



Le guide du CTO pour la recherche avec IA de bout en bout pour l'e-commerce



Comment la recherche avec Intelligence Artificielle native permet de réduire la charge de travail tout en améliorant la compréhension et l'expérience utilisateur...

La simplicité est essentielle pour les moteurs de recherche, mais parvenir à cette simplicité est complexe. Au cours des 20 dernières années, la plupart des solutions de recherche pour les sites web ont utilisé la correspondance de la requête du client avec un index de recherche. Commence alors le travail d'amélioration du moteur de recherche. Il faut ensuite améliorer le moteur de recherche car il ne comprend pas bien les intentions des utilisateurs.

La personne qui demande « qu'est-ce que rust » cherche-t-elle une information sur la signification du mot rust en anglais, ou un langage de programmation informatique ou encore autre chose ? Pendant des années, tous ceux qui ont eu à optimiser un moteur de recherche ont utilisé des règles manuelles, des synonymes, le bourrage de mots clés (keyword stuffing) ou d'autres techniques d'intelligence artificielle de base pour « ajuster » les correspondances afin d'obtenir une meilleure compréhension.

CORRESPONDANCE

Mots-Clés

(2000 à aujourd'hui)

Correspondance entre le mot-clé et l'index du catalogue produit

AJUSTEMENT

AI basique

(2013 à aujourd'hui)

- + règles, synonymes
- + AI Hacks
- + NLP add-ons

BALBUTIEMENT

Vecteurs

(2016 à aujourd'hui)

Première génération de Machine Learning / Intelligence Artificielle. Meilleure qualité de résultats, mais très onéreux et très lent

La création d'une solution avec IA de bout en bout était possible, mais le déploiement de ces solutions à grande échelle était inaccessible en raison de problèmes de coûts et de performances.

Algolia NeuralSearch change la donne, grâce à son Intelligence Artificielle native de bout en bout.

CORRESPONDANCE

Mots-Clés

(2000 à aujourd'hui)

Correspondance entre le mot-clé et l'index du catalogue produit

AJUSTEMENT

AI basique

(2013 à aujourd'hui)

- + règles, synonymes
- + AI Hacks
- + NLP add-ons

BALBUTIEMENT

Vecteurs

(2016 à aujourd'hui)

Première génération de Machine Learning / Intelligence Artificielle. Meilleure qualité de résultats, mais très onéreux et très lent

Neural Hashes

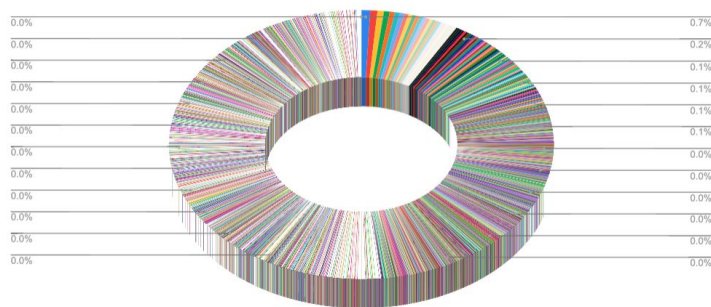
2023



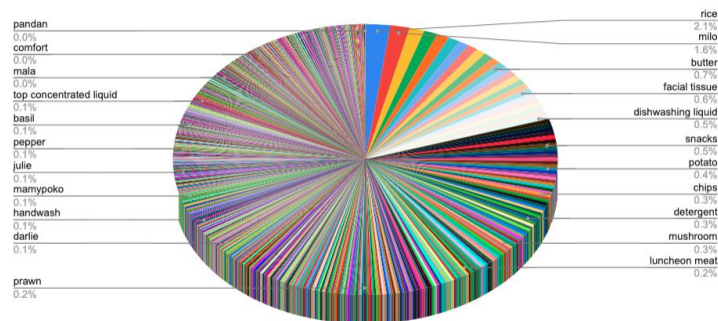
Personnalisation
Reclassement automatique (Learn to Rank)
AI évolutive
Contrôle par les équipes métiers

Pourquoi l'IA de bout en bout est importante

Il s'agit d'un sujet à part entière, mais avant d'aller plus loin, regardons pourquoi l'IA de bout en bout est importante. Les données montrent que 50 à 70 % du chiffre d'affaires est perdu à cause du manque d'optimisation des requêtes longues (en anglais, long tail) . En théorie, avec un temps et des ressources illimités, on pourrait tout optimiser manuellement, mais cela dépasse les capacités humaines.



Prenons l'exemple de deux grands détailliers au Royaume-Uni et aux États-Unis, dont les solutions de recherche sont représentatives de la majorité des entreprises en ligne. Quelqu'un peut-il gérer manuellement ce type de requêtes longues ? Est-il possible d'optimiser la recherche en intégrant l'IA de bout en bout dans le moteur de recherche ? Non, ce n'est pas réalisable.



Le nombre de variantes de produits que l'on trouve généralement sur les sites d'e-commerce est impressionnant, ce qui complique l'optimisation de chaque terme pour les requêtes longues, voire la rend impossible.

C'est là que l'apprentissage automatique entre en jeu. Les algorithmes d'apprentissage automatique peuvent optimiser tous les catalogues de recherche à 100 %. Jusqu'à présent, cela n'était pas facilement réalisable, mais avec la maturité des algorithmes d'IA, des données plus riches et un cloud computing plus puissant, il est désormais possible d'appliquer l'IA à des catalogues de toute taille. Le déploiement de l'IA de bout en bout, c'est à dire de la saisie de la requête par l'utilisateur à l'affichage des résultats, est la solution à ce défi.

Pourquoi l'IA de bout en bout est-elle nécessaire ? Jusqu'à présent, la plupart des solutions d'IA étaient destinées à résoudre des parties spécifiques de la requête de recherche. Par exemple, le traitement des requêtes, soit la première étape de résolution d'une recherche, ou le classement, qui vient généralement en dernier.

Le chaînon manquant était l'utilisation de l'IA pour la récupération d'informations. La récupération d'informations est le moment où le moteur de recherche tente de trouver les informations les plus pertinentes. C'est sans doute l'IA qui a le plus d'impact sur la qualité des résultats lors de la recherche. Grâce à la maturité de l'apprentissage automatique de l'IA aujourd'hui, il est enfin possible de résoudre ce problème.

Dans cet ebook, nous allons décrire comment chaque étape de la recherche avec IA permet d'obtenir de meilleurs résultats.

Compréhension des requêtes

Catégorisation des requêtes par l'IA
Aucune configuration pour les règles
et les synonymes

Récupération d'informations

Hachages neuronaux
Mots-clés populaires pour le langage libre

Classement

Fusion des résultats par mots-clés et par vecteurs
Personnalisation avec l'apprentissage par classement
(Learning to Rank)

Véritable IA de bout en bout

Algolia NeuralSearch soumet chaque requête à trois étapes afin d'obtenir les meilleurs résultats. Nous parlons de recherche avec IA de bout en bout pour nous différencier des solutions concurrentes qui n'offrent l'IA qu'à des étapes spécifiques du traitement (c'est-à-dire, une « IA de fortune »).

Ces étapes comprennent : la compréhension de la requête, la récupération des résultats et le classement. Chacune de ces fonctionnalités avec IA apporte de la valeur aux résultats finaux.



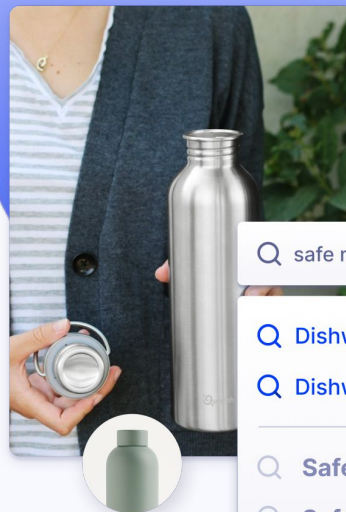
Compréhension des requêtes

La compréhension des requêtes comprend le traitement du langage naturel (NLP) et la compréhension du langage naturel (NLU). Le NLP s'occupe de l'analyse et du traitement du langage dans une barre de recherche afin de préparer une requête pour son exécution. Le processus peut être aussi simple que de comparer la requête telle qu'elle a été écrite au contenu de l'index. La recherche classique par mot-clé est en fait plus avancée que cela, car elle implique la création de jetons et la normalisation de la requête en éléments plus petits, c'est-à-dire en mots et en mots-clés. Ce processus peut être simple (lorsque les mots sont séparés par des espaces) ou plus complexe (comme dans les langues asiatiques, qui n'utilisent pas d'espaces, de sorte que la machine doit reconnaître les mots).

Le NLU se concentre davantage sur la détermination de l'intention de l'utilisateur.

Il utilise la classification des intentions par l'apprentissage automatique, la sémantique, les synonymes et l'expansion de l'ontologie pour comprendre les mots et les phrases. Cette étape permet d'améliorer l'interprétation et la compréhension de la requête.

Par exemple, deux requêtes peuvent utiliser les mêmes termes pour signifier des choses très différentes : « Pantalon pour femme **enceinte** » et « **Enceinte** Bluetooth ».



safe metal b |

Dishwasher safe metal bottle

Dishwasher safe metal flask

Safe metal box

Safe metal pot

Dans le premier cas, le terme « enceinte » est utilisé comme adjectif pour un pantalon, alors que dans le second cas, il s'agit d'un nom désignant un haut-parleur ou des enceintes destinées à diffuser de la musique. Dans ces exemples, la création de jetons par NLP permet de décomposer chaque phrase et le NLU aide à interpréter le sens de chaque requête. Cela vaut pour la recherche textuelle comme pour la recherche vocale, mais cette dernière peut s'avérer plus délicate avec des homophones tels que « vert », « verre » ou « vers », par exemple.

Recherche sémantique

La recherche sémantique nécessite toutefois une approche différente. Elle est généralement alimentée par de Grands Modèles de Langage (Large Language Models ou LLM) afin de créer des vecteurs. Les vecteurs sont des représentations mathématiques de mots, d'images, de vidéos ou de tout autre type d'objet que vous souhaitez comparer. Les modèles d'apprentissage automatique déterminent les relations sémantiques entre les objets d'un index. Les vecteurs peuvent avoir des centaines, voire des milliers de dimensions, qui peuvent être utilisées pour déterminer la similarité. Il est difficile de visualiser autant de dimensions, c'est pourquoi nous utilisons la plupart du temps des représentations tridimensionnelles, comme illustré ci-contre.

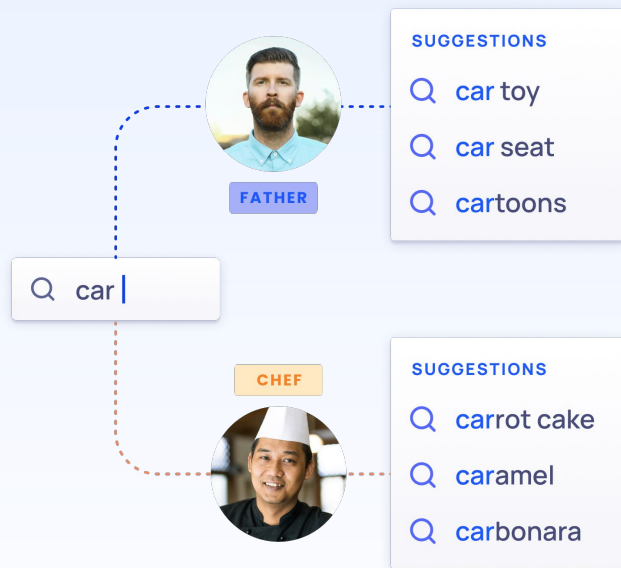
La recherche vectorielle permet d'établir des liens entre les mots, les vecteurs similaires sont ensuite regroupés.

Les vecteurs peuvent être combinés (ou soustraits) pour créer un nouveau sens et comprendre les relations. L'exemple classique est « Roi - homme = reine ».

Relations sémantiques



Relation syntaxique



Récupération de résultats

Une meilleure récupération de résultats peut améliorer à la fois la **précision** de la recherche (exactitude des résultats) et le **rappel** (exhaustivité des résultats sur l'ensemble de l'index). Il est difficile de trouver un équilibre entre la précision et le rappel. L'amélioration de l'un peut avoir l'effet inverse sur l'autre.

Les moteurs de recherche par mots-clés ont tendance à fournir de meilleurs résultats pour les termes principaux, tandis que la recherche vectorielle basée sur l'IA fonctionne bien pour les requêtes longues moins populaires. Vous devez résoudre les deux, ce que l'on appelle parfois la **recherche hybride**.

La recherche par mot-clé s'appuie généralement sur une correspondance statistique à l'aide des méthodes de pondération BM25 ou TF-IDF et des techniques connexes pour classer les résultats en fonction des termes de la requête qui apparaissent dans chaque document. Par exemple, la méthode TF-IDF s'intéresse à la fréquence inverse d'un mot dans un document (IDF) par rapport à la fréquence d'un terme (TF) pour déterminer son importance. Les stop words comme « le », « et » et « ou » apparaissent à tout bout de champ, alors que des mots comme « brosse à dents » ou « eau » apparaissent moins fréquemment, ils sont donc plus rares et informent probablement davantage sur le contenu. La fréquence des termes peut indiquer l'importance ou la pertinence du document. La plupart des modèles nécessaires à l'élaboration d'une recherche par mot-clé sont disponibles. De nombreuses entreprises utilisent encore Lucene, le vénérable projet Apache qui a donné naissance à Elasticsearch et Solr.

Les vecteurs sont énormes, nécessitent généralement des bases de données avec des processeurs GPU spécialisés, et sont coûteux à faire évoluer tout en restant performants. Des techniques telles que HNSW (Hierarchical Navigable Small World), IVF (fichier inversé) ou PQ (quantification des produits, une technique visant à réduire le nombre de dimensions d'un vecteur) comptent parmi les méthodes les plus populaires de recherche de similarité entre vecteurs par approximation des plus proches voisins (ANN). Chaque technique est dédiée à l'amélioration d'une propriété de performance particulière, telle que la réduction de la mémoire avec PQ ou des temps de recherche rapides, et précis, avec HNSW et IVF. Le mariage de plusieurs composants permet de produire un indice « composite » afin d'obtenir des performances optimales pour un cas d'utilisation donné.

Même avec la plupart des techniques d'approximation des plus proches voisins (ANN), il n'y a pas **de moyen facile de concevoir un algorithme de recherche vectorielle** qui fonctionne pour la plupart des applications de production. Par exemple :

- Les fonctions d'insertion, de mise à jour et de suppression peuvent remettre en question les structures basées sur les graphes, telles que HNSW, qui **rendent la suppression très complexe**.
- Le partitionnement des index vectoriels **est très coûteux**. Les performances de HNSW diminuent de manière non linéaire avec l'augmentation du nombre de partitions.
- **Il est très difficile de combiner les résultats des mots-clés et des vecteurs** car le calcul de la pertinence dans chaque système est complètement différent, et il est donc difficile de les comparer. C'est pourquoi de nombreuses personnes ajoutent un index vectoriel ANN à un moteur de mots-clés. Les vecteurs sont ainsi utilisés uniquement comme solution de repli pour proposer des résultats lorsque les mots-clés ne suffisent pas.

Chez Algolia, nous sommes les pionniers d'une nouvelle technique

appelée hachage neuronal. Nous utilisons des réseaux neuronaux pour compresser les vecteurs en minuscules fichiers binaires qui conservent la plupart des informations. Le vecteur binaire est beaucoup plus petit et plus rapide à calculer que les vecteurs standards, ce qui nous permet d'exécuter la recherche par vecteur et par mot-clé simultanément et à grande échelle sur des CPU standard. Nous normalisons et combinons ensuite les résultats afin d'obtenir une mesure unique de la pertinence pour une même requête.

Ces compromis, auxquels s'ajoutent les coûts associés au déploiement de la recherche vectorielle, font que les entreprises n'utilisent pas l'IA pour définir la pertinence des résultats.

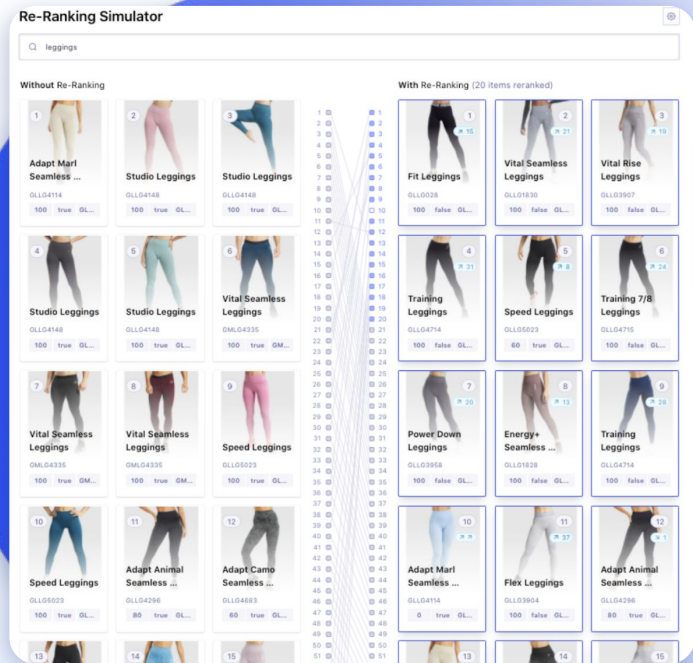
Les solutions de recherche pour sites web et applications ne représentent qu'une des applications des vecteurs. Il convient de mentionner ici que les vecteurs peuvent également être utilisés pour comparer de nombreux types d'objets, tels que des images, des vidéos, des symboles ou même des documents entiers. Ils ont de nombreux cas d'utilisation dans le monde réel, tels que les systèmes de recommandation de films, la reconnaissance faciale, la détection des fraudes, etc.

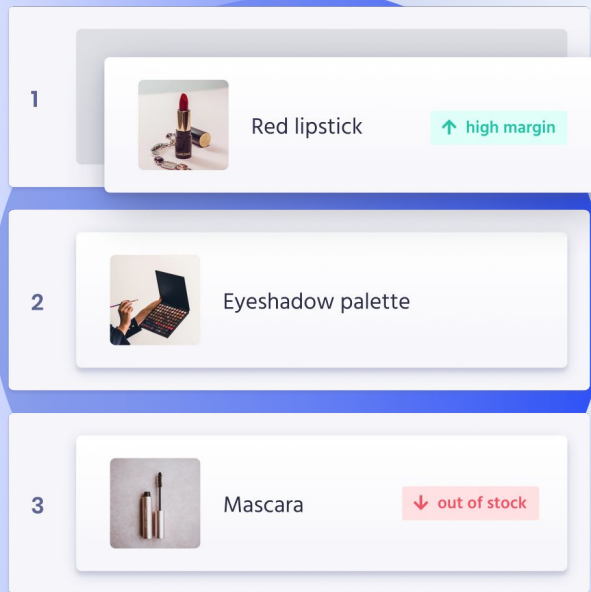


Classement

Le classement des produits peut faire une grande différence dans la performance d'une solution de recherche. Il existe un biais de position dans le classement des résultats, les trois premiers résultats obtiennent beaucoup plus de clics que les résultats moins bien classés. Il est également nécessaire de modifier constamment le classement des résultats. Les résultats qui sont bien classés aujourd'hui devront peut-être être déclassés demain si quelque chose de plus pertinent ou de plus populaire devient disponible.

Si votre récupération d'informations est bonne (voir ci-dessus), le classement commence à partir de meilleurs résultats. L'apprentissage par classement (Learning to Rank ou LTR), parfois également appelé classement par apprentissage automatique (MLR), est un type d'apprentissage automatique qui améliore le classement et contribue à la précision. Il comprend l'apprentissage supervisé, l'apprentissage non supervisé et l'apprentissage par renforcement. Il existe également des variantes telles que l'apprentissage semi-supervisé. Chacune de ces solutions repose sur le classement avec IA, qui permet d'améliorer les résultats par rapport aux méthodes statistiques plus simples.





Si l'apprentissage automatique semble être un défi plus facile à relever que celui de la pertinence, ce n'est pas forcément le cas. Il a fallu deux ans à AirBnB et une équipe de data scientists et d'ingénieurs très talentueux pour mettre au point une solution de classement par l'IA !

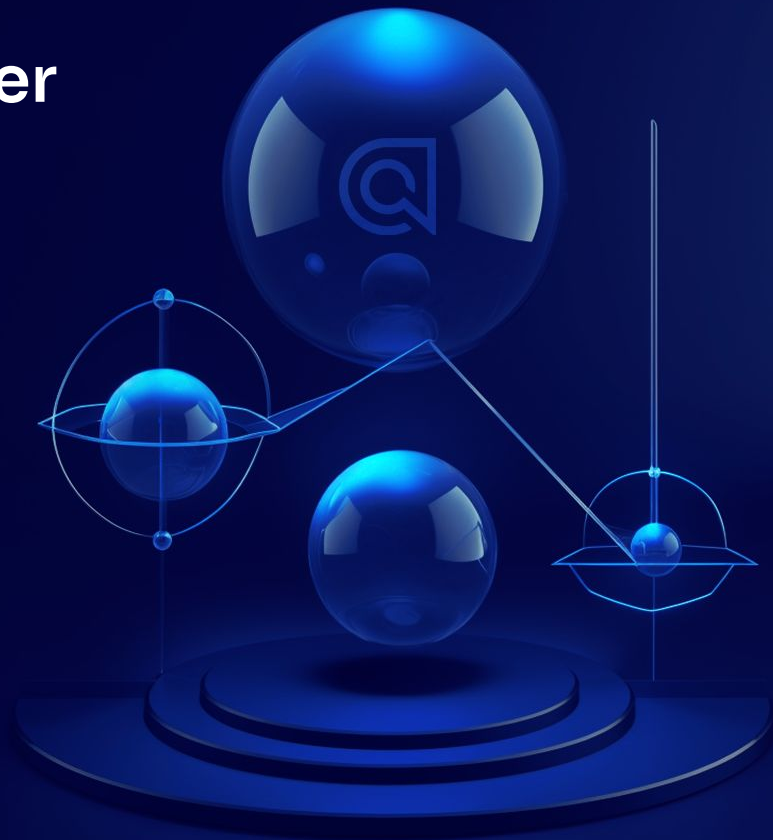
Ces dernières années, les techniques d'apprentissage par classement, telles que l'apprentissage par renforcement, sont devenues plus populaires parce qu'elles réduisent la maintenance et l'itération.

L'apprentissage par renforcement utilise des événements tels que les clics, les conversions, les achats, etc., pour reclasser les résultats du plus populaire au moins populaire. Il peut également avoir l'effet inverse, c'est-à-dire oublier les résultats qui ne sont pas performants. Le défi du classement réside dans les données. L'apprentissage par renforcement nécessite normalement beaucoup de données pour fonctionner, mais de nouveaux modèles apparaissent qui permettent de faire de meilleures prédictions avec moins de données.

Faut-il développer ou acheter une solution recherche avec IA de bout en bout ?

Chaque fois que vous avez été confronté à la décision d'acheter ou de développer en interne, votre équipe d'ingénieurs talentueux vous a dit : « Aucun problème, on peut le faire. » en expliquant « Pourquoi dépenser des centaines de milliers d'euros pour un fournisseur SaaS alors que notre équipe peut déployer une solution open source sur AWS et l'exploiter elle-même ? » Et cela a généralement fonctionné, permettant à l'entreprise d'économiser de l'argent qu'elle peut réinvestir dans l'embauche de nouveaux ingénieurs.

L'IA change tout.



Les 6 principales raisons pour lesquelles vous devriez acheter votre solution avec IA plutôt que de la développer vous-même

1. Talent

Il est très difficile de constituer une équipe d'excellents data scientists

Recruter une équipe de data scientists talentueux est un vrai défi. La course aux talents dans le domaine de l'IA a fait grimper les coûts de recrutement et de fidélisation du personnel, en particulier pour ceux issus des meilleures écoles et universités.

Les budgets de votre entreprise sont-ils aussi larges que ceux des géants technologiques et votre entreprise peut-elle rivaliser avec Google ou OpenAI?

Gagner la guerre des talents est un investissement significatif qui peut être souvent hors de portée de votre budget.

2. Vitesse

Il faut beaucoup de temps pour développer des compétences en science des données.

Il faut beaucoup de temps pour constituer une équipe à partir de zéro. Les délais d'embauche et l'intégration de data scientists dans une équipe prend du temps, car la complexité du travail et des données demande d'expérimenter pour acquérir de l'expérience.

Il faut encore plus de temps pour faire mûrir une nouvelle compétence comme la science des données au sein de votre organisation : il s'agit de créer des outils, des processus et des fonctionnalités qui n'existaient pas auparavant. Il faut compter 2 ans ou plus entre la décision d'embaucher et la mise en production des premiers produits. Votre entreprise peut-elle attendre aussi longtemps ?

3. Culture

Il faut acquérir une culture d'expérimentation scientifique

Pour réussir à développer une IA, il faut une culture qui favorise l'expérimentation scientifique. **L'apprentissage en production est la seule façon de construire et d'optimiser les modèles.** Cette culture est très différente de celle du développement logiciel, où la solution (et les délais de livraison) peut souvent être planifiés dans les moindres détails avant qu'une ligne de code ne soit écrite.

Pour l'IA, il faut des échecs (répétés) en production, et on ne sait jamais si la prochaine expérience fonctionnera ou non. De nombreuses entreprises ne sont tout simplement pas en capacité de développer cette culture d'expérimentation scientifique.

Les 6 principales raisons pour lesquelles vous devriez acheter votre solution de recherche avec IA plutôt que de la développer vous-même (suite)

4. Données

Vous n'avez pas assez de données pour être compétitif

Vous avez besoin de beaucoup de données pour entraîner des modèles d'IA à partir de zéro.

Il est fort probable que vous ne disposiez pas d'autant de données qu'un fournisseur capable d'exploiter les données de milliers de clients pour créer un modèle nettement plus performant que le vôtre.

Sans données à grande échelle, il faut beaucoup de temps pour tester vos nouvelles expériences et vos algorithmes ne s'amélioreront que lentement. Les données sont l'élément vital d'une équipe d'IA et la plupart des entreprises n'ont tout simplement pas l'envergure nécessaire pour rivaliser.

5. Agilité

L'IA évolue très rapidement, pouvez-vous suivre ?

Le secteur de l'IA évolue très rapidement. Si vous choisissez de développer vos propres fonctionnalités, pourrez-vous suivre les évolutions rapides de l'IA ? Vos équipes peuvent-elles se tenir au courant des derniers documents de recherche, assister aux conférences et développer votre solution d'IA en même temps ?

En choisissant d'acheter plutôt que de développer, vous pouvez évoluer plus rapidement et tirer parti des connaissances de pointe des équipes d'IA de vos fournisseurs.

6. Fiabilité

Il est difficile de faire fonctionner des systèmes d'IA en production

Il est extrêmement difficile de faire fonctionner des systèmes d'IA en production. **Ils nécessitent des pipelines de données à grande vitesse et en temps réel, ainsi que de grandes quantités de ressources informatiques.** En cas de pic de la demande, êtes-vous en mesure d'augmenter rapidement votre production et de la réduire avant que les coûts ne s'accumulent ? Disposez-vous d'ingénieurs spécialisés en fiabilité (en anglais reliability engineers) qui savent comment exécuter et contrôler la qualité des modèles d'IA ?



L'avantage du fournisseur d'IA

Les fournisseurs peuvent regrouper la demande de plusieurs clients et répartir la charge avec des pics se produisant à des moments différents. Les fournisseurs peuvent embaucher des ingénieurs spécialisés dans la fiabilité pour faire fonctionner les systèmes d'IA à grande échelle.

Même Microsoft, une entreprise qui dispose des ressources et des talents nécessaires pour développer l'IA, a choisi d'acheter une partie de sa solution d'IA pour Bing.

Acheter une solution d'IA plutôt que de la développer soi-même est un avantage concurrentiel. Vous pouvez égaler le talent et les performances des géants de la tech tout en améliorant la compréhension et la fiabilité des systèmes pour rester compétitif à l'ère de l'IA.

ChatGPT vs Recherche

Depuis qu'OpenAI a annoncé le lancement de ChatGPT et que les gens ont commencé à le tester, nombreux sont ceux qui ont annoncé l'aube d'une nouvelle ère. ChatGPT allait révolutionner la recherche.

Mais, est-ce que c'est vrai ? Non. Et voici pourquoi.

Formulation des requêtes

Les réponses de ChatGPT dépendent fortement de la formulation de la requête (prompt). OpenAI indique qu'il s'agit d'une « limitation ». Les fautes de frappe et les petits écarts de langage peuvent faire échouer ChatGPT. La recherche doit comprendre les utilisateurs, même s'ils font des fautes d'orthographe, sinon c'est un retour en arrière. Il est difficile de comprendre et de faire correspondre les différentes orthographes.

Expérience

Les schémas de recherche sont généralement les mêmes depuis des décennies. Plus précisément : un jeu de résultats classés du plus pertinent au moins pertinent (quelle que soit la manière dont on le mesure). Il est important que les personnes qui font les recherches puissent avoir des options, puissent filtrer par un clic et, d'une manière générale, puissent entrer dans une « phase de découverte ». ChatGPT n'est pas fait pour cela.

Rapidité et coût

La rapidité est importante dans le domaine de la recherche. Amazon et Google ont tous deux réalisé des études montrant que des décalages de 100 millisecondes seulement peuvent coûter des millions de dollars. ChatGPT sera un jour plus rapide, mais pour l'instant, on en est loin.

En résumé : ChatGPT est excellent, en particulier pour les cas d'utilisation du commerce conversationnel, mais pour une expérience prévisible et rapide, les clients choisiront toujours la recherche.

Faites passer votre site web au niveau supérieur.

Obtenez une démo personnalisée de nos experts
en recherche avec IA dès aujourd'hui.

DEMANDER UNE DÉMO

www.algolia.com/fr

