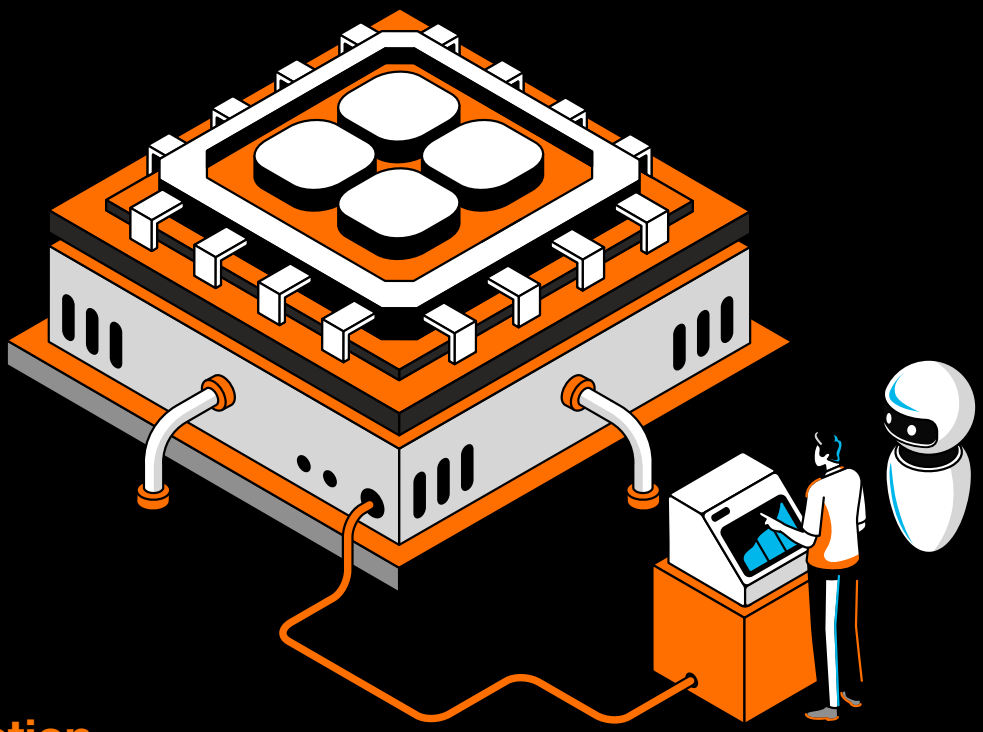
Business

Intelligence artificielle.
Valeur réelle.

Le directeur des opérations de Thomas souhaite développer une approche unifiée de l'IA générative. Mais Thomas sait qu'une stratégie hybride est plus rentable. Découvrez ici ce que lui et d'autres clients d'Orange Business ont compris.

Un appétit insatiable :

faire face à l'explosion des coûts des services IA



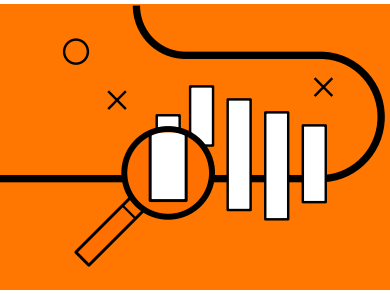
Introduction

La plupart des projets de type preuve de concept (PoC) en IA générative débutent sur des clouds publics comme Amazon Web Services (AWS), Microsoft Azure ou Google Cloud Platform. Ces environnements constituent la solution la plus simple et la plus logique pour héberger un service naissant, en raison de leur puissance de calcul facilement accessible, de leur capacité à évoluer rapidement et de l'accès aux modèles d'IA avancés qu'ils proposent.

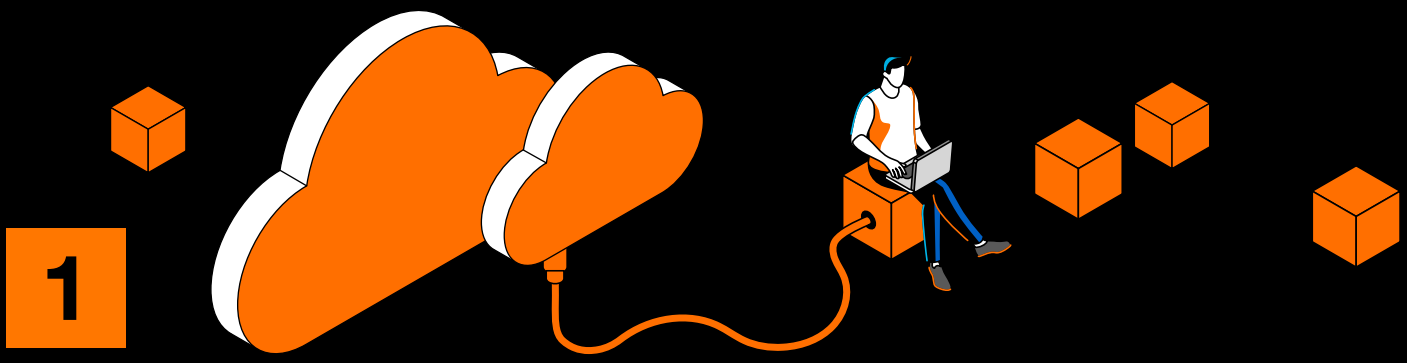
Cependant, lorsque vous passez à l'opérationnalisation, c'est-à-dire à la transformation de la PoC en un service GenAI pleinement fonctionnel à l'échelle de l'entreprise, vous ajoutez des utilisateurs, des données, et la puissance de traitement nécessaire pour répondre aux besoins du modèle augmente de manière exponentielle. Par ailleurs, à mesure que vous approfondissez votre connaissance du modèle, celui-ci doit être réentraîné et ajusté, des étapes qui requièrent chacune davantage de ressources de calcul. Résultat : vos coûts peuvent s'envoler très rapidement.

Il faut également garder à l'esprit que la facilité d'utilisation de la GenAI représente à la fois une opportunité et un risque. Grâce à son interface en langage naturel, nul besoin d'être data scientist ou ingénieur IA pour l'exploiter. Mais l'IA générative est aussi la ressource la plus gourmande, et donc la plus coûteuse, de votre environnement IT. Le fait que tout le monde puisse y accéder ne signifie pas que tout le monde doive l'utiliser.

Dans ce contexte, les décideurs les plus avisés commencent à se demander s'il n'existerait pas une manière plus rentable d'obtenir les mêmes résultats. Quant aux PME, la question est souvent plus fondamentale : est-il possible de faire quoi que ce soit avec la GenAI dans un cadre budgétaire limité ? Chez Orange Business, nous savons que, tout au long du parcours de déploiement de l'IA générative, certaines décisions peuvent avoir un impact majeur sur le coût final du service. Elles permettent de renforcer la création de valeur pour les grandes entreprises et de rendre les services GenAI accessibles au plus grand nombre.



La première de ces décisions consiste à évaluer précisément le cas d'usage. Chaque application ne nécessite pas une solution basée sur l'IA générative : des technologies d'IA plus classiques, voire des outils de visualisation de données, peuvent parfois répondre au besoin à un coût bien inférieur. Il n'est pas non plus nécessaire d'interroger l'ensemble du corpus de données à chaque requête : des économies significatives peuvent être réalisées en formant les utilisateurs à formuler des prompts plus ciblés. Enfin, l'adoption d'une approche FinOps permet d'obtenir la visibilité nécessaire pour suivre les dépenses cloud et désactiver les services inutiles.



1

Appliquer l'IA générative aux bons cas d'usage : identifier les domaines où elle crée le plus de valeur

Considérée comme la technologie la plus révolutionnaire depuis l'arrivée du smartphone, l'utilisation de l'IA générative comme solution de réponse à certaines problématiques suscite un fort engouement. Il est donc tentant de vouloir l'utiliser à tout propos. Mais faut-il vraiment mobiliser un modèle avancé pour rédiger un mail de deux lignes ? Est-ce que cela apporte une véritable valeur ajoutée ?

Le propre de l'IA générative est de produire, ou d'inférer, des résultats à partir de données existantes. Gartner définit l'inférence comme le processus par lequel un modèle d'IA applique les connaissances acquises lors de son entraînement à de nouvelles données, pour générer des contenus tels que du texte, des images ou du code. Il formule des prédictions ou des conclusions à partir des entrées reçues, ce qui revient à l'activer pour produire un résultat. Ce processus, et l'utilisation de l'unité de traitement graphique (graphics processing unit, GPU) pour traiter les données, a un impact à la fois économique et environnemental.

Dans un contexte où le marché des processeurs graphiques (GPU) connaît une croissance annuelle supérieure à 30 %, de nombreuses entreprises rencontrent des difficultés pour accéder à la puissance de calcul nécessaire au fonctionnement de leurs services GenAI. Si cela vous semble familier, sachez qu'il est possible de louer des GPU à la demande auprès d'Orange Business, via notre plateforme Cloud Avenue, une autre manière de maîtriser les coûts liés au cloud.

Tous les cas d'usage ne nécessitent pas cette intensité de calcul, ni ne justifient un tel niveau d'investissement. L'engouement autour de GenAI tend à occulter le fait que des technologies plus anciennes, comme l'IA traditionnelle ou les outils de business intelligence, peuvent fournir les mêmes résultats, à bien moindre coût. Il est donc indispensable d'identifier les cas où l'utilisation de l'IA générative est réellement pertinente. C'est pourquoi partir du cas d'usage reste essentiel.

Prenons un exemple issu de notre propre expérience : Orange Business a testé Microsoft Copilot à l'échelle de l'ensemble de l'entreprise, avant de conclure que la valeur générée ne justifiait pas un déploiement global. Nous avons donc choisi de le restreindre à deux cas d'usage : d'une part, les fonctions impliquées dans la création et la modification de contenus ; d'autre part, les chefs de projet, pour qui l'outil facilite la coordination et les comptes rendus de réunion.

Gartner estime qu'au moins 30 % des projets d'IA générative seront abandonnés après la phase de PoC d'ici fin 2025. Les raisons évoquées ? Une mauvaise qualité des données, des dispositifs de gestion des risques insuffisants, une envolée des coûts ou encore une absence de création de valeur claire. Une PoC GenAI peut coûter entre 15 000 et 50 000 dollars, voire plus, selon les fonctionnalités attendues, les exigences en matière de données et le niveau d'expertise nécessaire : les coûts financiers, mais aussi d'opportunité, liés à ces échecs sont donc considérables. Miser sur le bon cas d'usage dès le départ est crucial. Dans notre cas, nous avons sélectionné ceux pour lesquels nous pouvions démontrer un impact concret et calculer un retour sur investissement.

Recommandations concrètes

Selon Gartner, il existe deux catégories d'IA générative : l'« IA générative du quotidien », centrée sur la productivité, permet aux collaborateurs de gagner en rapidité et en efficacité dans leurs tâches habituelles ; tandis que l'« IA générative transformatrice » repose avant tout sur la créativité et peut profondément bouleverser les modèles économiques, voire des industries entières.

Si vous voulez utiliser l'IA générative facilement au quotidien, Orange Business a développé Live Intelligence : une solution prête à l'emploi qui repose sur plusieurs modèles de langage IA et fonctionne en mode SaaS (Software as a Service). Grâce à une interface simple et intuitive, les collaborateurs disposent d'une bibliothèque de prompts prédéfinis pour répondre aux besoins les plus courants : analyser ou résumer un document, extraire des informations essentielles d'un fil de mails, rédiger un compte rendu de réunion, préparer un ordre du jour, structurer un entretien ou encore relire un article.

Pour les cas d'usage plus ambitieux et à fort potentiel transformatif, Orange Business propose une méthodologie sur mesure, qui prend en compte tous les principaux défis liés à l'opérationnalisation de l'IA générative : création de valeur, sécurité, modernisation de l'infrastructure, gouvernance et conduite du changement.

2

Hébergement local ou cloud : choisir la bonne stratégie adaptée à votre usage d'IA générative

La définition en amont de votre cas d'usage est essentielle pour déterminer la stratégie d'hébergement la plus adaptée : en périphérie (Edge) ou dans le cloud. Aujourd'hui, les hyperscalers proposent des tarifs compétitifs en matière de stockage et de puissance de calcul. Si le cloud répond à votre usage, il peut représenter une solution plus économique que le déploiement et la gestion d'une infrastructure dédiée.

De manière générale, un hébergement en périphérie est à privilégier lorsque votre cas d'usage nécessite un traitement en temps réel, ou lorsque l'infrastructure de base est indisponible ou ne permet pas d'assurer une connectivité haut débit. Orange Business accompagne par exemple un acteur du secteur minier sur un projet GenAI lié à la sécurité. Les résultats doivent être produits en temps réel, et l'environnement souterrain rend difficile l'installation d'une infrastructure IT classique : dans ce contexte, le choix d'un hébergement Edge s'est imposé. Cette option peut également s'avérer pertinente lorsqu'un haut niveau de résilience ou de sécurité est requis, en garantissant la continuité de service en cas de coupure de connexion et en évitant toute transmission de données sensibles vers l'extérieur.

À l'inverse, dans un cas d'usage de type Copilot déployé dans un environnement de bureau bien connecté, sans exigence de résultats en temps réel, un hébergement dans le cloud s'avère parfaitement adapté. (Pour aller plus loin, consultez notre article de blog : « L'IA en périphérie ? Trouver le juste équilibre entre hébergement cloud et sur site pour vos services IA ».)

Un autre élément clé à prendre en compte est la conformité, et notamment la question de la souveraineté des données. Les clouds privés offrent un meilleur contrôle sur la localisation des données, avec la possibilité de les héberger physiquement dans le pays ou la région imposés par la réglementation. Les fournisseurs de cloud public disposent souvent de centres de données répartis dans plusieurs zones géographiques, rendant plus complexe le fait de garantir que certaines données restent bien dans la juridiction requise.

Les législations en matière de souveraineté des données tendent à favoriser les solutions de cloud privé, qui offrent un contrôle renforcé sur la localisation, l'accès, la sécurité et la gouvernance des données. Néanmoins, le choix entre cloud privé et public dépend aussi d'autres critères comme le coût, l'évolutivité et les besoins spécifiques de l'entreprise. Il est donc essentiel d'évaluer les exigences propres à votre activité, en tenant compte des exigences réglementaires applicables sur la souveraineté des données, pour adopter le modèle d'hébergement le plus approprié.

La législation sur la souveraineté des données privilégie souvent les solutions de cloud privé, en raison du contrôle accru qu'elles offrent sur la localisation, l'accès, la sécurité et la gouvernance des données.

En moyenne, 80 % des coûts des centres d'appel sont liés au personnel. Selon Gartner, d'ici 2029, l'IA agentique résoudra de manière autonome 80 % des demandes courantes de service client, sans intervention humaine, ce qui entraînerait une réduction de 30 % des coûts opérationnels globaux.

Une fois votre décision prise, reste à déterminer l'emplacement d'hébergement des données, une décision stratégique qui influencera fortement vos coûts globaux. Prenons l'exemple d'un chatbot GenAI déployé dans un centre de contact. Ce type de cas d'usage présente des bénéfices significatifs pour l'entreprise comme pour ses clients. En moyenne, 80 % des coûts des centres de contact sont liés au personnel. Selon Gartner, d'ici 2029, l'IA agentique résoudra de manière autonome 80 % des demandes courantes de service client, sans intervention humaine, ce qui entraînerait une réduction de 30 % des coûts opérationnels globaux. De plus, les chatbots ne subissant aucune contrainte de temps et disposant d'un accès complet aux informations nécessaires, ils sont en mesure d'apporter des réponses plus exhaustives, améliorant ainsi la satisfaction client et le Net Promoter Score (NPS).

Mais, pour garantir la pertinence des réponses, le chatbot doit pouvoir accéder aux dernières interactions du client. Or, ces données sont souvent réparties entre différents clouds et soumises à des frais de sortie (egress charges) facturés par les fournisseurs pour le transfert de données hors de leur cloud, en supplément des coûts de stockage et de calcul. Ces frais peuvent rapidement s'accumuler.

Une solution consiste à rapprocher les données de la périphérie, si cela s'avère pertinent pour votre usage. Une autre alternative consiste à adapter votre stratégie cloud en mettant en place une copie ou un miroir sécurisé des données issues des interactions clients, ce qu'on appelle un proxy de données, hébergé dans un cloud privé non soumis à ces frais de sortie.

Cette approche s'inscrit dans une logique FinOps (voir ci-dessous), qui vise à garantir une visibilité complète sur les coûts liés au cloud. En tant que partenaire des principaux fournisseurs de services cloud, Orange Business est régulièrement sollicité par ses clients pour les aider à comparer les modèles économiques, gérer les différentes parties prenantes et optimiser leurs coûts cloud.

Recommandations concrètes

Vous devez concentrer vos efforts sur la fourniture de solutions GenAI à la fois sécurisées et fiables, en apportant une vigilance particulière à la souveraineté des données et à la conformité aux exigences réglementaires. Une approche figée, tout cloud ou tout sur site, ne permet pas d'adresser efficacement l'ensemble des cas d'usage. Nous recommandons une stratégie hybride, qui permet de tirer parti des deux modèles grâce à une matrice de décision claire, fondée sur des critères tels que la latence, la sensibilité des données, les coûts et les exigences réglementaires. Votre architecture doit également intégrer une logique de modularité, pour garantir la pérennité du service et permettre l'intégration progressive de nouvelles technologies.

La solution Live Intelligence d'Orange Business offre un cadre complet pour répondre à ces défis, tout en conservant la flexibilité nécessaire pour s'adapter à l'évolution de vos besoins et des avancées technologiques.

Utiliser le modèle de langage adapté à votre cas d'usage

Tous les modèles de langage GenAI ne se valent pas. Quels composants sont nécessaires ? Lesquels sont les plus coûteux ?

Les besoins en ressources varient d'un Large Language Model (LLM, grand modèle de langage) à l'autre, certains étant plus efficaces que d'autres.

Orange Business a par exemple évalué l'utilisation de deux générations différentes d'un même modèle dans un cas d'usage précis. Les résultats étaient identiques dans les deux cas, mais la version précédente du modèle s'est révélée 90 % moins coûteuse que la plus récente. Il n'existe cependant aucune règle générale : bien que sorti après ChatGPT-4, ChatGPT-4o Mini est moins onéreux que ce dernier (même constat pour DeepSeek). Il est donc judicieux de comparer. Demandez-vous si vous avez réellement besoin du modèle le plus puissant pour votre cas d'usage.

Une autre alternative consiste à utiliser un Small Language Model (SLM, petit modèle de langage). Certains de ces modèles ont été spécifiquement optimisés pour réduire leur empreinte, via un processus appelé « distillation ». Cette méthode consiste à transférer les connaissances d'un modèle complexe (le « professeur ») vers un modèle plus léger et plus efficace (l'« élève ») et permet de réduire significativement les coûts : les besoins en puissance de calcul sont moindres, tout comme l'infrastructure nécessaire à l'exécution du SLM, ce qui abaisse sensiblement le coût global.

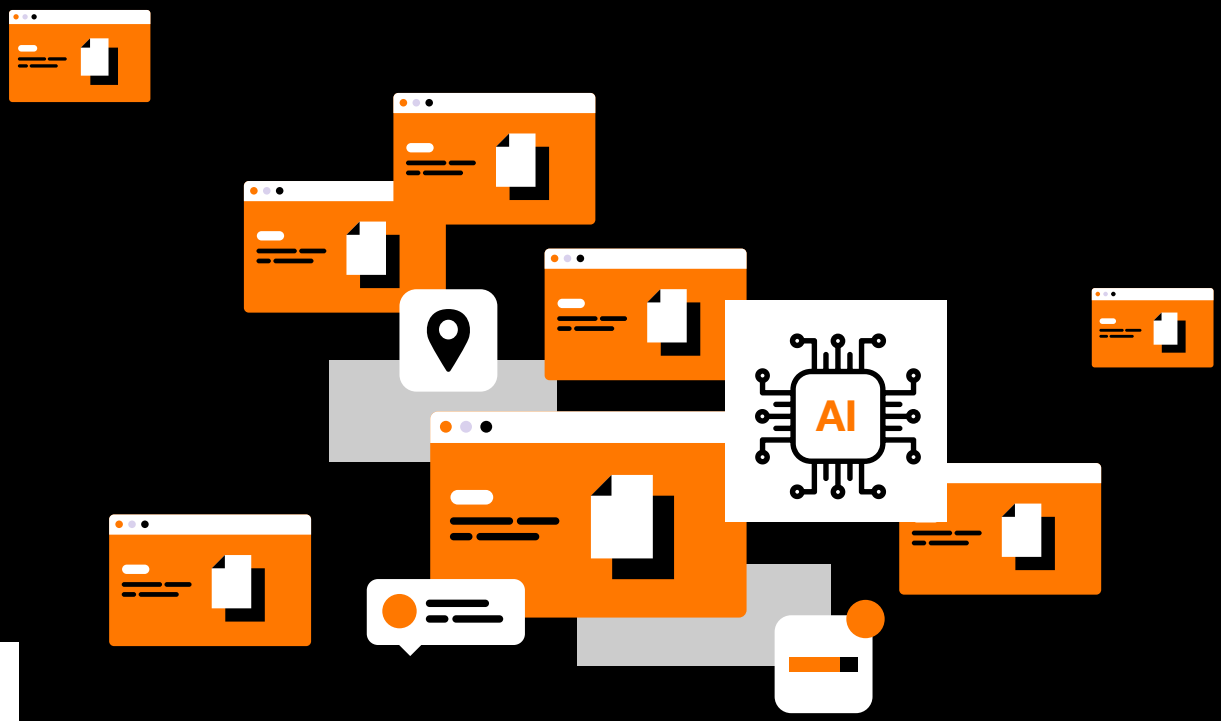
Il est important de noter que, si un modèle d'IA est entraîné de manière répétée à partir de données produites par d'autres modèles d'IA, sa capacité à générer des résultats pertinents, diversifiés et fiables diminue : on parle alors de dégradation de modèle ou « model collapse ».

Recommandations concrètes

Pour évaluer si un LLM ou un SLM est adapté à votre cas d'usage, vous devez prendre en compte plusieurs facteurs : la complexité de la tâche à accomplir, la vitesse de réponse attendue, les ressources informatiques disponibles, la volumétrie de données et le budget.

En règle générale, les SLM sont plus adaptés à des tâches simples, ciblées, nécessitant une réponse rapide et peu de ressources. Les LLM, eux, excellent dans des tâches plus complexes et nuancées, qui exigent des volumes de données importants et une puissance de calcul élevée. Il est essentiel de réaliser des tests de performance comparatifs entre différents modèles avant de faire un choix, en veillant à bien documenter les métriques de performance, les coûts associés et la qualité des résultats produits.

Compte tenu de l'évolution rapide des technologies GenAI, et de la multiplication des modèles disponibles, il peut être pertinent de faire appel à des experts indépendants pour vous accompagner dans le processus de sélection. La plateforme Live Intelligence d'Orange Business intègre déjà une sélection de LLM et de SLM rigoureusement évalués, mise à jour en continu, garantissant ainsi un accès aux modèles les plus performants, sans maintenance côté client : la plateforme gère automatiquement la sélection et les mises à jour des modèles, tout en optimisant les coûts et en préservant la flexibilité.



4

Gouvernance documentaire, concentrer le modèle sur les documents les plus pertinents

La logique voudrait que plus un LLM a accès à un grand volume d'informations, meilleures seront ses performances. En réalité, c'est souvent l'inverse : plus vous fournissez de données à un LLM, plus il les exploitera, et stocker de nombreux documents à faible valeur ajoutée peut conduire à une surexploitation des données, avec des résultats médiocres et des coûts élevés. En réalité, ce n'est pas le volume de données disponibles qui compte, mais leur pertinence.

Ainsi, si la gouvernance des données reste une composante essentielle de la gestion de l'information, la gouvernance documentaire doit être considérée. Là où les modèles d'IA traditionnels s'appuient sur des données structurées, la GenAI peut traiter un langage naturel, ce qui lui permet d'analyser des données non structurées comme des mails, des présentations ou d'autres types de documents. La gouvernance documentaire permet donc de s'assurer que seules les informations utiles sont exploitées, ce qui réduit les coûts et améliore la pertinence des résultats.

Elle constitue également un pilier fondamental de la génération augmentée de récupération (RAG), que de nombreuses entreprises adoptent pour maîtriser les coûts liés à la GenAI.

L'approche RAG repose sur la combinaison d'un système de recherche d'informations avec un modèle de langage génératif. Lorsqu'un utilisateur soumet une requête, le système recherche d'abord les informations pertinentes dans une base de connaissance externes, puis les transmet au modèle de langage afin de générer une réponse plus précise et contextualisée. Cette méthode est plus économique que le réentraînement d'un LLM sur des données internes, car elle permet de tirer parti de modèles pré-entraînés tout en les enrichissant avec des sources externes, y compris du matériel non structuré comme les mails, présentations et documents.

La gouvernance documentaire constitue un pilier fondamental de la génération augmentée de récupération (RAG), que de nombreuses entreprises adoptent pour maîtriser les coûts liés à la GenAI.

Cependant, même cette méthode suppose d'interroger un grand volume de données à chaque prompt. Or, plus ce volume est élevé, plus les besoins en puissance de calcul augmentent, ce qui est un facteur majeur d'augmentation des coûts. Une autre manière de gérer ces coûts consiste donc à réduire la quantité de données interrogées, **soit par des leviers technologiques, soit par la gouvernance**.

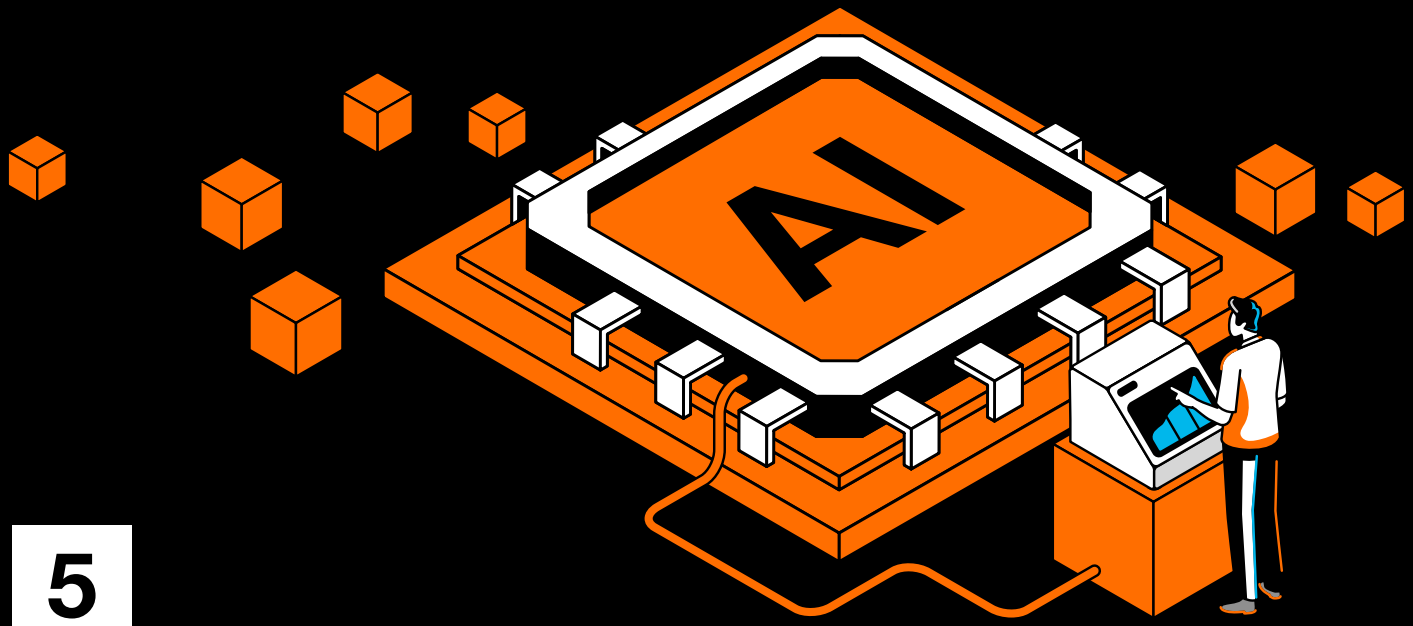
- L'approche technologique consiste à faire en sorte que le modèle GenAI crée un sous-ensemble des données disponibles, en présélectionnant uniquement les documents les plus pertinents par rapport à la requête, puis à faire travailler le LLM sur cet ensemble réduit de données.
- La seconde approche relève de la gouvernance. Prenons l'exemple d'un commercial utilisant la GenAI pour automatiser la réponse à un appel d'offres. Dans ce cas, le modèle conserve des informations sur tous les contrats en cours (ce qui permet, par exemple, aux équipes juridiques de mettre à jour les clauses). Mais un nouveau modèle tarifaire a été introduit il y a six mois. Tout prompt lié aux prix devrait donc exclure les contrats antérieurs à cette date. Cela permettrait non seulement de fournir des informations plus précises, mais aussi de réduire considérablement les coûts en limitant le volume de documents analysés.

Recommandations concrètes

Comme pour tout type de données, la règle du « garbage in, garbage out » s'applique : il est toujours plus économique d'assurer la qualité des documents dès le départ que de corriger les erreurs une fois le modèle en production. C'est également le meilleur moyen d'optimiser la valeur générée par votre service GenAI.

Pour garantir que seuls les documents pertinents sont intégrés à vos référentiels RAG, il est essentiel de faire collaborer ceux qui créent les données avec ceux qui vont les exploiter. Il s'agit de confier les responsabilités aux bonnes personnes et de mettre en place une boucle de retour d'expérience qui assure un haut niveau de qualité des résultats. Il est conseillé d'instaurer un comité de gouvernance des données et documents, chargé de superviser la pertinence des contenus. La mise en place d'une normalisation des métadonnées, de processus réguliers de revue, d'archivage ou de suppression des documents obsolètes est également recommandée. Des règles claires pour la gestion du cycle de vie des documents, ainsi que des audits réguliers du référentiel documentaire, sont essentiels pour garantir la pertinence des contenus. Orange Business dispose d'une solide expérience dans ce domaine et accompagne déjà de nombreux clients dans la mise en œuvre de ces bonnes pratiques.

Un nettoyage de données structuré et rigoureux consiste à identifier et supprimer les doublons, les valeurs aberrantes, les incohérences et les informations non pertinentes, tout en veillant à un formatage adéquat des données. Il implique également une surveillance régulière afin de détecter l'apparition de biais ou d'informations obsolètes dans les jeux de données utilisés pour l'entraînement.



5

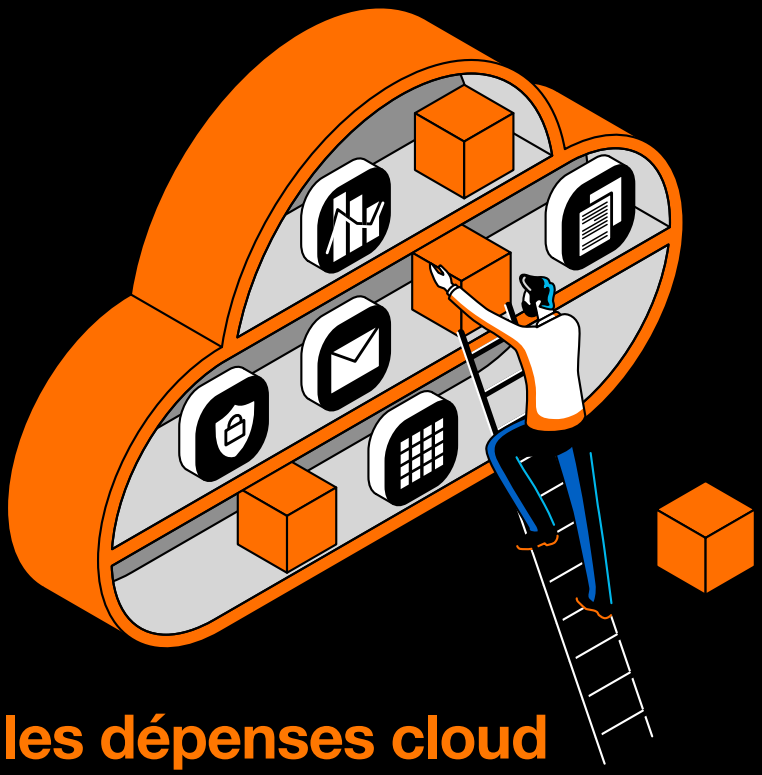
Bibliothèques de prompts : chaque requête a un coût, confiez aux utilisateurs les bons outils

Comme évoqué précédemment, le volume de données interrogées par un LLM génère des coûts importants. Si un utilisateur doit formuler cinq ou six prompts avant d'obtenir une réponse satisfaisante, les coûts en puissance de calcul augmentent sensiblement. À l'inverse, réduire le nombre de tentatives permet de réaliser des économies substantielles. Il est donc judicieux de mettre à disposition une bibliothèque de prompts : des modèles de requêtes optimisés pour obtenir des résultats pertinents dès les premières tentatives.

Recommandations concrètes

Ici encore, la collaboration est essentielle. Les spécialistes des prompts doivent travailler main dans la main avec les utilisateurs finaux pour créer une boucle d'amélioration continue, identifier les requêtes les plus fréquentes et valider les formulations les plus efficaces. Ce processus peut constituer le cœur d'un centre d'excellence dédié à la gestion des prompts, centralisant les bonnes pratiques et facilitant l'optimisation de l'usage des ressources. Vous pouvez aussi réfléchir à la mise en place d'un système de notation qui permet d'identifier et de valoriser les prompts les plus performants.

La solution Live Intelligence d'Orange Business intègre une bibliothèque de prompts préconfigurés et éprouvés dans l'industrie, que les clients peuvent utiliser immédiatement et enrichir en fonction de leurs besoins spécifiques. Cette base solide permet de réduire considérablement les délais de mise en œuvre, tout en garantissant l'efficacité des requêtes



6

FinOps, optimiser les dépenses cloud grâce à une meilleure visibilité des usages GenAI

Les coûts directs de votre service GenAI, de stockage et de calcul sont clairement identifiables. Mais les coûts indirects, liés à l'infrastructure nécessaire au bon fonctionnement du service, sont souvent mal évalués et peuvent se révéler bien supérieurs aux attentes.

L'approche FinOps vise à assurer une gouvernance financière et opérationnelle des budgets liés au cloud. Il s'agit d'une méthode de gestion qui favorise la transparence et la collaboration entre les équipes techniques, financières et commerciales. L'objectif est de prendre des décisions éclairées grâce aux données sur l'usage des ressources cloud, pour optimiser la valeur commerciale tout en maîtrisant les coûts avec efficacité.

Prenons un exemple concret : si vous déployez un service GenAI hébergé par un hyperscaler, il peut rapidement devenir très populaire en interne et entraîner une explosion des dépenses. C'est précisément ce que nous avons constaté chez Orange Business lorsque nous avons autorisé nos employés à utiliser ChatGPT. Nous avons alors mis en place un plafond d'usage pour éviter les dérives budgétaires.

Recommandations concrètes

Il existe plusieurs stratégies fondées sur le modèle FinOps que vous pouvez mettre en œuvre pour réduire les coûts cloud liés à vos services GenAI. Tout d'abord, il convient d'établir un cadre solide de gouvernance des dépenses cloud et d'optimiser l'utilisation des ressources, notamment en désactivant celles qui ne sont pas utilisées. Vous pouvez également vous appuyer sur des outils de gestion des coûts tels qu'AWS Cost Explorer ou Azure Cost Management afin de mieux visualiser vos dépenses et de mettre en place des modèles de refacturation, dans lesquels les coûts cloud sont répartis entre les différents départements.

Cependant, ces outils ne couvrent pas l'ensemble du périmètre des coûts liés à la GenAI, qui incluent les données, l'usage, l'infrastructure, la cybersécurité et le cloud, autant de dimensions souvent en dehors du contrôle direct des hyperscalers. C'est pourquoi il est essentiel de faire appel à une expertise indépendante, afin d'adopter une approche globale de la gestion des coûts.

Business

Conclusion

Comme vous pouvez le constater, il n'existe pas de solution unique pour réduire les coûts relatifs au cloud. En revanche, une série de décisions ciblées peut vous permettre de baisser sensiblement vos dépenses tout en maximisant la valeur générée de vos services GenAI.

Ce qui est certain, c'est que l'infrastructure requise demain pour la GenAI ne ressemblera pas à celle que vous utilisez aujourd'hui. Une étude récente commandée par Orange Business a révélé que moins de la moitié des entreprises interrogées disposaient, ou allaient disposer, de l'infrastructure nécessaire à l'opérationnalisation de leurs projets GenAI. Le projet DeepSeek a d'ailleurs montré qu'il est possible de faire fonctionner des services GenAI avec bien moins de ressources qu'on ne le pensait.

Les efforts que vous investirez pour adapter votre infrastructure numérique à vos besoins peuvent vous amener à réduire votre empreinte cloud et à renforcer vos capacités sur site, ou inversement. En maîtrisant votre infrastructure cloud et Edge, vous gagnez l'agilité nécessaire pour trouver le bon équilibre entre le volume de données à gérer et les lieux où elles sont stockées. Cela vous permet de pérenniser votre activité tout en optimisant la gestion de vos coûts.

C'est précisément pour répondre à cette exigence de flexibilité que nous avons conçu notre Evolution Platform. Il s'agit d'une plateforme composable, offrant un haut niveau d'interopérabilité avec les partenaires de notre écosystème de référence, afin de vous permettre de faire évoluer rapidement et à moindre coût votre infrastructure numérique. Elle est proposée selon un modèle « network as a service », semblable à celui du cloud, avec une tarification à l'usage qui vous permet d'ajuster la taille de votre réseau en fonction des circonstances ou des choix stratégiques.

Comme le disait Peter Drucker, « la meilleure façon de prédire l'avenir, c'est de le créer ». Mais n'oublions pas le proverbe suédois : « Celui qui achète ce dont il n'a pas besoin se vole lui-même. » En suivant les recommandations de ce livre blanc, vous devriez pouvoir concilier les deux.