XAVIER BAYS
GAEL LEDERREY
JOSÉ LAMAS-VALVERDE

# APPLICATIONS DE L'IA POUR LES PME: SAISIR LES OPPORTUNITÉS EN MAÎTRISANT LES RISQUES

## Comment créer de la valeur à partir des données grâce aux algorithmes?

**La valorisation des données liée au business model est un enjeu stratégique. En Suisse, des solutions pour PME se développent, parmi lesquelles le conseil et la conception d'algorithmes prédictifs sur mesure. Deux applications concrètes d'IA illustrent ces possibilités: la prédiction à court terme de la consommation électrique régionale et la recherche de documents dans une base de fichiers.**

### 1. INTRODUCTION

L'intelligence artificielle (IA) occupe aujourd'hui le devant de la scène. Et pour cause, une révolution technologique est en marche. Cette innovation aussi disruptive que l'avènement d'Internet transforme profondément le rapport à la machine, les interactions avec les données et la manière de communiquer avec les ordinateurs.

L'IA n'est pourtant pas un concept né en 2022 avec la démocratisation de ChatGPT. Ses origines remontent aux années 1950. Au fil des décennies, divers phénomènes technologiques ont successivement suscité l'attention: *data mining, statistical learning, machine learning, big data, deep learning, data science, intelligence artificielle, agents IA,* etc.

**1.1 Pourquoi créer des algorithmes?** Dans cet article, l'attention se portera sur la volonté de dégager de la valeur des données. Pour y parvenir, il est nécessaire de créer des algorithmes, soit une suite d'actions programmées qui permettront à un ordinateur de traiter et d'affiner les données afin d'en révéler toute la richesse.

Selon l'European AI Act [1], un «système d'IA» est un système basé sur une machine, conçu pour fonctionner avec différents niveaux d'autonomie et pouvant faire preuve d'adaptabilité après son déploiement et qui, pour des objectifs explicites ou implicites, déduit, à partir des données qu'il reçoit, comment générer des résultats tels que des prédictions, du contenu, des recommandations ou des décisions susceptibles d'influencer des environnements physiques ou virtuels.

La Suisse ne dispose actuellement d'aucune réglementation en vigueur concernant l'IA. Les débats en cours opposent deux visions: la nécessité de protéger la population et la crainte que des réglementations trop strictes freinent les progrès technologiques. La nouvelle loi sur la protection des données (nLPD), entrée en vigueur le 1er septembre 2023, encadre ainsi l'utilisation des données personnelles en renforçant les droits de l'individu.

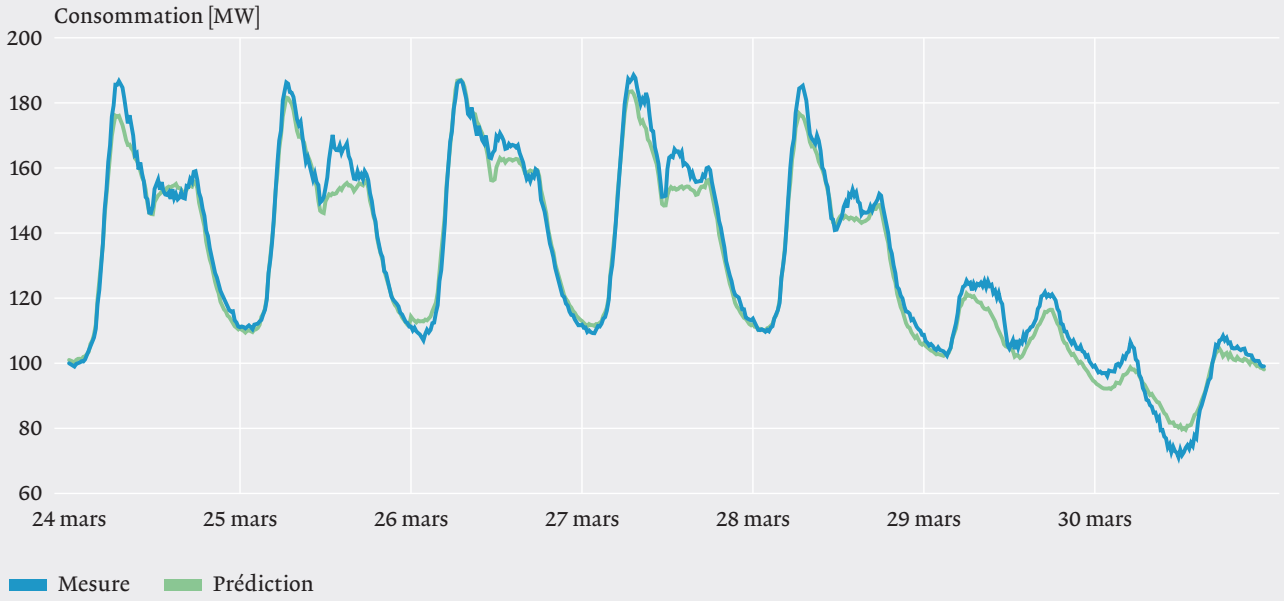
Depuis les années 2010, les progrès de l'informatique ont favorisé l'essor de cette discipline issue des statistiques, requérant une certaine puissance de calcul et de mémoire pour traiter de grandes quantités de données. La convergence de ces domaines a ouvert des perspectives phénoménales pour la compréhension et la prédiction du monde environnant.

Deux exemples d'application de l'intelligence artificielle dans le domaine de la prédiction sont présentés: le premier, dit «traditionnel», porte sur la consommation électrique; le second, dit «moderne», concerne la recherche intelligente de document. Les similarités et les différences de ces deux exemples seront finalement mises en lumière.

XAVIER BAYS,
INGENIEUR EN MATHEMATIQUES APPLIQUEES EPFL, COFONDATEUR ET HEAD OF SERVICES, SWISS STATISTICAL DESIGN & INNOVATION

GAEL LEDERREY,
PH.D. MACHINE LEARNING, SENIOR DATA SCIENTIST, SWISS STATISTICAL DESIGN & INNOVATION

Figure 1: **MESURE ET PRÉDICTION DE LA COURBE DE CONSOMMATION**




## 2. EXEMPLE 1: PRÉDIRE LA CONSOMMATION ÉLECTRIQUE

Savez-vous comment votre foyer ou votre entreprise bénéficie toujours d'électricité alors même que celle-ci ne peut être stockée une fois produite?

Des équipes mobilisent leurs efforts, notamment à l'aide d'algorithmes (comme il sera présenté ultérieurement), afin de garantir un approvisionnement constant en électricité, ni excédentaire ni insuffisant. Autrement dit, l'équilibre entre l'offre et la demande doit être maintenu. Un excès de production par rapport à la demande saturerait le réseau, provoquant une élévation de la tension susceptible d'entraîner son effondrement. À l'inverse, une demande trop importante pourrait conduire à une pénurie. Selon le droit de l'énergie en Suisse [2], les gestionnaires de réseau de distribution ont l'obligation d'alimenter leurs clients en électricité. La consommation individuelle (particulier, entreprise, etc.) détermine ainsi le volume d'énergie requis, auquel le réseau doit s'adapter. L'essor des énergies renouvelables a complexifié la gestion du réseau, en raison notamment de la difficulté à anticiper une production soumise aux conditions météorologiques. Toutefois, veiller à l'équilibre entre production et consommation électrique constituait déjà un défi avant l'introduction des panneaux solaires et des éoliennes.

JOSÉ LAMAS-VALVERDE, PH.D. PHYSIQUE, EXPERT EN ANALYSE DES RISQUES, MÉTHODES D'ANALYSE QUANTITATIVE ET ANALYSE DÉCISIONNELLE, FONDATEUR-DIRECTEUR, GESTRISK, JOSE.LAMAS@GESTIONDESRISQUES.CH

Le *trader* d'électricité, parfois aussi appelé gestionnaire de portefeuille, a pour mission d'élaborer un plan de production et de piloter les opérations d'achat et de vente d'énergie afin de répondre à la demande. Il doit donc prévoir la consommation électrique à venir et s'assurer de livrer la même quantité d'énergie. Pour atteindre cet objectif, deux options s'offrent à lui: produire sa propre électricité (ou éteindre une centrale de production), ou bien acheter (ou vendre) de l'énergie sur le marché.

**2.1 La fonction prédictive des modèles.** Dans ce contexte, la qualité des prévisions apparaît clairement comme un enjeu central. Le *trader* va ainsi essayer de prévoir l'avenir en s'appuyant sur une sélection de modèles [3] d'IA et sur son expertise. Il travaille sur différents horizons de temps, allant de l'*intraday* (ajustement le jour même) au *day-ahead* (prévision pour le lendemain), jusqu'à des prévisions à plus de trois ans. La *figure 1* illustre un exemple de prévisions en *day-ahead*, la courbe bleue indique la consommation réelle, la courbe verte la consommation estimée par le modèle.

Un ensemble de modèles d'intelligence artificielle a été conçu afin de soutenir le *trader* dans sa prise de décision à court terme (*intraday* et *day-ahead*). Les données d'entrée sont les mêmes que celles utilisées par le *trader*: historique de la consommation, historique des données météorologiques sur les régions concernées, prévisions météorologiques pour les jours à venir, calendrier civil, etc. Il en ressort une courbe prédictive estimant la consommation et/ou la production pour les cinq prochains jours, avec une granularité de 15 minutes.

**2.2 Adapter le modèle à la pratique.** Pour mettre en place un tel modèle, il a été nécessaire de passer du temps avec les *traders* pour comprendre comment ils réfléchissent, sur quelles données ils se fondent et quelles informations ils ex-

Figure 2: **MESURE, CORRECTION ET PRÉDICTION DE LA COURBE DE CONSOMMATION**

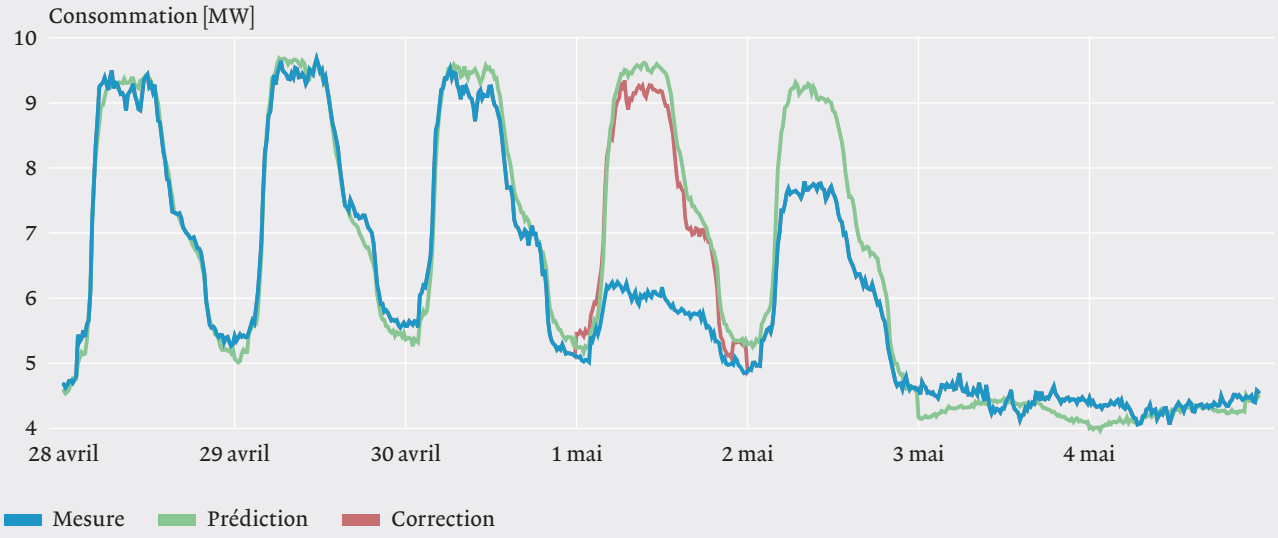
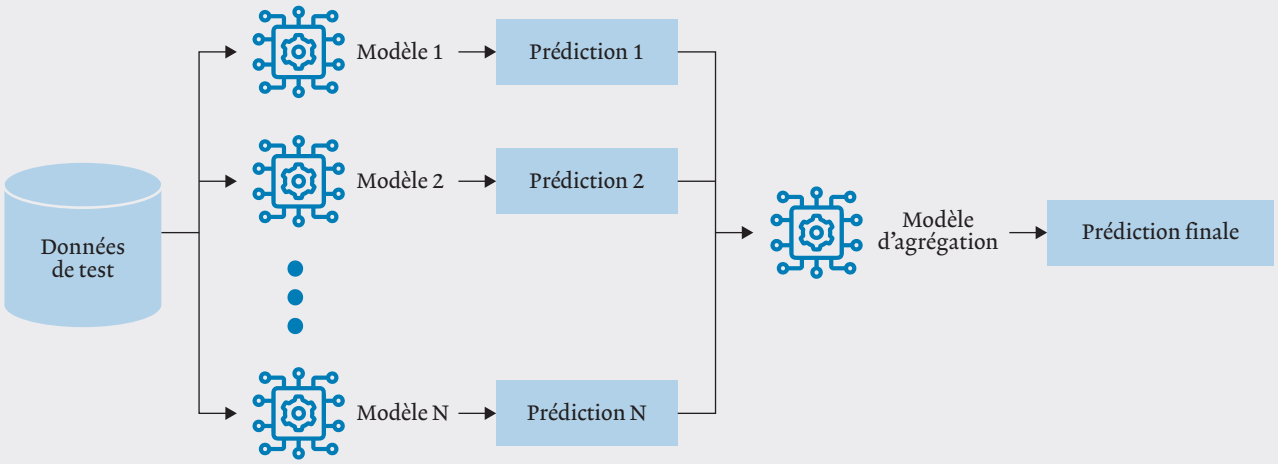



Figure 3: **DIAGRAMME ILLUSTRANT L'ENSEMBLE LEARNING**




traient de ces données. «L'idée, ici, est de reproduire la réflexion humaine», qui s'appuie sur deux constats majeurs:
→ un fort impact du court terme: la journée d'hier va influencer la prévision de demain;
→ et un fort impact du long terme: la même date de l'année passée et des années précédentes va influencer la prévision de demain [4].

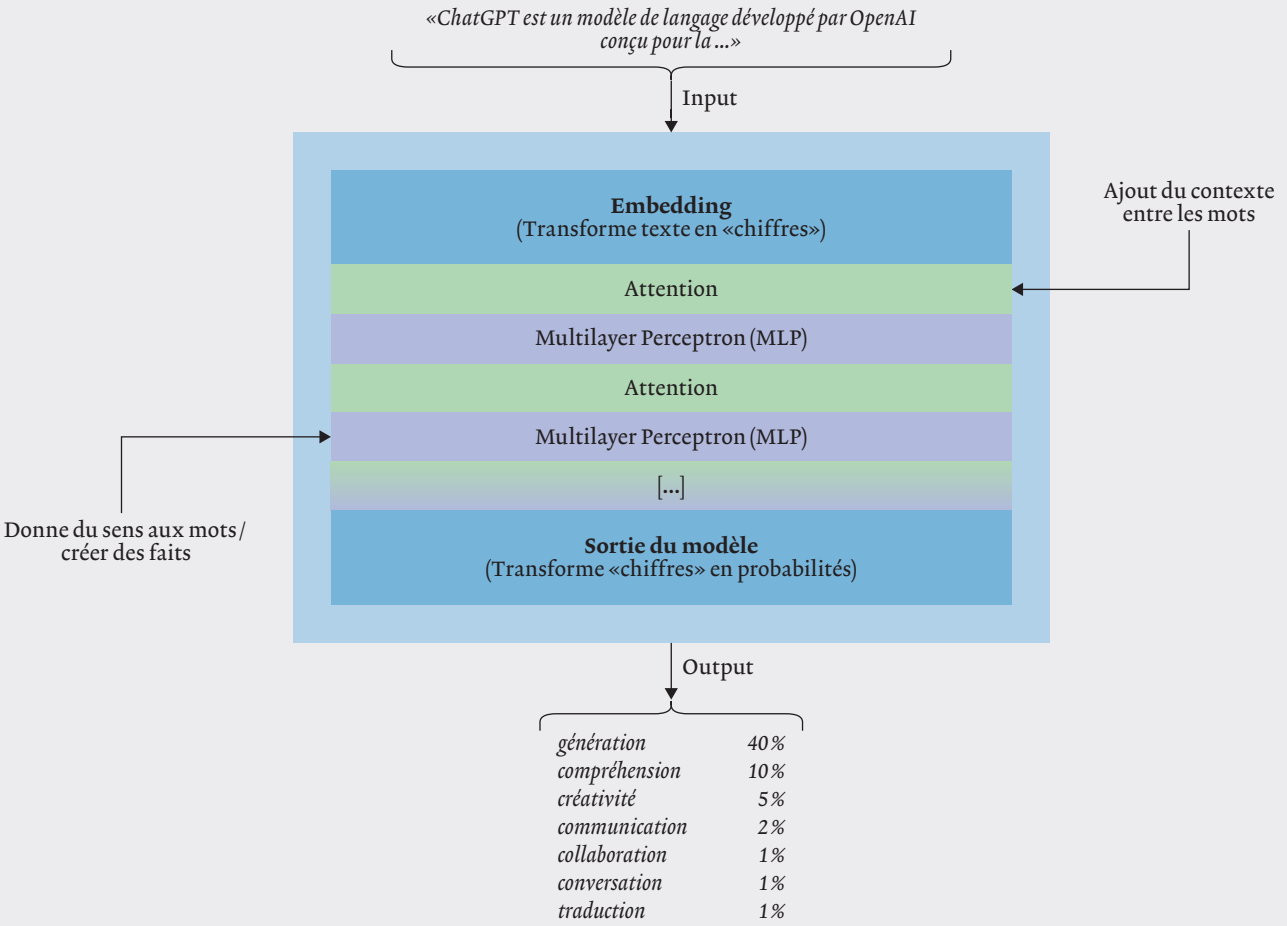
Pour compléter la méthode humaine, une analyse des données a également été réalisée pour identifier des corrélations susceptibles d'échapper à l'œil des *traders*.

Fort de toutes ces informations, le *data scientist* (scientifique des données) va passer une grande partie de son temps à réaliser le *feature engineering*. Cette étape consiste à créer des variables intéressantes et compréhensibles pour l'ordinateur à partir des données brutes. Le *feature engineering* est l'ingrédient secret qui distingue un bon modèle d'un excellent modèle.

**2.3 L'importance de s'assurer de la qualité des données.** Un grand travail a également été mené sur la qualité des données d'entrée. Il peut sembler surprenant que les données fournies par le réseau ne soient pas définitives et qu'elles puissent varier considérablement pendant six mois. Un «modèle de détection d'anomalies» a ainsi été mis en place pour identifier les cas où la qualité des données d'entrée est insuffisante pour une utilisation fiable. Ce modèle analyse l'écart (la déviation) entre la dernière prévision et la valeur mesurée, ainsi que la comparaison de la dernière mesure avec celles des jours précédents.

En présence d'une anomalie, les données du modèle (données synthétiques) remplacent les données brutes réelles de l'historique. La *figure 2* montre un exemple de correction effectuée le 1er mai, consécutivement à la détection d'une dégradation de la qualité des mesures du réseau.

Figure 4: **SCHÉMA SIMPLIFIÉ DU MODÈLE GPT3 D'OPENAI**




**2.4 Un assemblage de modèles.** La solution combine en réalité un ensemble de modèles. Plusieurs technologies sont mobilisées, comme des modèles dits de *machine learning* classique (p. ex. *Gradient Boosting*) ou de réseaux de neurones (p. ex. LSTM pour *Long short-term memory*). Un modèle final est conçu pour déterminer, en fonction du contexte, quel modèle spécifique doit être activé. Cette technique est connue sous le nom d'*ensemble learning*. La *figure 3* schématise le concept de l'*ensemble learning*.

**2.5 Clés de succès: vérification et validation.** Deux facteurs de succès peuvent être dégagés de ce projet. Le premier réside dans le monitoring de la solution, une fois mise en production. Il est nécessaire de garantir le bon déroulement de l'automatisation des prévisions, en veillant à la réception des données, à la disponibilité des serveurs et à la qualité de la prévision. Différents systèmes d'alerte ont été créés à cette fin. Ils permettent, par exemple, de détecter une déviation lente de la qualité attendue, une perte soudaine de qualité, ou encore une insuffisance persistante dans la qualité des données d'entrée, etc.

Le second facteur de réussite est l'utilisation exclusive de données effectivement disponibles à un instant *t* lors du développement de la solution. Bien que cette remarque puisse paraître triviale, elle constitue une source d'erreurs fréquente dans le cadre de la modélisation sur des séries temporelles. En effet, le *data scientist* a accès à l'ensemble des données (réelles et prédictives) lors de l'élaboration de la solution: des prévisions à différents horizons, des historiques de qualités variables selon l'horizon, etc.

Que se passe-t-il si un modèle est entraîné à partir de données provenant du futur, autrement dit d'informations qui ne sont pas disponibles au moment où la prévision est censée être faite? Lors du passage en production, le modèle devient instable et sa qualité se dégrade brutalement. Ce qui semblait être un modèle précis et efficace se révèle être inutilisable au dernier moment.

## 3. EXEMPLE 2: RECHERCHE INTELLIGENTE D'UN DOCUMENT

Avez-vous déjà perdu du temps à rechercher un document sur votre ordinateur? Les personnes les plus expérimentées de votre entreprise sont-elles vos principales sources d'information pour retrouver d'anciens projets? Souhaitez-vous tirer parti de tous ces documents Word et PDF qui dorment dans la base de données de votre organisation? Si vous avez répondu oui à l'une de ces questions, alors les LLM et les RAG vous seront utiles.

**3.1 Qu'est-ce qu'un Large Language Model (LLM)?** Les LLM sont des modèles d'IA spécialement conçus pour traiter le langage naturel. Le langage naturel désigne la manière dont les humains communiquent spontanément entre eux, sans effort particulier sur la formulation ou la structure du discours. Le terme LLM regroupe tous les modèles équivalents à ChatGPT.

Les LLM sont conçus pour recevoir en entrée des données exprimées en langage naturel et fournir en sortie une réponse également en langage naturel, généralement du texte, mais parfois aussi du son. Leur développement a connu une accélération majeure à partir du 30 novembre 2022, date à laquelle la version 3 de ChatGPT a été rendue accessible au grand public. Cette avancée a marqué une rupture significative dans la capacité à traiter le langage naturel, ouvrant le champ des possibles de manière inédite.

La création et l'utilisation des *Transformers* par OpenAI constituent une avancée majeure. Cette architecture de modèle permet une compréhension fine des données en s'appuyant sur des couches successives d'interprétation sémantique et de contextualisation. Cette architecture est présentée dans la *figure 4*.

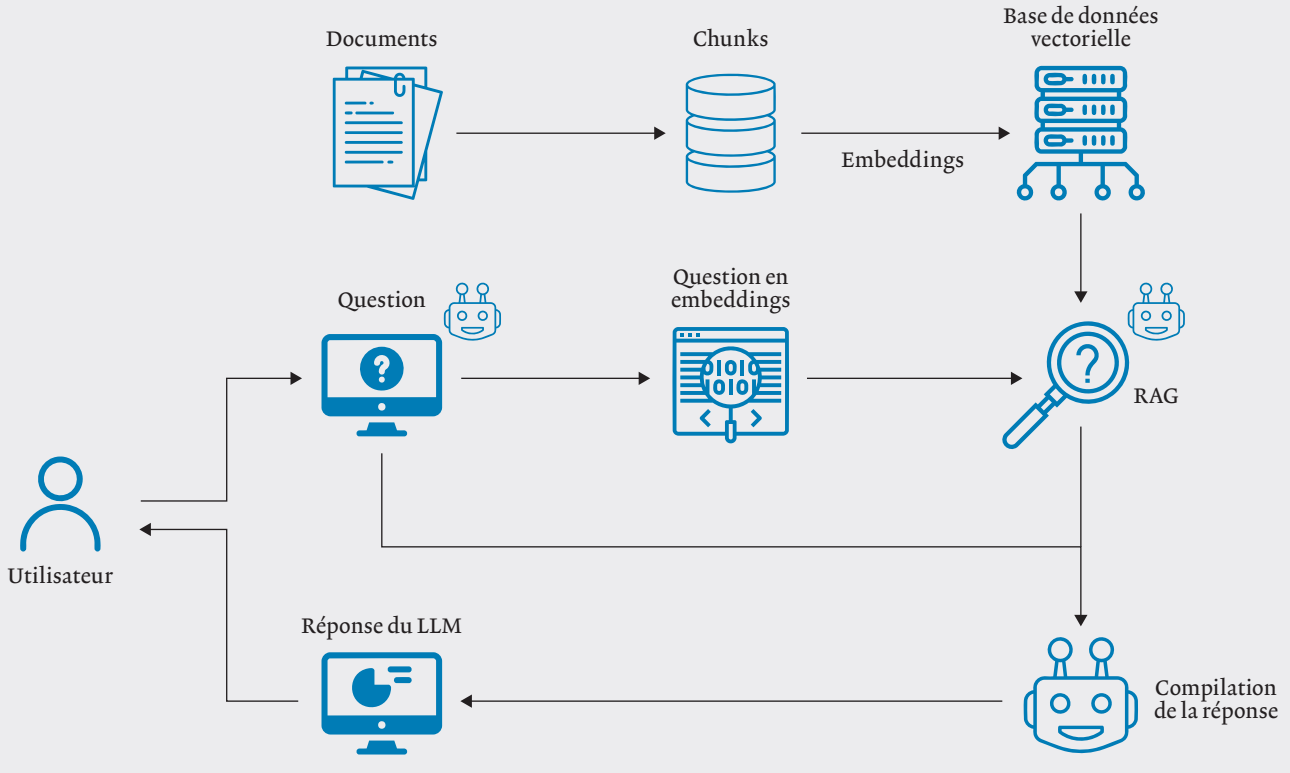
Pour permettre le traitement automatique du texte, il est nécessaire de transformer les données textuelles en chiffres. Cette étape, appelée l'*embedding*, permet de convertir des données textuelles en une information compréhensible par un ordinateur. Créer un nouvel *embedding* peut être extrêmement coûteux et se chiffrer en dizaines voire en centaines de millions de dollars. Cependant, de nombreux *embeddings* sont disponibles gratuitement en open source.

Dans le cas de notre exemple, il s'agit de retrouver un document contenant une information précise parmi une vaste base documentaire. Pour accomplir cette tâche, un système dit *Retrieval Augmented Generation* (RAG) peut être mis en place, permettant d'extraire de manière intelligente des contenus pertinents en réponse à une requête précise, à partir d'un corpus documentaire.

**3.2 Comment fonctionne un RAG?** Le RAG fonctionne de la manière suivante: l'ensemble des documents du corpus est référencé et découpé en petites parties appelées *chunks*, qui sont ensuite transformées en vecteurs numériques grâce à l'*embedding*. La requête formulée subit le même traitement de transformation. Une fonction de similarité est alors utilisée pour identifier les passages les plus pertinents en fonction de la requête. Le système retourne ensuite une référence à un ou plusieurs documents de votre corpus. À partir de là, il devient possible de traiter l'information de multiples façons: fournir un extrait du document, générer une réponse en langage naturel, ou encore proposer une synthèse des contenus retrouvés. La *figure 5* illustre le processus du RAG couplé avec un LLM.

**3.3 Valeur ajoutée du système IA.** En résumé, il est désormais possible de converser avec une IA (LLM) qui s'occupe de retrouver le document que vous cherchez (RAG), mais aussi d'aller plus loin. L'IA peut, par exemple, extraire les points clés, comparer plusieurs documents, ou même générer une première version d'un rapport ou d'un courriel à partir des informations collectées. Elle peut également faciliter le par-

Figure 5: **SCHÉMA D'UN SYSTÈME COMBINANT LLM ET RAG**



tage de connaissances au sein de l'entreprise, automatiser certaines tâches documentaires, ou encore adapter ses réponses selon le contexte ou le profil de l'utilisateur. Les possibilités offertes par ces technologies sont vastes et évoluent rapidement.

Ce cas d'usage fait ses preuves dans de nombreuses entreprises et est appelé à devenir la nouvelle norme pour retrouver des informations dans un système informatique. L'époque où la barre de recherche classique ne retrouvait pas les documents est désormais révolue. Cette application rencontre actuellement un franc succès en entreprise, notamment dans les PME.

## 4. LIENS ET DIFFÉRENCES DE CES EXEMPLES

Un observateur non averti pourrait penser que les deux exemples présentés n'ont rien en commun. En effet, les types de données traitées, l'application métier et les algorithmes diffèrent largement. La plupart des PME ignorent ainsi souvent vers quels interlocuteurs se tourner pour développer ces différentes applications. Or, ce sont les mêmes professionnels qui réaliseront ces deux projets car, pour les spécialistes de la donnée, les concepts sous-jacents sont identiques. Les entreprises disposant d'une expérience significative dans le traitement et l'analyse des données sont les mieux placées pour déployer les solutions d'intelligence artificielle.

**4.1 Similarités.** Dans les deux cas, le flux de données suit une logique similaire: une donnée d'entrée (plus ou moins complexe) est fournie et une donnée en réponse est attendue. La presse et les médias cherchent à différencier les nouvelles IA en les appelant génératives ou *genIA*, en référence au fait qu'elles génèrent des données, sous la forme de texte, de son ou de figures. Cependant, il convient de rappeler que tout algorithme d'IA ou de *machine learning* a pour objectif de produire une donnée en sortie. L'expert en IA va traiter les deux problématiques avec la même approche: compréhension des enjeux et du métier, nettoyage des données, application d'un algorithme, tests et validation de la solution, monitoring de la qualité, etc.

**4.2 Différences.** Une différence majeure subsiste toutefois entre les deux exemples proposés. Dans l'exemple du RAG, les modèles utilisés sont préalablement entraînés *(pre-trained)*. L'*embedding* utilisé est soit fourni *(open source)*, soit intégré à la solution choisie. Cette approche, qui consiste à exploiter un algorithme génératif pré-entraîné (en anglais *Generative Pre-Trained*, d'où l'acronyme GPT), représente une nouvelle manière de travailler avec les algorithmes existants. Il était jusqu'alors nécessaire de procéder soi-même à l'entraînement. Ce changement de paradigme, couplé à la grande capacité des LLM à traiter le langage naturel, a contribué à la démocratisation actuelle de l'IA. En outre, l'utilisation de modèles pré-entraînés réduit considérablement l'effort de développement requis pour déployer une telle solution.

Dans l'exemple de la prédiction énergétique, le secteur est particulièrement concurrentiel, avec des *traders* qui, très précis dans leur travail, s'appuient déjà sur des outils analytiques avancés. Le développement du modèle (hors interfaces visuelles et intégration client) pour un tel projet nécessite en moyenne six mois de travail pour une équipe de deux ou trois personnes. Dans ce contexte, un développement sur mesure, d'une grande précision, est nécessaire et demande évidemment davantage d'efforts. Cependant, les économies en termes de coûts, d'effort et de sécurité générées par une meilleure prévision énergétique justifient largement les investissements.

Dans le cas d'un système RAG, le temps de développement se réduit généralement à une dizaine de jours. Cette rapidité s'explique par le fait que la solution s'apparente davantage à l'assemblage de briques technologiques existantes plutôt qu'à un développement entièrement sur mesure. Il serait d'ailleurs irréaliste, tant en termes de coûts que de délais, de recréer un modèle de type ChatGPT pour chaque nouveau client. L'approche RAG permet ainsi de bénéficier de la puissance des grands modèles de langage tout en adaptant la solution aux besoins spécifiques de chaque entreprise, de façon rapide et efficiente.

## 5. CONCLUSION

L'intelligence artificielle, et notamment les solutions reposant sur les LLM et le RAG, ouvre de nouvelles perspectives aux PME. Elle permet une exploitation intelligente des données, accessible sans nécessiter d'investissements démesurés ni de délais interminables. Comme le dit l'adage, «mieux vaut avancer à petits pas que de rester immobile». Pour les PME, il est temps de saisir ces opportunités et de transformer leurs données en un véritable levier de croissance. ■

**Notes: 1)** Loi sur l'intelligence artificielle dans l'EU, entrée en vigueur en août 2024 https://artificialintelligenceact.eu/fr/article/3 **2)** Loi sur l'approvisionnement en électricité (LApEl) en particulier l'Article 6. **3)** Il existe de nombreux modèles de machine learning utiles à cette tâche: régression linéaire, gradient boosting, random forest, LSTM NN, etc. **4)** La consommation d'un mercredi ressemble à celle d'un autre mercredi, mais pas à celle d'un dimanche. La météo et le climat ont également un fort impact sur la consommation. Il est donc essentiel de parvenir à comparer le jour à prédire avec d'autres jours similaires (calendrier civil, conditions météorologiques, historique, etc.).