dans GPT-4 via Fine-Tuning (arxiv.org).

74 Bitkom (2024), L'IA générative dans l'entreprise. Questions juridiques relatives à l'utilisation de l'intelligence artificielle générative dans l'entreprise.

75 Les lois existantes et proposées sur la protection de la vie privée réglementent implicitement le développement de l'IA, mais sont considérées par les experts comme insuffisantes pour réglementer suffisamment la course actuelle aux données - la réglementation va donc encore évoluer. Pour un aperçu, voir : King et Meinhardt (2024), White-Paper-Rethinking-Privacy-AI-Era.pdf (stanford.edu).

76 Trois façons dont les chatbots AI sont un désastre pour la sécurité | MIT Technology Review.

77 ChatGPT peut fuir les données de formation, violer la vie privée, dit Google's DeepMind | ZDNET.

78 Use Open Source for Safer Generative AI Experiments (mit.edu).

79 Voir l'analyse de : OpenAI Begins Tackling ChatGPT Data Leak Vulnerability - Embrace The Red.

d'attaque. Il reste à voir quelle sera l'efficacité de ces mesures à long terme et si la sécurité des données des PME pourra être garantie.

Le respect des droits d'auteur fait l'objet d'un débat académique animé et de plusieurs batailles juridiques qui ne sont toujours pas terminées à ce jour. La plus connue est sans doute la plainte déposée par le New York Times (NYT) contre Microsoft et OpenAI, qui affirme que des services d'IA comme ChatGPT utilisent illégalement des contenus du journal. Les plaignants exigent que tous les LLM formés sur leurs articles soient retirés. Au cœur de la plainte se trouve l'accusation du Times selon laquelle les LLM sont des « machines à copier en masse » qui produisent à la demande des « copies presque conformes » de parties importantes d'articles de NTY[80]. Selon les observateurs du procès, il semble que la plainte présente mal le fonctionnement des LLM et utilise des exemples sélectifs qui fournissent un récit moralement attrayant, mais pas d'argument juridique solide[81]. En fait, l'IA générative n'est pas basée sur un algorithme prédéfini, mais sur des méthodes statistiques. Pour simplifier, on peut dire que les modèles linguistiques n'apprennent pas par cœur des textes originaux, mais seulement des probabilités. Le fait que le modèle reproduise parfois presque mot pour mot certains articles du NYT est donc plutôt dû au fait que ces textes ont été soit très souvent copiés ou partagés sur Internet (ils font donc partie du modèle statistique), soit qu'ils sont très spécifiques (thématiquement, linguistiquement) et peuvent donc être déclenchés par des « prompts » tout aussi spécifiques. Selon les informaticiens, il n'est donc pas utile de comparer dans quelle mesure l'output de ChatGPT correspond exactement aux articles originaux. Si la décision des juges se concentre sur ce point, cela pourrait, selon ces observateurs, rendre plus difficile la résolution du problème sous-jacent, à savoir l'absence de participation financière des auteurs au savoir qu'ils ont créé[82]. De même, l'opinion dominante dans le domaine juridique ne considère pas pour le moment l'extraction d'informations à partir d'œuvres protégées et l'adaptation des valeurs de pondération du réseau neuronal sur lequel repose une technologie de modèles de langage d'IA comme ChatGPT comme une reproduction punissable des œuvres entraînées[83]. Même si le tribunal se prononce finalement contre le NYT, les PME devraient garder un œil sur cette question juridique, car elle aura un impact sur les modèles qu'elles peuvent utiliser et sur la possibilité d'entraîner elles-mêmes des systèmes d'IA à l'aide de données accessibles au public sur Internet. L'autorité italienne de protection des données Garante a récemment terminé son enquête sur ChatGPT, qui avait conduit à l'interdiction temporaire du chatbot l'année dernière, et a constaté plusieurs violations des règles de protection des données[84]. Des enquêtes sur les pratiques d'OpenAI en matière de protection des données sont également en cours en Espagne, en France et en Allemagne. Le développement de politiques internes visant à respecter les lois existantes en matière de protection des données est donc crucial pour les PME afin de conserver la confiance de leurs clients à long terme malgré l'utilisation de l'IA générative.

Les nouvelles règles de l'UE en matière d'IA, sur lesquelles les négociateurs des États membres se sont mis d'accord fin 2023 après d'intenses discussions, pourraient également contribuer à renforcer la

[80] NYT_Complaint_Dec2023.pdf (nytimes.com).
[81] Le procès en matière de droits d'auteur du New York Times contre OpenAI menace l'avenir de l'IA et de l'utilisation équitable - Center for Data Innovation.
[82] La fin de course de Generative AI autour du droit d'auteur ne sera pas résolue par les tribunaux (aisnakeoil.com).
[83] Bitkom (2024), L'IA générative dans l'entreprise. Questions juridiques relatives à l'utilisation de l'intelligence artificielle générative dans l'entreprise, p. 38.
[84] ChatGPT : Garante privacy, notificato a OpenAI l'atto di contestazione... - Garante Privacy.

confiance des clients[85]. Ces règles visent à garantir que les modèles d'IA sont utilisés en Europe de manière éthique, sûre et respectueuse et qu'ils protègent les droits fondamentaux. Le respect des règles est obligatoire pour tous les fournisseurs, distributeurs ou exploitants de systèmes et de modèles d'IA mis sur le marché dans l'UE[86]. Les exigences varient en fonction du niveau de risque et comprennent quatre catégories de risque, allant d'inacceptable à minimal, chacune étant assortie d'obligations spécifiques et de délais de six à 36 mois. Par exemple, alors que les filtres anti-spam ne représentent qu'un risque faible, les tests de solvabilité seraient considérés comme un risque élevé, car ils comportent un risque de discrimination. Des obligations spécifiques s'appliquent à l'IA générative, selon qu'il s'agit ou non d'un modèle open source. Les PME qui prévoient d'intégrer la technologie de modèles de langage devraient comprendre la catégorie de risque de leur système d'IA et se préparer dès à présent à se conformer aux règles européennes en la matière.

En ce qui concerne le sujet traité ici, il est surtout pertinent que tous les modèles de base d'IA (indépendamment du risque) doivent satisfaire à des exigences de transparence avant d'être mis sur le marché, par exemple en ce qui concerne l'utilisation, l'architecture, les données d'entraînement et d'autres documents techniques[87]. Une réglementation plus stricte a été introduite pour les modèles de base d'importance systémique (définis provisoirement en fonction du nombre de FLOPS). Il s'agit de modèles de base qui sont entraînés avec de grandes quantités de données et dont la complexité et les performances sont largement supérieures à la moyenne, ce qui peut entraîner la propagation de risques systémiques le long de la chaîne de valeur. Ces modèles de base d'importance systémique (appelés systèmes d'IA à haut risque) doivent préalablement faire l'objet d'une procédure d'évaluation de la conformité (art. 8 et suivants et art. 43 du règlement européen sur l'IA). Pour les PME qui envisagent d'intégrer un chatbot basé sur l'IA, il est important de savoir que le règlement introduit de nouvelles possibilités de divulgation et de traçabilité du contenu généré artificiellement ainsi que d'information des utilisateurs finaux sur le fait qu'ils ont affaire à un chatbot basé sur l'IA (art. 50 du règlement UE sur l'IA). Étant donné que les PME sont susceptibles de recourir fréquemment à des modèles de base fournis par des développeurs externes, souvent américains (voir section 2.2 ci-dessus), il est important que la loi UE sur l'IA formule des règles permettant aux utilisateurs ultérieurs des modèles de base de mieux les comprendre et d'obtenir toutes les informations nécessaires à une mise en œuvre sûre (art. 13 et 53 et suivants du règlement UE sur l'IA). Toutefois, en cas de modification substantielle du modèle initial, il peut arriver que le fournisseur qui a initialement mis le système d'IA sur le marché ne soit plus considéré comme tel - la responsabilité est alors transférée à l'utilisateur du modèle qui a effectué les modifications.

Dans une perspective globale dépassant les réglementations individuelles, la loi sur l'IA crée un système de gouvernance excessivement complexe avec un degré élevé d'insécurité juridique. Comme l'a récemment fait remarquer Kai Zenner, qui a participé aux négociations de la loi sur l'IA, ce mélange de complexité et d'insécurité juridique, avec de nombreux termes juridiques indéterminés, « pourrait augmenter considérablement les coûts de mise en conformité pour les fournisseurs et les utilisateurs de l'IA. En particulier, les PME et les start-ups de l'UE pourraient finir par trouver trop risqué de

[85] Pour une évaluation de l'accord, voir : cep - Centrum für europäische Politik : EU AI Act : A Milestone Met, But Key Challenges Remain in Standardisation and Competition. Le texte final se trouve ici : AM_Ple_LegConsolidated (europa.eu).

[86] Pour un aperçu de la loi sur l'IA visant à aider les entreprises à se conformer au règlement, voir : Compliance AI Act - Feb 24 (wavestone.com). Le résumé présenté ici est basé sur ce guide.

[87] Loi sur l'intelligence artificielle : le Conseil et le Parlement s'efforcent de trouver un accord sur les premières règles en matière d'IA dans le monde - Consilium (europa.eu)

développer ou d'utiliser l'IA ... ou pourraient être obligées de recourir à des audits et des certifications coûteux par des tiers afin d'éviter de lourdes amendes » [traduction personnelle][88]. Pour éviter ce scénario, les PME devraient suivre de près la mise en œuvre concrète de la loi sur l'IA avec sa série de règlements d'application et d'actes délégués qui seront adoptés conformément à l'entrée en vigueur progressive de la loi sur l'IA entre 2025 et 2027. En outre, la loi sur l'IA permet de solliciter des « *sand-boxes* » réglementaires (articles 57 et suivants du règlement européen sur l'IA). Dans celles-ci, les PME peuvent engager un dialogue étroit avec les autorités nationales compétentes et tester et améliorer leurs systèmes d'IA dans des conditions réelles et sans incertitude juridique. Dans tous les cas, lors de l'intégration de l'IA générative, les PME devraient faire appel dès le début à des experts compétents du service juridique et de la conformité, intégrer une gouvernance propre de l'IA dans le processus d'acquisition et, le cas échéant, mettre en place leur propre système de gestion des risques au sens de la loi sur l'IA de l'UE[89].

2.8 Tests internes : évaluer son propre « caractère d'IA »

Avant la mise en œuvre complète d'un système d'IA dans le quotidien de l'entreprise, le modèle linguistique choisi doit être soigneusement testé afin de garantir sa précision, sa fiabilité et son efficacité. Cela implique des tests en conditions réelles et l'évaluation des modèles à l'aide d'indicateurs de performance spécifiques. Des contrôles de qualité doivent également être effectués régulièrement après la mise en œuvre afin de garantir un niveau de performance élevé. En substance, il s'agit de mieux comprendre les propriétés ou le « caractère » du modèle d'IA utilisé - une tâche complexe, rendue plus difficile par le fait que ces propriétés peuvent changer au fil du temps ou avoir des effets secondaires involontaires. Ainsi, le style de communication des personnalités pilotées par l'IA, par exemple dans les discussions avec les clients, est parfois perçu par les humains comme tellement authentique, professionnel et attentionné que des effets secondaires psychologiques peuvent se produire[90].

Les PME disposent déjà de premiers instruments systématiques pour de tels tests internes. Ainsi, des chercheurs ont développé un nouveau cadre logiciel qui facilite la planification d'expériences entre le LLM et l'intégration de LLM dans des expériences avec des sujets humains (comme des collaborateurs ou des clients)[91]. Cette boîte à outils est disponible gratuitement et permet par exemple de réaliser des « dilemmes du prisonnier » - un scénario typique de la théorie des jeux avec de nombreuses applications pratiques dans l'économie - de différents types, dans lesquels l'interaction de plusieurs LLM entre eux ainsi que les interactions homme-machine peuvent être étudiées de manière systématique et empirique. Les résultats montrent que le comportement des modèles linguistiques de l'IA peut changer fortement et parfois de manière surprenante au fil du temps ou lors d'interactions répétées[92], ce qui souligne l'urgence de tels tests avant la mise en ligne publique. Des chercheurs américains ont en outre développé un premier cadre pour une évaluation globale des risques, qui permet d'évaluer le

88 Quelques réflexions personnelles sur le EU AI Act : une fin douce-amère (linkedin.com).
89 Bitkom (2024), L'IA générative dans l'entreprise. Questions juridiques relatives à l'utilisation de l'intelligence artificielle générative dans l'entreprise, p. 19.
90 Voir l'étude : AI Embraces the « Evil » Side of Online Dating (bsi.ag).
91 Voir : GitHub - mrpg/ego : Code for Engel, Grossmann & Ockenfels.
92 Engel, Christoph and Grossmann, Max R. P. and Ockenfels, Axel, Integrating Machine Behavior into Human Subject Experiments : A User-Friendly Toolkit and Illustrations (January 3, 2024). MPI Collective Goods Discussion Paper, No. 2024/1.

risque marginal de la mise à disposition d'un modèle - c'est-à-dire le risque supplémentaire - par rapport au risque des modèles existants ou à la renonciation totale aux technologies d'IA[93].

Lors des tests internes, il ne faut pas seulement considérer le système en soi, mais aussi le contexte de l'application[94]. La littérature met par exemple en garde contre les conséquences d'une confiance excessive dans des modèles erronés lors de conseils juridiques ou médicaux (*automation and confirmation bias*).[95] Afin de mieux comprendre le potentiel de telles dépendances et erreurs dans le propre contexte de l'entreprise, il est judicieux d'étudier empiriquement les modèles d'IA génératifs - et leur utilisation par les collaborateurs - sur une plus longue période[96]. Une étude a par exemple montré que de nombreux modèles peuvent être étonnamment sujets à des erreurs dans des contextes politiques, comme lorsqu'il s'agit de demander des informations sur certaines élections[97]. Il est tout aussi important de prêter attention aux erreurs dans la pratique de l'ingénierie - y compris la construction, la validation, l'intégration et la maintenance - en plus des éventuelles lacunes dans la conception théorique[98]. De tels problèmes lors de la mise en œuvre concrète sont souvent négligés dans le débat actuel sur la sécurité de l'IA, alors que c'est justement là que les problèmes peuvent être directement abordés. Enfin, l'interaction avec les parties prenantes externes à l'entreprise, comme les clients ou les autorités, devrait également être prise en compte. Les systèmes d'IA qui touchent des domaines sensibles et des espaces publics nécessitent une consultation et une validation plus larges.

Sur la base des tests internes avec son propre système d'IA, il convient de développer des protocoles d'interaction et des mécanismes de feedback afin de garantir une collaboration efficace et fluide entre les travailleurs et l'IA[99]. Il s'agit notamment de définir des directives claires sur la manière et le moment où le système d'IA doit solliciter l'intervention de l'homme et inversement. De même, des formations devraient être organisées pour familiariser les collaborateurs avec les outils d'IA générative et leurs spécificités caractéristiques (par exemple par rapport aux fournisseurs populaires comme OpenAI), tout en leur donnant la possibilité de donner un feedback constructif.

2.9 Durabilité et énergie : prendre en compte les coûts de mise à l'échelle de l'IA

Comme le développement de l'IA générative se caractérise par une forte consommation d'énergie, cette technologie a récemment été associée à une crise écologique dans le secteur technologique. L'aveu de la crise énergétique imminente par le CEO d'OpenAI, Sam Altman, lors du Forum économique mondial de cette année, illustre bien cette tendance à aborder la dimension écologique de l'IA - au niveau politique et entrepreneurial[100]. Cette dimension ne se limite pas à l'énergie ; les systèmes d'IA générative nécessitent également de grandes quantités d'eau douce à des fins de refroidissement, et

93 Voir : On the Societal Impact of Open Foundation Models (stanford.edu).
94 Dobbe (2022), System Safety and Artificial Intelligence, 2202.09292.pdf (arxiv.org).
95 O'Neil, C. (2016), Weapons of Math Destruction : How Big Data Increases Inequality and Threatens Democracy. Crown ; Logg, J. M., Minson, J. A., & Moore, D. A. (2019), Algorithm appreciation : People prefer algorithmic to human judgment. Organizational Behavior and Human Decision Processes, 151, 90-103 ; Goddard, K., Roudsari, A., & Wyatt, J. C. (2012), Automation bias : a systematic review of frequency, effect mediators, and mitigators. Journal of the American Medical Informatics Association, 19(1), 121-127.
96 Narayanan et Kapoor (2024), AI safety is not a model property (aisnakeoil.com).
97 A la recherche d'informations électorales fiables ? Don't Trust AI (proofnews.org).
98 Raji and Dobbe (2022), Concrete Problems in AI Safety, Revisited, 2401.10899.pdf (arxiv.org).
99 Votre organisation n'est pas conçue pour travailler avec GenAI (hbr.org).
100 Khalaf (2024), The environmental cost of AI (ft.com).

les entreprises technologiques de premier plan enregistrent des pics de consommation importants pour le développement et l'entraînement de leurs modèles[101]. Ces tendances suscitent des inquiétudes quant à la durabilité de la croissance rapide de l'IA générative, puisque les besoins en ressources prévus pourraient atteindre ceux d'une nation entière dans un avenir proche. En raison de cette situation, les experts appellent désormais au développement de systèmes d'IA plus durables, à l'établissement de rapports environnementaux rigoureux et au passage à des sources d'énergie renouvelables, ainsi qu'à des mesures législatives[102].

Du point de vue des entreprises individuelles, ces considérations éthiques prennent de plus en plus d'importance lors du déploiement de l'IA, tant vis-à-vis des employés que des clients. Les préoccupations éthiques font référence aux principes moraux et aux valeurs qui guident le comportement humain et qui sont désormais de plus en plus intégrés dans l'utilisation des systèmes d'IA afin de proposer des produits d'IA axés sur la durabilité[103]. Par exemple, l'utilisation de systèmes de maisons intelligentes soulève des questions d'accès et d'utilisation des données qui devraient être mises en balance avec les possibilités d'économie d'énergie. Bien qu'il existe de nombreux concepts et recommandations pour l'utilisation éthique des systèmes d'IA, tels que les lignes directrices éthiques de l'UE pour une IA de confiance et les recommandations éthiques mondiales sur l'IA de l'UNESCO, de l'OCDE et de l'Institute of Electrical and Electronics Engineers, le développement de réglementations concrètes ne démarre que progressivement[104]. Le concept de conception sensible à la valeur (Value Sensitive Design, VSD) souligne que la technologie n'est pas neutre, mais qu'elle est imprégnée de certaines valeurs et normes, et vise à les intégrer dans le processus de conception technologique[105]. Le développement de l'IA sur la base de la VSD nécessite une réflexion critique sur les valeurs et les besoins de toutes les parties prenantes et le développement de systèmes d'IA qui respectent ces valeurs. Pour mettre en œuvre ce concept au niveau des PME, on peut penser par exemple à la boîte à outils open-source « AI Fairness 360 » d'IBM, qui propose des instruments d'évaluation des applications d'IA en termes d'équité et de justice[106]. De manière générale, la littérature s'accorde à dire que les systèmes d'IA responsables devraient répondre à des normes éthiques telles que l'équité, l'explicabilité et la transparence[107] et concilier la durabilité environnementale avec la rentabilité et la responsabilité sociale. Les PME doivent donc mettre l'accent sur la vérification des données d'entraînement pour garantir l'exactitude des outils d'IA et prendre des mesures précoces pour éviter les biais et garantir l'explicabilité des décisions automatisées[108]. Le suivi de l'impact de l'IA sur les objectifs de durabilité et le respect des futures normes industrielles sont essentiels pour mettre en place des outils d'IA de manière durable dans le domaine de l'approvisionnement et d'autres domaines d'application.

Lors de la sélection et de la mise en œuvre des technologies de modèles de langage, les PME devraient donc tenir pleinement compte de leur durabilité et de leur impact sur l'environnement, à la fois pour

---

101 [2304.03271] Making AI Less « Thirsty » : Uncovering and Addressing the Secret Water Footprint of AI Models (arxiv.org).
102 d41586-024-00478-x.pdf (nature.com).
103 Questions éthiques d'une IA orientée vers la durabilité - Année scientifique 2019 : Intelligence artificielle.
104 Cas (2023), « Intelligence artificielle » et durabilité sociale. Principes éthiques pour les technologies d'IA en tant que solutions pour la réduction de la pauvreté et des inégalités ?, Magazin erwachsenenbildung.at 49, p. 51-60.
105 Voir : IA et éthique : la durabilité comme facteur central (suso.academy).
106 Voir l'aperçu sur : IBM Research Trusted AI, AI Fairness 360 (ibm.com).
107 Voir : Équité, explicabilité et transparence des applications de l'IA dans le domaine de la sécurité - une mission impossible ? - Union Humaniste (humanistische-union.de).
108 Ceci et d'autres questions chez : IA et achat durable : La morale de la machine | Sustainability | Haufe.

des raisons d'éthique, de réputation et de pression politique[109]. et en raison des coûts croissants liés à leur mise à l'échelle. Tant les start-ups que les grandes entreprises sont actuellement confrontées à des coûts de déploiement croissants lorsqu'elles passent d'une preuve de concept pour quelques utilisateurs à un déploiement à grande échelle de la technologie de modèles de langage[110]. Il est important de comprendre que la structure des coûts des logiciels pilotés par l'IA est très différente de celle des logiciels traditionnels[111]. La microarchitecture des puces et l'architecture du système jouent un rôle crucial dans l'évolutivité de la technologie de modèles de langage. C'est pourquoi l'optimisation de l'infrastructure de l'IA est essentielle pour pouvoir utiliser l'IA générative de manière durable. Dans un avenir proche, il y aura probablement des normes qui prendront en compte non seulement l'évaluation de la consommation d'énergie directe et des émissions de CO2 des technologies, mais aussi les aspects écologiques tout au long de la chaîne d'approvisionnement de l'IA.

Ces derniers mois, plusieurs études empiriques ont été publiées, qui tentent de mesurer non seulement les émissions, mais aussi d'autres impacts environnementaux et sociaux de l'IA générative, et de développer des normes pour les rapports à ce sujet[112]. Elles peuvent être consultées comme premier « benchmark ». Toutefois, la majeure partie de cette littérature ne traite que des exigences énergétiques liées à l'*entraînement* de l'IA. Pour les PME qui intègrent des modèles déjà entraînés dans leurs activités, la recherche de HuggingFace, qui a quantifié les besoins énergétiques de l'*utilisation* de l'IA générative, est particulièrement pertinente[113]. Ces besoins sont plus importants qu'on ne le pense généralement et peuvent rapidement s'accumuler, selon le modèle d'entreprise et le cas d'application. Pour les PME, les trois conclusions suivantes de cette recherche sont essentielles : 1. les tâches impliquant la prédiction de catégories sont moins gourmandes en énergie que les tâches génératives. En d'autres termes, les applications d'IA les plus gourmandes en énergie et en CO2 sont celles qui génèrent de nouveaux contenus, notamment la génération d'images et (dans une moindre mesure) la génération de textes ; 2. Même si l'apprentissage de l'IA reste de plusieurs ordres de grandeur plus gourmand en énergie et en CO2 que l'application individuelle de l'IA (ce que l'on appelle l'inférence), l'utilisation généralisée de modèles d'IA génératifs permet d'atteindre rapidement la parité en matière de consommation d'énergie pour de nombreux modèles courants ; 3. l'utilisation de modèles polyvalents (comme ChatGPT) est plus gourmande en énergie pour la classification de textes et la réponse à des questions que les modèles spécifiques à une tâche.

2.10 Tirer profit des expériences internes des utilisateurs et de la sagesse des foules externe

Enfin, les PME devraient surveiller et évaluer régulièrement les performances des technologies linguistiques mises en œuvre. Il est communément admis qu'une réflexion et une évaluation continues aident à réagir aux évolutions du marché ou des technologies, à adapter les stratégies en conséquence et à identifier les domaines d'amélioration. Dans le contexte de l'IA générative et des PME dont il est

[109] Voir : Measuring AI's Environmental Impacts Requires Empirical Research and Standards | TechPolicy.Press.
[110] Voir : Intelligence artificielle : Microsoft, Google, Nvidia Win as Computing Costs Surge - Bloomberg.
[111] Cet argument est basé sur l'analyse de : Google AI Infrastructure Supremacy : Systems Matter More Than Microarchitecture (semianalysis.com).
[112] Voir par exemple : Luccioni et al. (2022), [2211.02001] Estimating the Carbon Footprint of BLOOM, a 176B Parameter Language Model (arxiv.org) ; AI is harming our planet : addressing AI's staggering energy cost (2023 update) (numenta.com).
[113] Luccioni et al. (2023), [2311.16863] Power Hungry Processing : Watts Driving the Cost of AI Deployment ? (arxiv.org).

question ici, deux points concrets sont pertinents : la surveillance de l'expérience d'utilisation interne par la conception de modèles « *human-in-the-loop* » afin d'éviter une « surréalisation de l'IA » ; et l'ajout d'une intelligence collective externe (« *crowd wisdom* ») par le biais des réseaux sociaux, du web et de la littérature pré-imprimée afin d'adapter en permanence la technologie linguistique choisie à l'état actuel de l'application et de la sécurité. Qu'est-ce que cela signifie exactement ?

Au niveau organisationnel, l'introduction d'outils vocaux d'IA faciles à utiliser et intuitifs, comme ceux de type « ChatGPT », peut poser des problèmes à long terme, car les personnes ont tendance à faire moins d'efforts et à être moins attentives à mesure que la qualité de l'IA augmente. Une expérience de terrain menée avec des recruteurs professionnels qui examinaient des CV a révélé que ceux qui travaillaient avec des outils d'IA de moindre qualité donnaient des évaluations plus précises, car ils faisaient plus d'efforts et interagissaient plus efficacement avec l'IA[114]. De plus, les gens acceptent souvent la décision recommandée par un système d'IA, même si elle est erronée - un problème que la littérature qualifie d'*overreliance* de l'IA, ou de « confiance aveugle » dans l'IA. L'interaction entre l'homme et la machine est difficile à évaluer ex ante, car les humains ne réagissent pas toujours de manière rationnelle aux recommandations d'un ordinateur[115]. Ainsi, lors d'une expérience, des participants ont continué à suivre les conseils délibérément mal programmés de l'algorithme, même lorsqu'ils auraient dû mieux savoir depuis longtemps[116]. Certains chercheurs espèrent réduire la confiance aveugle dans les systèmes d'IA en les obligeant à expliquer leurs décisions. Mais les tests montrent que de telles explications ne font qu'augmenter la probabilité que les gens acceptent la recommandation de l'IA, qu'elle soit correcte ou non[117]. Une solution qui fonctionne, du moins expérimentalement, consiste à encourager les gens à se confronter à ces explications, y compris sur le plan cognitif[118]. L'expert en organisation Gianni Giacomelli a noté, en ce qui concerne le développement des modèles *« human-in-the-loop* » à l'ère de l'IA générative, que « la capacité à exploiter les nouvelles possibilités en transformant nos processus organisationnels et commerciaux et en développant les pratiques humaines qui y sont associées sera probablement aussi importante que le travail sur le côté technologique de l'IA [propre traduction] »[119]. Selon une méta-analyse récemment publiée par Microsoft, qui regroupe une soixantaine d'études sur le sujet, les mesures les plus importantes pour réduire l'« *overreliance* » de l'IA sont la fourniture d'un feedback en temps réel, des explications efficaces pour favoriser la confiance et la possibilité pour les utilisateurs de contrôler eux-mêmes le rythme et l'utilisation des recommandations de l'IA[120].

Outre ce pilier « interne » pour l'apprentissage continu avec et sur la technologie linguistique de l'IA, il faudrait en outre utiliser des offres d'information « externes ». Les PME en particulier, avec leurs équipes et leurs données parfois limitées, pourraient rapidement se heurter à des limites organisationnelles en raison du développement rapide de la technologie et de problèmes inattendus. L'utilisation d'un feedback continu de la part d'utilisateurs externes de la technologie de modèles de langage mise en œuvre ou d'utilisateurs de systèmes d'IA similaires est donc décisive pour son efficacité (et son

[114] Dell'Acqua, F. (2022), Falling+Asleep+at+the+Wheel+-+Fabrizio+DellAcqua.pdf (squarespace.com).
[115] Algorithmic Risk Assessment in the Hands of Humans (iza.org).
[116] Biermann, Jan et Horton, John J. et Walter, Johannes, Algorithmic Advice as a Credence Good (2022). ZEW - Centre for European Economic Research Discussion Paper No. 22-071, http://dx.doi.org/10.2139/ssrn.4326911.
[117] [2006.14779] L'effet des explications de l'IA sur la performance des équipes complémentaires (arxiv.org).
[118] [2212.06823] Des explications peuvent réduire la surreprésentation des systèmes AI lors de la prise de décision (arxiv.org).
[119] Giacomelli, G. (2024), Au-delà de « l'humain dans la boucle » : l'IA fiable dans les workflows d'entreprise (linkedin.com).
[120] Passi, S. et Vorvoreanu, M. (2022), Overreliance on AI Literature Review (microsoft.com).

acceptation) à long terme. Il existe de nombreuses auditions en ligne spécialisées qui s'occupent quotidiennement de tester les vulnérabilités LLM (« *Red Teaming* ») et qui identifient souvent les problèmes plus rapidement et mieux que les experts internes. La littérature académique ne peut plus suivre ce rythme depuis longtemps ; des découvertes importantes se trouvent en pré-impression sur ArXiv et sont discutées sur des plateformes de médias sociaux comme Twitter et Reddit. Les PME devraient observer ces discours et, le cas échéant, les modérer activement afin de garantir une application optimale et de pouvoir mettre à jour rapidement leurs propres systèmes.

3 Conclusion : développer des options stratégiques, saisir concrètement les opportunités

Le développement rapide de grands modèles linguistiques tels que ChatGPT constitue un défi majeur pour l'Europe. Compte tenu de l'application jusqu'ici insuffisante dans l'économie, il est urgent de passer des « projets phares » et de l'aversion au risque à une mise en œuvre plus pragmatique et généralisée de la technologie de modèles de langage. Malgré l'intention stratégiquement judicieuse de l'Union européenne de sécuriser la chaîne de valeur de l'IA à long terme avec des initiatives telles que les espaces de données, les subventions pour les usines de puces et les supercalculateurs d'IA, la dynamique de la technologie d'IA ne permet pas de retarder davantage son application. Le recours à des modèles commerciaux et à source ouverte, en particulier pour les petites et moyennes entreprises (PME), est essentiel pour maintenir la compétitivité et exploiter le potentiel de l'IA en matière d'automatisation des tâches à forte intensité de connaissances et de promotion de l'innovation. Dans ce contexte, cet input du cep a décrit dix facteurs que les PME devraient prendre en compte lors de la mise en œuvre de la technologie linguistique de l'IA. Ceux-ci peuvent être résumés comme suit :

1. **Effectuer une analyse des besoins et une planification stratégique** : Les PME doivent d'abord effectuer une analyse approfondie afin de comprendre conceptuellement comment l'IA générative peut améliorer leurs processus internes et externes. Un objectif clair est essentiel pour la mise en œuvre.
2. **Réduire les dépendances stratégiques** : Le choix entre des services basés sur le cloud et des installations privées sur site a des conséquences directes sur l'évolutivité, la flexibilité et les engagements financiers à long terme. L'utilisation de modèles de base ouverts peut aider à minimiser les dépendances stratégiques.
3. **Personnaliser les modèles grâce au fine tuning et à la RAG** : Les PME peuvent optimiser et différencier leurs applications d'IA en adaptant les modèles pré-entraînés à des cas d'utilisation spécifiques grâce au fine tuning et à la génération augmentée de récupération.
4. **Développer des compétences internes en PNL** : La compréhension du fonctionnement et des limites des modèles linguistiques ainsi que le développement de compétences dans le domaine du prompt design et du design en ligne sont essentiels pour exploiter le potentiel de l'IA par rapport aux concurrents.
5. **Adapter la tolérance aux erreurs de l'IA** : La tendance des modèles d'IA à « halluciner » et les risques qui en découlent nécessitent d'adapter la technologie à la tolérance aux erreurs de l'entreprise et de développer des contre-mesures efficaces.
6. **Intégrer des agents d'IA** : L'ajout d'agents IA dans les processus, les produits et les services permet d'améliorer l'efficacité des flux de travail, mais comporte également des risques qui doivent être réduits par un examen et des tests minutieux.

7. **Suivre le cadre juridique** : La protection des données, les droits d'auteur et la législation sur l'IA doivent être pris en compte lors de la mise en œuvre de l'IA générative afin de minimiser les risques juridiques et de pouvoir remplir les obligations de transparence. La loi sur l'IA peut être une opportunité pour les PME d'obtenir plus de transparence sur les modèles de boîte noire sous-jacents et leurs données d'entraînement.
8. **Effectuer des tests internes** : Avant la mise en œuvre complète, les modèles linguistiques et leur « caractère » doivent être testés de manière approfondie afin de garantir leur précision et leur fiabilité et de détecter rapidement les effets secondaires indésirables.
9. **Penser à la durabilité et à l'efficacité énergétique** : l'impact écologique du déploiement à grande échelle des technologies d'IA doit être pris en compte dès le départ afin de garantir un déploiement durable et rentable après la mise à l'échelle.
10. **Utiliser les mécanismes de feedback** : L'évaluation continue de la technologie de modèles de langage par les propres expériences des utilisateurs ainsi que par la sagesse des foules externe est décisive pour maintenir la technologie à jour et identifier les vecteurs d'attaque à un stade précoce.

Dans l'ensemble, ces dix facteurs constituent une base conceptuelle pour les PME afin de développer une stratégie interne d'IA et d'exploiter efficacement le potentiel dans le domaine des modèles linguistiques tout en identifiant les risques techniques et les défis stratégiques. La politique en Europe devrait contribuer à ce que la technologie de modèles de langage puisse être mise en œuvre rapidement, mais en toute sécurité, dans l'environnement domestique des entreprises en renforçant la sécurité juridique (par exemple en adoptant rapidement des lignes directrices en matière de conformité après l'adoption de la loi européenne sur l'IA) et en fournissant davantage de subventions et de points de contact. Parallèlement, les entreprises devraient surmonter leur scepticisme, tirer parti des avancées méthodologiques et des dernières découvertes en matière de recherche, et réduire leur dépendance vis-à-vis des fournisseurs étrangers afin de minimiser les coûts de conversion ultérieurs. Cette transition vers une application à grande échelle de l'IA générative est essentielle non seulement pour accroître l'efficacité et l'innovation dans tous les secteurs, mais aussi pour assurer la résilience de l'Europe en période d'instabilité. Si elle est utilisée à grande échelle tout en gérant soigneusement les risques, la technologie de modèles de langage a le potentiel de renforcer l'Europe sur le marché mondial et d'accélérer la transformation vers un ordre économique numérique et durable.

Centrum für Europäische Politik
FREIBURG | BERLIN

Auteur :

Dr. Anselm Küsters, LL.M., chef du département Numérisation et nouvelles technologies
kuesters@cep.eu

Traduit depuis l'allemand par Thomas Plancq, chargé de communication

Centre de politique européenne FREIBURG | BERLIN
Kaiser-Joseph-Straße 266 | D-79098 Fribourg
Schiffbauerdamm 40 Salle 4315 | D-10117 Berlin
Tél. + 49 761 38693-0