Note de synthèse mensuelle sur l'IA Générative – Mai 2024

De : **AI Builders**

Abstract

Dans cette édition du mois de mai de notre note de synthèse mensuelle sur l'Intelligence Artificielle Générative, voici les 5 informations majeures à retenir :

- Microsoft réalise un **investissement massif en France de 4 milliards d'euros** pour le développement de **solutions, de talents et d'infrastructures** en lien avec l'IA Générative, avec pour objectif de favoriser l'émergence de **solutions plus souveraines** *(Rubrique : Les chiffres clés du mois)*
- **« H » lève 220 millions de dollars** pour développer une **IA Générative plus déterministe** que les modèles de langage traditionnels, ouvrant la voie à de **nouveaux cas d'usage spécifiques à l'entreprise** *(Rubrique : Les chiffres clés du mois)*
- **UPS** parvient à **réduire de 50% le temps de traitement** des agents de son centre de relation client grâce à un outil interne basé sur l'IA Générative et cherche à **généraliser cet usage** à l'ensemble de l'entreprise *(Rubrique : Les Tops / Flops du mois)*
- Le **Groupe Rocher** intègre dans ses assistants basés sur de l'IA Générative des modules **d'auto-formation à destination de ses collaborateurs afin de favoriser son adoption** *(Rubrique : Les 3 cas d'usage marquants du mois)*
- **OpenAI** présente son nouveau **modèle multimodal GPT-4o**, **moins cher et plus performant** que ses concurrents directs, favorisant ainsi de **nouveaux cas d'usage conversationnels**. *(Rubrique : Synthèse des évolutions technologiques du mois)*

I. Les chiffres clés du mois

4 milliards d'euros, c'est le montant investi par Microsoft pour développer les filières IA et IA Générative en France[1].

Microsoft a annoncé en mai 2024, en marge de l'événement Choose France, sa volonté d'engager de **nombreux investissements en France d'ici à 2027**. L'ambition de la firme américaine est triple : **renforcer l'infrastructure** avec l'extension de ses data centers existants et la construction d'un centre de données de nouvelle génération, **former et sensibiliser** un million de particuliers et professionnels à l'IA Générative et enfin, **financer et accompagner 2 500 startups** basées en France via le programme « Microsoft GenAI Studio ».

À retenir : l'attractivité de la France se renforce avec une accélération notable des investissements étrangers ces derniers mois en matière d'IA Générative, à l'instar des futurs financements d'une trentaine d'acteurs étrangers dont Nvidia et Salesforce dans le prochain tour de table de 600 millions d'euros de la startup française Mistral AI. Ces investissements

[1] https://news.microsoft.com/fr-fr/2024/05/13/microsoft-annonce-son-investissement-en-france-le-plus-important-a-ce-jour-pour-accelerer-ladoption-de-lia-les-competences-et-linnovation/

seront cruciaux pour développer et pérenniser les usages en entreprise, grâce au renforcement d'une infrastructure et d'offres souveraines.

220 millions de dollars levés par « H » pour développer des modèles de fondation de nouvelle génération pour les entreprises[2].

Fondée par d'anciens ingénieurs de Google DeepMind, la startup parisienne « H » a levé un tour de financement record en mai 2024. « H » ambitionne de fournir aux entreprises du monde entier une **nouvelle génération d'intelligences artificielles génératives**, capables de **planifier et exécuter des tâches complexes**, tout en garantissant **transparence**, **explicabilité** et **respect des données personnelles**. Pour ce faire, H mise sur le développement de « LAM » (Large Action Models) basées sur des systèmes intelligents (ou « agents », cf. concept vulgarisé du mois) conçus pour atteindre un **niveau de compréhension supérieur aux LLM actuels** et de **faciliter la prise de décision**.

À retenir : bien que les IA Génératives actuelles aient démontré une réelle valeur ajoutée, celles-ci demeurent difficiles à maîtriser en raison de la nature parfois aléatoire de leurs réponses. L'introduction des agents de type Reasonning and Acting (ReAct) dans les outils d'IA générative ainsi que dans Copilot pour sa personnalisation, laisse penser que les LLM orchestrateurs aujourd'hui génériques, seront remplacés par des LLM spécialisés, d'où leur nom donné par « H », Large Action Models.

+30% d'émissions carbone pour Microsoft en 2023 à cause de l'IA Générative[3].

Si l'IA Générative a permis à Microsoft de dynamiser significativement sa performance financière, son **bilan extra-financier** a quant à lui suivi une trajectoire inverse. En 2023, **les émissions de carbone de l'entreprise ont en effet augmenté d'un tiers**, compromettant ainsi l'objectif de neutralité carbone visé par la firme américaine d'ici à 2030. Cette hausse est principalement due aux **investissements massifs** consentis dans le **développement des data centers** nécessaires à ses services d'IA Générative, lesquels sont particulièrement **énergivores**.

À retenir : l'impact carbone de l'IA Générative demeure un défi majeur. L'augmentation de l'empreinte carbone des fournisseurs de solutions d'IA Générative peut compromettre les efforts environnementaux de leurs clients. Toutefois, des alternatives émergent, offrant des modèles de langage à la fois performants et moins consommateurs de ressources. Ainsi, les petits modèles de langage ultra spécialisés ou « SLM » (comptant jusqu'à 7 milliards de paramètres), voire les très petits modèles de langage ou « STLM » (jusqu'à 100 millions de paramètres) se révèlent être des solutions suffisamment prometteuses pour certains cas d'usage spécifiques.

II. Les Tops / Flops du mois

Le Top : UPS réduit de 50% le temps de traitement des agents de son centre de relation client grâce à l'IA Générative[4].

Le géant américain du transport de marchandises UPS a lancé en juillet dernier un projet interne basé sur les grands modèles de langage visant à **automatiser la gestion de la relation client**. Baptisé **MeRa** (Message Response Automation), ce projet vise à l'**amélioration de l'efficacité opérationnelle** du centre de relation client d'UPS en **réduisant le temps de traitement des 52 000 demandes entrantes** reçues quotidiennement par e-mail. Les premiers résultats de MeRa s'avèrent concluants, avec une **réduction de moitié du temps passé par les agents**. UPS compte par la suite généraliser cet

[2] https://capitalfinance.lesechos.fr/deals/capital-risque/ia-the-h-company-leve-220-m-2096373

[3] https://www.novethic.fr/economie-et-social/transformation-de-leconomie/microsoft-ses-emissions-de-co2-bondissent-de-30-en-raison-de-lintelligence-artificielle

[4] https://www.cio.com/article/2096052/ups-delivers-customer-wins-with-generative-ai.html

outil à d'autres fonctions au sein de l'entreprise, notamment la vente, les ressources humaines et la finance.

À retenir : l'augmentation de l'efficacité opérationnelle des centres de relation client grâce à l'IA Générative devient un cas d'usage standard, comme en témoignent les pilotes lancés avec succès par UPS (-50% de temps de traitement) ou Klarna (-80%, équivalent à 700 ETP, cf. la note de synthèse du mois de mars). L'industrialisation de ces usages de relation client implique pour les entreprises des impacts RH majeurs. Nous assistons à une profonde prise de conscience des directions des ressources humaines quant à la nécessité d'anticiper l'impact de l'IA.

Le Flop : le fondateur de Google supplie un employé de ne pas rejoindre OpenAI[5].

L'accès aux talents est au cœur de la bataille que se livrent les géants de la tech dans la course au développement de l'IA Générative. Alors qu'Elon Musk qualifiait la situation actuelle de « plus folle course aux talents de tous les temps », il semblerait que **Google rencontre des difficultés à retenir certains de ses employés clés.** Le co-fondateur de Google, Sergey Brin, aurait ainsi personnellement contacté un ingénieur de Google DeepMind, la filiale du géant américain spécialisée dans le deep learning, pour **tenter – sans succès – de le convaincre de ne pas rejoindre OpenAI**.

À retenir : la pénurie de talents exacerbée par la demande croissante en matière d'IA Générative est une réalité indéniable. Cependant, pour nos entreprises, les data scientists, grâce à leur formation généraliste et polyvalente, sont tout à fait à même de devenir d'excellents spécialistes de l'IA Générative. Il convient alors de leur permettre de suivre intégralement la courbe d'apprentissage, sans brûler d'étapes, à travers la maîtrise de divers projets d'IA Générative en interne. La mise en œuvre en interne d'équipes de data scientists spécialisées dans l'IA Générative devient ainsi une nécessitée.

III. Les 3 cas d'usages marquants du mois

Santé : Moderna injecte l'IA Générative dans tous ses métiers en utilisant les technologies d'Open AI[6].

Le spécialiste des biotechnologies **Moderna a identifié l'IA Générative comme un catalyseur essentiel** pour atteindre son ambition de lancer 15 nouveaux produits d'ici à 2030. Pour ce faire, Moderna a développé une série d'outils internes basés sur l'IA Générative, visant à **automatiser les tâches, améliorer la productivité des employés et accélérer la recherche et le développement**. Parmi ces outils figure **GPT Dose ID**, qui utilise la fonction d'analyse de données d'OpenAI « Advanced Data Analytics » pour mieux évaluer la dose optimale de vaccin lors des études cliniques et générer des graphiques informatifs illustrant les résultats clés. Pour **garantir une intégration réussie**, Moderna a opté pour un **déploiement progressif** : une première cohorte de 100 « AI power users » a été identifiée via un hackathon interne, puis l'usage a été étendu à plus de 2 000 participants. Cette stratégie de déploiement par vagues a porté ses fruits, avec un **taux d'adoption de l'outil atteignant 80%.**

À retenir : ce cas d'usage illustre l'importance de développer une méthodologie de déploiement maîtrisée, progressive et adaptée, pour favoriser l'industrialisation des outils d'IA Générative et garantir un taux d'adoption élevé. Moderna démontre en effet qu'il est crucial de commencer par des utilisateurs enthousiastes, puis d'optimiser la solution avant de la déployer à plus grande échelle. Un déploiement trop rapide, auprès d'utilisateurs non-ambassadeurs, peut entraîner un effet déceptif, compromettant irrémédiablement le taux d'adoption de la solution.

[5] https://www.msn.com/en-in/money/technology/sergey-brin-personally-called-a-google-employee-to-convince-them-to-turn-down-a-job-at-openai-report/ar-BB1kDAzM
[6] https://openai.com/index/moderna/

Grande distribution : Le Groupe Rocher mise sur des modules d'auto-formation intégrés aux outils d'IA Générative[7].

Après avoir identifié des **cas d'usages générateurs de valeur** pour diverses fonctions du groupe (marketing, IT, ressources humaines et communication), le Groupe Rocher a mis en place un **outil d'auto-formation au prompting** pour accompagner l'utilisation d'une première cohorte de 600 utilisateurs internes. Pour ce faire, le groupe propose à ses collaborateurs un **outil pédagogique intégré à ChatGPT et à Copilot for Microsoft 365** pour les guider vers une meilleure utilisation de l'IA Générative et pour en maximiser la valeur.

À retenir : la formation au prompting est un enjeu décisif en entreprise, tant la maîtrise de cette compétence est corrélée avec la qualité des résultats fournis par un LLM (cf. concept vulgarisé du mois d'avril). Alors que la plupart des entreprises tentent d'accompagner leurs collaborateurs dans un mode de formation "pull" au travers de séminaires, e-learnings, etc. le Groupe Rocher mise sur une formation "push" visant à délivrer le bon accompagnement au bon moment, permettant ainsi une formation en continu aux utilisateurs.

Services financiers : Visa s'attaque aux transactions frauduleuses grâce à l'IA Générative[8].

Le leader américain des systèmes de paiement a dévoilé en mai 2024 son outil **Visa Account Attack Intelligence** (VAAI) basé sur l'IA Générative capable d'**identifier en 20 millisecondes les transactions suspectes en ligne**. Entraîné sur 15 milliards de transactions passées, VAAI limite les vulnérabilités à chaque paiement enregistré leur attribuant un score de risque. L'outil enregistre une **performance supérieure aux précédents modèles de prédiction des risques utilisés par Visa**, avec une réduction mesurée de **85%** du nombre de faux positifs. VAAI devrait ainsi permettre de réduire le **coût opérationnel lié à la fraude**, estimé à 1,1 milliard de dollars chaque année. Dans le sillage de Visa, son principal concurrent **Mastercard**, a également annoncé ce mois-ci le déploiement d'une technologie interne fondée sur l'IA Générative visant à doubler la détection de cartes compromises.

À retenir : l'IA Générative permet un accès simplifié à des sources externes de données. Cet enrichissement substantiel de données améliore notoirement les modèles prédictifs. Dans le secteur bancaire comme dans d'autres domaines, la gestion des risques s'en trouve ainsi grandement améliorée et gagne en précision.

IV. Synthèse des évolutions technologiques du mois dans le domaine de l'IA Générative

1) **OpenAI** a récemment dévoilé son **nouveau modèle multimodal**, améliorant l'accessibilité de ses outils[9].

Lors de sa dernière conférence officielle, OpenAI a présenté son nouveau modèle multimodal **GPT-4o**. Il intègre désormais le **son** en plus du texte et de l'image. En ajoutant à son nouvel outil une **rapidité d'exécution améliorée**, OpenAI franchit une étape supplémentaire vers une **IA conversationnelle toujours plus anthropomorphique.** Son nouveau modèle est présenté comme étant plus efficace et moins cher que les précédents, avec la particularité que l'utilisateur peut interrompre la réponse du modèle à voix haute, ce qui pourrait présenter une **amélioration notable des callbots** utilisés par les services client. Après avoir publié ses services en ligne, OpenAI propose une **application** tout d'abord portée dans l'univers des devices Apple en s'intégrant à l'iOS et bientôt dévoilée dans le Windows Store.

[7] https://www.cio-online.com/actualites/lire-le-groupe-rocher-deploie-la-formation-a-l-ia-generative-au-coeur-du-prompt-15640.html
[8] https://investor.visa.com/news/news-details/2024/Visa-Announces-Generative-AI-Powered-Fraud-Solution-to-Combat-Account-Attacks/default.aspx
[9] https://openai.com/index/gpt-4o-and-more-tools-to-chatgpt-free/

2) **Google intègre de l'IA Générative** à ses outils historiques[10].

Pour sa part, Google a lancé deux versions améliorées de son modèle phare, **Gemini 1.5**, pour tenter de concurrencer la firme de Sam Altman. Ces versions proposent des **fenêtres de contexte plus larges** pouvant même intégrer de longues vidéos. Google améliore son assistant Google Workspace en utilisant Gemini 1.5 et ses nouvelles performances.

À l'image de Bing, Google **prend le virage du prescriptif** dans son moteur de recherche, en proposant des **réponses générées par des modèles de langages** en complément d'une liste de liens hypertextes pertinents. Cette combinaison de typologies de réponses pourrait remettre en question l'efficacité et la valeur perçue des liens sponsorisés au cœur des stratégies SEA et du business de Google.

3) **Microsoft facilite la création de copilotes personnalisés** au poste de travail[11].

Le 21 mai s'est ouverte la conférence de Microsoft avec de nombreuses annonces visant à **intégrer l'IA Générative dans le poste de travail**. À l'aide de l'outil Copilot Studio il est désormais possible de personnaliser des Copilots en fonction du besoin métier. Cet outil offre différents niveaux de personnalisation **dépendant des compétences de l'équipe IT**. En effet, une personnalisation de base peut être réalisée avec du low code et approfondie grâce à la plateforme Azure AI Studio. Copilot Studio propose un **workflow complet** permettant de suivre le cycle de vie des Copilots. Le déploiement des Copilots peut être personnalisé et piloté grâce à un ensemble de statistiques permettant de **suivre l'utilisation du copilote pour les différents métiers**. La personnalisation de Copilot est rendue possible par la **mise à disposition de LLM et SLM verticaux** dans une marketplace spécifique à l'entreprise intégrée à Copilot Studio. Cette capacité de personnalisation et d'intégration de multiples modèles et outils (consultation du web, connexion à des bases de données de suite métier,…) est rendue possible par l'adjonction d'agents de type ReAct (Reasonning and Acting) nativement dans les Copilots.

4) **Microsoft** lance un **PC** intégrant toute sa **suite produit liée à l'IA Générative**[12].

Microsoft s'allie à Intel afin d'utiliser la **dernière génération de microprocesseurs** « Intel Lake Meteor » qui intègre un processeur de type NPU (Neural Processing Unit) sur lequel il fait tourner localement son modèle **Phi-3** adapté à cet environnement et dénommé pour l'occasion Phi-Silica.

Une révolution dans le monde l'IA Générative où l'accès aux modèles était auparavant réalisé dans le cloud uniquement de manière connectée, un avant-goût du *AI at the Edge*. Ces **nouvelles capacités embarquées** dans les PCs permettent d'ores et déjà de faire tourner Copilot mais aussi de nouvelles applications. Recall en l'occurrence est certainement l'une des premières d'une longue série. Elle permet de **recréer une mémoire visuelle** avec des captures d'écran effectuées toutes les 5 secondes et d'avoir accès à toutes les tâches réalisées sur le PC au cours des 1 à 2 derniers mois.

[10] https://blog.google/intl/fr-fr/nouvelles-de-lentreprise/google-io-2024-sundar-pichai-io-pour-nouvelle-generation/

[11] https://news.microsoft.com/fr-fr/2024/05/21/whats-next-microsoft-build-poursuit-levolution-et-lexpansion-des-outils-dia-au-service-des-developpeurs/

[12] https://blogs.microsoft.com/blog/2024/05/20/introducing-copilot-pcs/

V. Le concept technique vulgarisé du mois

Il est maintenant possible de créer des **applications répondant à des besoins métiers utilisant plusieurs LLM spécialisés et des outils** de différents types (Recherche sur Internet, connexion à une base de données métiers type CRM, ...). En effet, il est apparu que certains LLM de plus de 100B, appelés LLM orchestrateurs, étaient dotés de capacités d'ordonnancement permettant d'analyser une requête et d'en **déduire les LLMs et/ou les outils à utiliser** pour apporter une réponse. Ces agents de type **ReAct** (Reasonning and Acting) décomposent la requête et réalisent une forme de routage.

La mise en place d'agents est largement **facilitée aujourd'hui dans des outils tel que Copilot Studio** ou grâce à des frameworks tel **LangChain** ou **LlamaIndex**. Les récents investissements de l'ordre de 200 millions de dollars dans la société « H » qui souhaite se spécialiser dans le développement de LLM orchestrateurs montre tout l'intérêt porté à ce type de modèle nécessaire pour faire fonctionner les agents.