Collection des essentiels

Amarc
ASSOCIATION POUR LE MANAGEMENT DE LA RÉCLAMATION CLIENT

inetum.
Positive digital flow

L'apport des **IA génératives** au management de l'insatisfaction client

Laurence Houdeville &
Jean-Paul Muller

L'IA au service du management de l'insatisfaction client

Comment inculquer de l'émotion à une IA ? L'IA pourrait-elle disposer d'un esprit critique ? Une IA générative peut-elle valider la pertinence et la véracité de ce qu'elle produit ? L'arrivée de l'IA dans nos vies suscite de nombreuses questions. Pour répondre à plusieurs d'entre elles, l'Amarc (Association pour le Management de la Réclamation Client) et Inetum se sont rapprochés pour comprendre le phénomène.

Des managers impliqués dans leurs organisations, tous spécialistes de la relation client, nous ont aidés à intégrer des cas concrets. Au-delà des entretiens, des tests ont été réalisés avec eux afin d'explorer le potentiel des IA génératives et de tenter de répondre à la question « **Quel est l'apport des IA génératives au management de l'insatisfaction client ?** ».

Ce guide enrichi de leurs témoignages vise à rassembler les informations à connaître avant de se lancer dans un projet d'IA.

Amarc | inetum. Positive digital flow

Sommaire

L'IA générative, une révolution en marche

1

Les IA génératives, Kesako ?

Qui n'a pas entendu parler de ChatGPT ? Introduit par OpenAI à la fin de l'année 2022, ce chatbot a créé une rupture technologique, véritable onde de choc dont les impacts sont observables dans un grand nombre de métiers. La réclamation client n'est pas en reste. Au collaborateur soucieux d'apporter une réponse juste et dans les meilleurs délais à un client mécontent, vient s'ajouter un conseiller virtuel dont le rôle pourrait être appelé à s'étendre.

Derrière ChatGPT se trouve un modèle de langage (GPT 4) qui s'est vu rapidement concurrencé par d'autres modèles (Bard, Llama, Alpaca, Claude, Hugginface).

Ces modèles entrent dans la famille des LLMs (Large Language Models) dont les usages sont de trois natures :

- La création d'informations en langage naturel
- La recherche d'informations et sa restitution
- L'analyse sémantique pour identifier la proximité entre plusieurs contenus

Il est important de rappeler que la démocratisation des IA génératives n'étant pas dénuée de risques (disparition de certains métiers, divulgation d'informations sensibles, plagiat, problèmes éthiques : propagation de stéréotypes…), elle nécessite de réfléchir à des formes de régulation. Un nouveau cadre normatif, l'AI Act, voit le jour en Europe. Il répond à des enjeux de confiance et s'impose à toute IA qui se déploiera, demain, sur le territoire européen.

IA génératives : comment ça marche ?

Nous aborderons ici quelques composants qui permettent de comprendre le fonctionnement des IA génératives dans le domaine de la réclamation client.

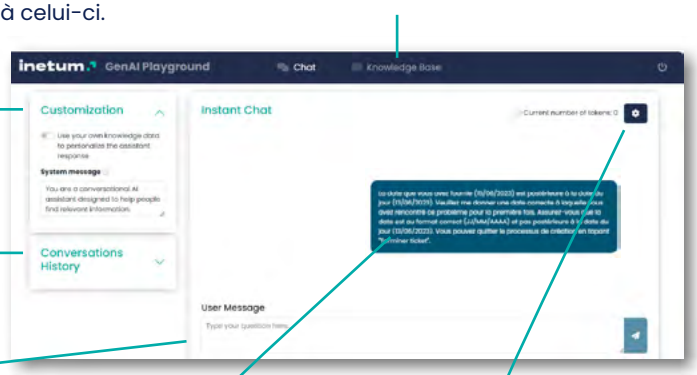
En premier lieu les **GAN** (Generative Adversarial Networks). Ces réseaux de neurones produisent du contenu et font intervenir simultanément un générateur de contenu et un discriminant qui valide si le contenu est réel. Basé sur ces deux piliers, le modèle s'améliore au fil de l'entraînement.

Autre composant essentiel des IA génératives les **VAE** (Variational Auto-Encoders). Ils disposent d'un réseau de neurones proposant en entrée un encodage de la donnée et en sortie une reconstruction de la donnée à partir de son encodage (décodeur). Cela permet de disposer, dans un même espace de toutes les données d'entrées encodées et de pouvoir les utiliser.

Sans être exhaustifs, nous aborderons en dernier lieu, les **Transformers.** Utiles principalement pour la traduction et la génération de texte (Texte-To-Speech ou Speech-To-Text), ils forment un ensemble de modèles de Deep Learning qui appartiennent à une classe spéciale d'architectures de réseaux de neurones récurrents. Ces modèles apprennent, via un mécanisme d'attention, ce qui est important et ce qui ne l'est pas. Ils observent chaque séquence et décident quelles parties de cette séquence sont à conserver.

Comment utiliser un playground (espace de travail) ?

L'utilisateur peut personnaliser son expérience via les onglets de personnalisation **(Customization)** et de base de connaissances **(Knowledge Base)**, permettant d'introduire du contenu propre à l'entreprise et ainsi de contextualiser la réponse du modèle par rapport à celui-ci.



Gestion de l'historique des demandes. Cet historique peut-être utilisé par l'IA générative pour compléter, ajuster ses réponses au fil de la conversation.

Le prompt : instruction ou question pour orienter l'IA dans sa réponse. Les prompts en entrée fournissent des informations aux IA génératives alors que les prompts de sortie définissent les exigences attendues pour la réponse apportée par l'IA.

Espace de chat. Les conversations apparaissent avec les jalons d'historisation (dates).

Dans les modèles de langage, **les tokens** sont la première opération pour transformer le texte dans une structure numérique que la machine pourra manipuler par la suite. Dans le cas de la génération de texte, les modèles produisent des tokens, et non du texte directement. De fait, plus le texte est long ou complexe, plus il y a de tokens, c'est pour cette raison qu'il est toujours indiqué un nombre de tokens géré au niveau des modèles de LLMs. Ils représentent la taille du prompt qu'on peut leur envoyer, c'est-à-dire, la taille globalement du contexte qu'ils sont capables d'analyser. Un même LLM peut avoir des versions différentes uniquement basées sur le nombre de tokens qu'il est en capacité de gérer. Voir le lexique en page 15.

Le périmétrage

Ce qu'on désigne par « périmétrage » est l'action d'injecter un ensemble d'informations directement dans le prompt, parmi lesquelles le modèle va pouvoir trouver les informations nécessaires à la formulation de sa réponse, en précisant explicitement à ce dernier qu'il ne doit utiliser que cet ensemble d'informations. En agissant ainsi, on supprime tout ce que le modèle aura appris en termes de contenus, mais pas en termes de capacité de tâches (réponse, résumé, extraction d'information, etc.), en lui imposant un « périmètre » pour construire sa réponse.

Cela nécessite, en amont du **LLM,** d'aller chercher les informations en fonction de l'intention de l'utilisateur, afin de l'intégrer dans le prompt contenant ladite intention.

Cette étape peut s'appuyer sur une vectorisation de la base documentaire (« référentiel de vérité ») découpée en petits morceaux de texte **(chunks)** et complété d'un système de mesure de similarité sémantique, ou directement sur un moteur de recherche sémantique.

Cette méthode pour la construction des réponses d'un LLM est appelée **RAG** (Retrieval Augmented Generation).

Le prompting à la loupe

Le prompt engineering peut se définir comme une simple demande et/ou une série d'instructions que l'on fournit à un modèle d'IA générative pour initier une conversation et générer des résultats spécifiques. Celui-ci se compose a minima des éléments suivants, pouvant varier selon les modèles :

Persona du répondant et situation
Instruction au LLM désignant le persona qu'il doit adopter pour répondre

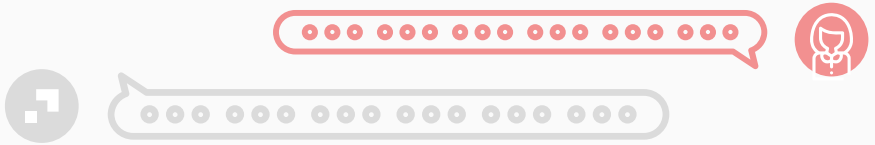
Mémoire
Contient les échanges précédents dans le cas d'un dialogue

Périmètre source
Ensemble des données sources à partir desquelles le modèle construira sa réponse (et pontentiellement des exemples pour l'aider dans sa réponse)

Tâche
Correspond à l'intention de l'utilisateur

Contraintes de réponse
Élements permettant de forcer certaines réponses dans les cas décrits

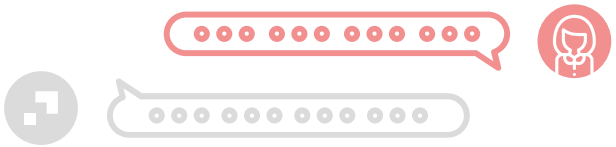
Spécifications du format de sortie
Description du formattage attendu de la réponse



Les règles du prompting

Thèmes	Composants d'un prompt	Exemples
Persona, Compétences rôles	Indiquer quel persona, quel rôle doit jouer l'agent virtuel. Fournir les compétences, expertises particulières sur le domaine que l'agent doit connaître.	Tu es le Responsable administratif d'un groupe bancaire international. Spécialiste des marchés financiers, des fusions acquisitions.
Périmètres traités	Indiquer les domaines couverts, fournir les indications complémentaires sur le ou les domaines qui peuvent guider la génération.	Tu ne dois répondre qu'aux demandes relatives au domaine de....
Ton, style, angle	Adopter un ton spécifique adapté au public cible : officiel, professionnel, humoristique, amical, enthousiaste, attentionné, châtié, convaincant, etc. Orienter le mode de conversation vers le vouvoiement ou le tutoiement. Orienter la réponse avec un point de vue particulier. Utilisation d'éléments de langage, de mots-clés, de termes spécifiques, utiliser des analogies, des tournures spécifiques...	Tu utiliseras un langage technique en fournissant le maximum de détails. Tu utilises systématiquement le vouvoiement avec l'utilisateur. Tu te mettras toujours dans une position d'observateur, tu ne porteras aucun jugement. Tu utiliseras des analogies.
Format	Réponse courte/longue, contraintes de mots, de phrases, de paragraphes.	Si tu as de nombreuses informations, utilise des listes. Tu dois répondre avec un maximum de 2 phrases de moins de 20 mots.
Fournitures d'exemples	Intégrer des exemples de discussions entre l'agent et l'utilisateur qui orienteront le modèle dans sa génération.	Tu es le responsable administratif d'un groupe bancaire international. Spécialiste des marchés financiers, des fusions acquisitions.
Langue	Indiquer les contraintes de langue.	Quelle que soit la langue de l'utilisateur, tu répondras en français.

Thèmes	Composants d'un prompt	Exemples
Contraintes liées au contenu de la base de connaissance	En fonction des contenus de la base de connaissance, il peut, parfois, être intéressant d'intégrer des conditions additionnelles.	Aucune URL ou lien externe ne doit figurer dans ta réponse. Si tu le peux, utilise le moins possible les sigles dans ta réponse.
Format	Réponse courte/longue, contraintes de mots, de phrases, de paragraphes.	Si tu as de nombreuses informations, utilise des listes. tu dois répondre avec un maximum de 2 phrases de moins de 20 mots.
Certitude	Orienter la réponse en assurant une plus grande fiabilité.	En fonction des informations dont tu disposes, si tu as un doute sur certains aspects de la réponse, indique-le.
Suggestions	Proposer des exemples de demandes pour un « second tour » de réponses.	La prochaine de mes questions pourrait être…
Informations personnelles	Informer l'utilisateur lorsqu'il intègre dans sa demande des informations qui semblent à caractère personnel.	En complément de ta réponse, avertis l'utilisateur qu'il doit….
Informations sensibles	Gérer les propos relatifs à la religion, à la spiritualité, les actualités / liés à des positions politiques, les discussions autour des sentiments, de la sexualité, etc. De manière générale, prendre en compte que les demandes dites « chitchat » sont nombreuses dans les conversations avec les agents.	
Insultes / sujets hors propos	Gérer les termes inappropriés, la manière de répondre à des demandes insultantes.	



Chiffres clés :

ChatGPT

2 **milliards de visites par mois**

68%

des Français qui utilisent **l'IA générative** dans leur entreprise le cachent à leur supérieur hiérarchique.

68%

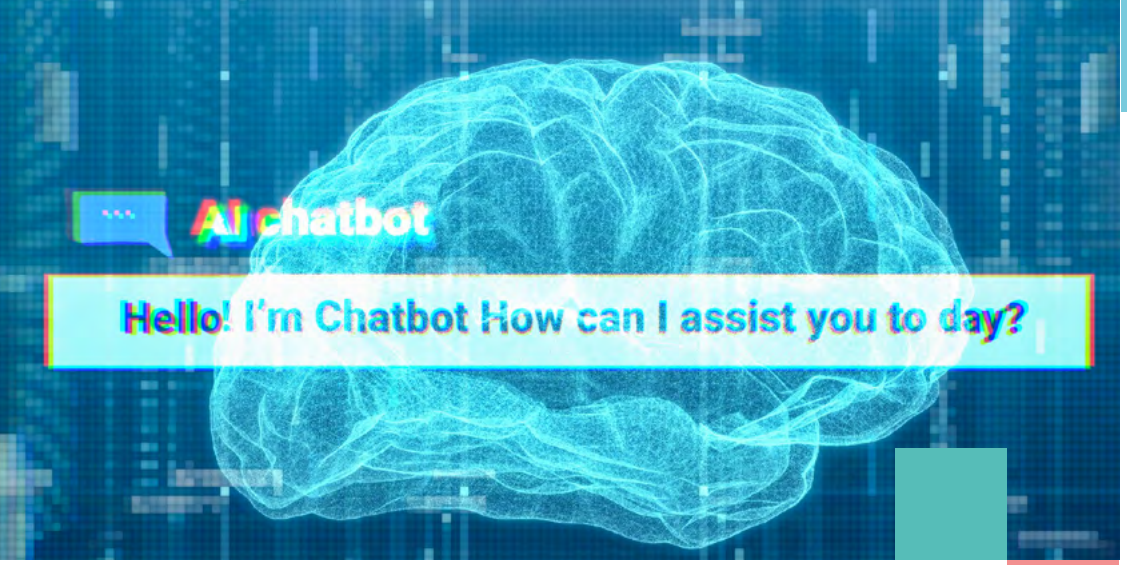
ont des craintes vis-à-vis de l'émergence des **IA génératives.**

72%

des personnes interrogées estiment **ne pas avoir** de connaissances suffisantes pour l'utiliser.

L'entraînement de GPT-3 sur des GPU a nécessité **190 000 kWh,** soit **85 000 kg d'équivalents CO_2,** ce qui équivaut à parcourir environ **700 000 kilomètres en voiture**, ou encore un **aller-retour Terre-Lune.**

128 000 tokens est la taille du **prompt** du dernier modèle en date d'**OpenAI**, **GPT4 Turbo**, soit l'équivalent d'un livre standard.



Lexique :

Intelligence Artificielle : domaine d'étude ayant pour objet la reproduction artificielle de facultés cognitives de l'intelligence humaine dans le but de créer des systèmes ou des machines capables d'exécuter des fonctions relevant normalement de celle-ci.

IA générative : est un type d'intelligence artificielle capable de créer de nouveaux contenus (texte, image, son) et de nouvelles idées.

GPT (Generative Pre-trained Transformer) : est un outil conversationnel développé par OpenAI capable de générer du contenu écrit en réponse à des prompts.

Modèles de fondation : modèles d'IA entraînés à partir de données massives et pouvant servir des usages différents.

LLM (Large Language Models) : modèles de fondation générant du texte à partir d'autres données (texte, image, son).

Hallucination : les hallucinations font référence au modèle générant des sorties syntaxiquement et sémantiquement correctes mais déconnectées de la réalité et basées sur de fausses hypothèses.

Un token dans les modèles de langage est une unité de base qui représente une partie du texte. Les tokens sont généralement des mots ou des caractères individuels. Par exemple, dans la phrase «J'aime les pommes», il y a cinq tokens : « J' », « aime », « les », « pommes », et le symbole de fin de phrase.

Le découpage en tokens est plus complexe qu'un simple découpage par les espaces ou les tirets car il y a de nombreux cas où des groupes de mots peuvent ne représenter qu'un token car ils représentent un concept unitaire (la pomme de terre, 3 mots, 1 seul concept). À l'inverse il peut être intéressant de découper certains mots en plusieurs tokens, notamment au niveau des préfixes ou des suffixes qui représentent en tant que tels des concepts unitaires, comme la négation (im-probable) ou la taille (tarte-lette), etc. De fait, le nombre de tokens pour une phrase est totalement dépendant de la langue.

Sources :

Explodingtopics.com
http://tinyurl.com/yck9ysse

Sondage Ifop (leptidigital.fr)
http://tinyurl.com/yc47h8wx

The Register – AI me to the Moon... Carbon footprint for 'training GPT-3' same as driving to our natural satellite and back
http://tinyurl.com/3ps5w2mz

Linc (cnil.fr) - Dossier IA générative - Quelles régulations pour la conception des IA génératives ?
http://tinyurl.com/5t57ywnt

Sustainable AI – Environmental Implications, Challenges and Opportunities
http://tinyurl.com/53m7cyjm

Office Québécois de la Langue Française – « Une intelligence artificielle bien réelle »
http://tinyurl.com/y6wkwkjp

Que sont les hallucinations LLM? Causes, préoccupation éthique et prévention
http://tinyurl.com/4n7pesk2

Quelle projection dans 5 ans ?

« La vision à 5 ans
embarque les objectifs qui nous sont assignés à savoir une baisse des frais généraux sans impact sur la qualité de nos services. Cet objectif ne peut passer que par plus d'automatisation. »

« En matière d'insatisfaction,
nous devons mieux détecter, catégoriser, traiter la réclamation. Ceci en apportant une réponse fiable, rapide, complète et empathique. »

2

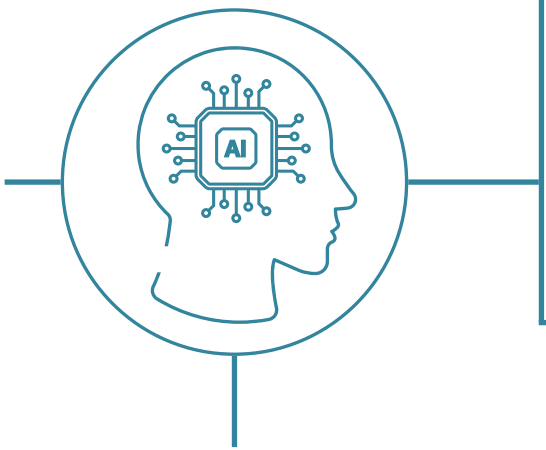
Projection dans 5 ans : un agent virtuel doté de sentiments

Les profils seront nécessairement plus experts. Les postes répétitifs seront laissés aux IA génératives. L'automatisation prendra en charge le récurrent. Les personnels engagés dans la relation client devront nécessairement développer une plus grande intelligence situationnelle afin de mieux préparer et encadrer les IA.

La vision dessinée par les entreprises interrogées fait apparaître un agent virtuel doté de sentiments, véritable coach des conseillers physiques. Les enjeux cités par les collaborateurs sont de plusieurs natures :

Evolution du SI

Pour industrialiser des cas d'usage intégrant des IA génératives, dans le domaine de la réclamation client il faut au préalable connaître ses données, les applicatifs et bases de données qui les hébergent. La labellisation des données, la constitution d'un dictionnaire, la gestion des référentiels, l'uniformisation et la mise en qualité de la base clients, le réalignement des processus métiers, sont autant de chantiers qui accompagnent l'urbanisation du SI, enjeu important en amont de tout projet d'IA.



Un agent virtuel

Cet agent virtuel agirait comme un « coach » qui libérerait l'agent physique de multiples petites tâches à réaliser et lui permettrait d'être plus performant (ex : réduction des temps d'attente au téléphone via la recherche anticipée de documents / éléments de réassurance déterminés en fonction du profil psychologique).
L'agent virtuel intégrera un parcours en 3 étapes :

- Attribution d'un rôle
- Formations à suivre
- Insertion de l'agent virtuel dans le quotidien de l'agent physique.

Sentiment analysis

C'est probablement l'enjeu majeur des années à venir. Le sentiment fait référence aux pensées et attitudes, tenues ou exprimées, motivées par des émotions. L'extraction et l'analyse des sentiments des clients à partir des commentaires et des avis (CEM) va permettre d'établir des réponses appropriées à chaque client en fonction des émotions collectées. L'IA générative proposera un vocabulaire spécifique afin de rassurer le client « mécontent ».

Frugalité

Les IA frugales anticipent et optimisent la quantité de données, la création des algorithmes, le choix du matériel, la source d'énergie utilisée afin de réduire au maximum l'impact environnemental de l'IA sans amoindrir sa performance. Pour les IA génératives la consommation de la phase d'inférence (phase d'utilisation) peut être supérieure à celle de la phase d'entraînement. Dans une étude publiée par META il est montré que pour un modèle de langage utilisé plusieurs milliards de milliards de fois par jour, pendant deux ans, la phase d'apprentissage représente 35% de l'impact environnemental tandis que la phase d'inférence en représente 65%.

Sécurité / Privacy

Dans certains secteurs en matière de conformité, la Protection de l'Intérêt du Client (PIC) est une règle de bien vendre sur laquelle les entreprises prennent des engagements.
Plus globalement, pour assurer la sécurité des données et leur conformité RGPD, il est nécessaire de délimiter le périmètre des traitements réalisés. Les enjeux sont à deux niveaux :
- Lors de la conception du modèle de langage qui utilise un grand nombre de données. Les mesures et garanties en place sont-elles suffisantes pour se passer du consentement des personnes concernées ? C'est le principe de licéité qui est questionné.
- L'utilisation de LLMs qui traitent les données des utilisateurs et peuvent les réutiliser. Là encore, le RGPD encadre les conditions de réutilisation des données fournies par les utilisateurs, à travers une série de droits accordés aux personnes, mais aussi à travers d'autres obligations que le responsable du traitement doit anticiper. Le RGPD impose au Responsable de Traitement de démontrer sa conformité (principe d'accountability). Ce sujet est aussi lié à la nécessaire analyse d'impact sur la Protection des Données (AIPD).

Pain points, cas d'usages & KPIs

3

4 principaux pain points à résoudre

Spécificités des caractéristiques de l'insatisfaction

Le champ lexical, la tonalité des échanges (impatient, frustré, en colère, exaspéré…) et le temps qu'un conseiller passe avec un client mécontent sont autant d'indices de préoccupation. Les personnes les plus exposées à l'insatisfaction doivent être formées. Les IA génératives sont, dans ce domaine, intéressantes à analyser. Quelle sera leur performance en matière de détection et de classification des tonalités ? Comment introduire dans la réponse une émotion adaptée qui satisfasse le client mécontent ?

Nécessaire acculturation à la donnée

C'est un préalable à tout projet de déploiement massif des IA génératives dans des services dédiés à la relation et réclamation client. La sensibilisation actuelle est encore insuffisante car portée par et pour des spécialistes.

Données massives à traiter

« Le problème majeur rencontré est la quantité de contenus générés. La nécessité d'intégrer des outils de synthèse des attentes du client est devenue essentielle. La génération automatique de pistes de résolution des problèmes est un must sur lequel notre entreprise travaille. » Pourtant d'autres problématiques d'ingestion et de traitement de données massives apparaissent avec les IA génératives.
La quantité et la qualité des données à la source restent la problématique de fond.

Limites des IA génératives

« La véracité des réponses n'est malheureusement pas toujours assurée. L'IA n'a pas d'esprit critique. Nous veillons à ce que l'IA ne déresponsabilise pas nos agents et ne supprime pas la gestion des émotions si utile dans la relation client. » L'apparition successive et rapide de nouvelles générations de LLMs est une autre forme de limite : « Les modèles d'IA évoluent très rapidement et les équipes n'ont pas le temps de s'adapter et de réaliser les tests attendus entre deux versions. L'entreprise tente de définir des exigences et des recommandations qui permettront d'industrialiser des modèles et de partager les résultats obtenus avec les autres entités, avant d'upgrader les IA. »

Cartographie des cas d'usages

Macro use cases	Use cases	Contexte	Indicateurs
Automatisation des tâches répétitives (Improved operating efficiency)	- Classification automatique et routage. - Résolution automatisée de problèmes simples. - Suivi et gestion automatisée des dossiers clients. - Analyse d'images dans des contenus échangés avec des clients.	Aujourd'hui, une des préoccupations principales est la simplification des outils des frontliners. Il apparaît nécessaire de proposer aux agents des solutions dans le contexte de leurs échanges clients pour : - La reconnaissance automatique dans la lecture de documents. - La réalisation de tris. - L'amélioration du temps de réponse aux clients. - La génération d'emails. - L'analyse des images reçues lors d'un sinistre.	- NPS - SAT - CES - Once and Done. - Temps de traitement. - Temps de prise en charge.

Cartographie des cas d'usages

Macro use cases	Use cases	Contexte	Indicateurs
Personnalisation de l'expérience client (Reinvent Customer Management Experience)	- Analyse de l'historique. - Apprentissage continu. - Proposition de contenus.	Il est intéressant ici de mettre l'accent sur plusieurs services : - l'accompagnement du client (selfcare) Exemple : Dans l'espace client, la personne rencontre un souci de remboursement. Lui apporter une réponse personnalisée au-delà d'une simple FAQ. - La détection et l'intégration dans la réponse d'éléments correspondant au moment de vie du client. La détection via une IA générative des moments de vie, souvent difficiles (invalidité, décès), du client est une information importante. Elle permet d'intégrer un facteur de « fragilité » et de générer une alerte, à l'attention du conseiller. Des services particuliers pourront être poussés au client.	- Taux de création de leads. - Pourcentage de selfcare. - Pourcentage de transactions d'assistance sur le digital. **En rebond :** - Taux de rétention. - Churn.

Cartographie des cas d'usages

Macro use cases	Use cases	Contexte	Indicateurs
		Seront réalisées, des synthèses des conversations clients, de la captation de tonalité et de l'insertion, historisation dans le CRM. - La génération de processus. Certaines fonctionnalités permettent de déterminer, prioriser l'action à réaliser immédiatement pour un client. L'objectif est de retranscrire en temps réel les échanges entre un agent et un client et de pousser, en fonction d'une analyse de mots clés, l'information adéquate. Le mot « brochure » est capté, un lien vers le document attendu est généré. Un workflow dynamique est entretenu et permet de nourrir une base de connaissances sur les clients.	

Cartographie des cas d'usages

Macro use cases	Use cases	Contexte	Indicateurs
Évaluation de la satisfaction client (CSAT)	- Analyse des sentiments dans le cadre d'une optimisation des réponses aux clients mécontents. - Optimisation des solutions choisies (chatbot ou conseiller).	On retrouve ici : - L'identification des tonalités « Speech Analytics : analyse de la voix du client en ajoutant une dimension supplémentaire, l'analyse de l'émotion. Cela apporte une compréhension plus fine des motifs d'appels ou des irritants clients ». Seront réalisés : la classification de verbatims, le plan de taggage, la catégorisation des mots et de leur tonalité, et l'analyse des fréquences de contacts. - Des tests de performance des agents. Comparaison faite avec les chatbots.	- Mesure de la CSAT. - NPS - Emotion score

Cartographie des cas d'usages

Macro use cases	Use cases	Contexte	Indicateurs
		Certaines entreprises réfléchissent à l'analyse comparée des résultats d'un échange entre un client et un agent d'une part, et un client et un chatbot, d'autre part. L'objectif est de positionner le chatbot au bon moment de la relation, de mesurer le coût induit par chacune des solutions et leur ROI. Ceci devrait permettre à l'entreprise de mieux positionner la place qu'elle donne aux IA génératives dans le cadre de la relation client.	

Cartographie des cas d'usages

Macro use cases	Use cases	Contexte	Indicateurs
Prévention des réclamations	- Analyse des tendances et des problèmes récurrents.	Faire de la prévention pour générer moins d'insatisfaction. Pour cela veiller à bien analyser les besoins du client, afin de mieux les anticiper et de répondre de façon proportionnée à ses demandes. Peuvent être réalisés ici : - la captation de l'émotion et le contrôle des actions réalisées. L'analyse du contenu non verbal (au-delà de la parole) révèle les émotions telles que la colère ou l'enchantement. Lier ces émotions à des sujets ou motifs d'appels permet de mieux prioriser des actions d'amélioration des processus, des produits et services ou encore des arguments des conseillers.	- CSAT - DSAT - Suivi des volumes et items de SAT et DSAT. - Churn rate. - Taux de ré-ouverture des réclamations. - Taux de ré-appel (first call resolution. Once & Done). - NBO

Cartographie des cas d'usages

Macro use cases	Use cases	Contexte	Indicateurs
		. Cela nous aide aussi à vérifier que le client est bien positionné dans le bon parcours. - Autre action : l'analyse prédictive des comportements des clients et l'historique de leurs achats (Next Best Offer - NBO model).	
Formation des conseillers (training & onboarding)	Optimisation de la performance des conseillers en matière de réclamation, relation client.	Optimiser l'activité de l'équipe chatbot. Mieux appréhender les questions des clients. Analyser les conversations, repérer les usages où le bot ne sera pas compétent. Exemple : Formation Toyota « maximiser l'émotion avec Salesforce Voice ». « La formation réalisée est de bon niveau. Elle permet à l'agent de progresser. Il suit son score d'émotion, apprend à baisser la voix ou à la moduler en fonction des interactions avec son client. »	- Taux de complétude des réponses. - Taux de fiabilité des réponses. - Récurrence des sujets (proportion). - Taux de rétention des collaborateurs. - Taux de satisfaction des collaborateurs (NPE, Net promoter employee).

Cartographie des cas d'usages

Macro use cases	Use cases	Contexte	Indicateurs
Contrôle qualité du service client	Analyse en temps réel des interactions avec les clients et notation sur la qualité des échanges.		- Taux de compliance (mesure de conformité de la réponse).
Anticipation des besoins des clients	- Passer d'un centre de coût à un centre de profit. - Faire de la réclamation un levier de business.		- Taux de proactivité (durée et volume d'appels/ sollicitations des clients).

Solutions existantes et métriques

4

Choix de la solution : les questions à se poser

NB : les questions portent sur les spécificités de chaque solution.

- Quel playground ?
- Quelles IA génératives choisir ? Comment comparer leurs performances ?
- Quels outils de coordination interne ?
- Quelle équipe ?
- Comment recetter un projet ? Quel type de cahier de recettage faut-il prévoir ?
- Quelle enveloppe budgétaire prévoir ? Quelles conséquences sur mes besoins actuels et futurs en matière de stockage des données ?
- Quelle confiance je souhaite donner à mon projet d'IA génératives ? Quel contrôle ? Quels sont les impacts des erreurs du système IA ? Comment les mesurer ?
- Quel est le périmètre de mon projet d'IA génératives ? (Meilleur équilibre entre le gain/bénéfice généré et le coût/risques). Quel est le ROI de mon projet ?

Les 6 étapes à franchir

Formaliser les cas d'usages

Les premiers cas d'usages envisagés à partir d'IA génératives sont ceux qui enrichissent la gestion de la relation client (chatbots) ou sont entraînés sur de la documentation virtuelle pour apporter de l'aide aux conseillers (assistants virtuels).

Cadrer les coûts et le calcul du ROI

Le coût dépend d'un ensemble de paramètres : puissance des modèles utilisés, taille du prompt (nombre de tokens utilisés) et volume de données ingérées pour produire les réponses.

Sélectionner les LLMs pertinents

Le choix du modèle dépend de l'objectif et du niveau de performance à atteindre. Ainsi Google Bard pourra être plus performant dans la compréhension d'un texte. En fonction du cas d'usage, il faut passer en revue l'offre de LLMs.

Définir une stratégie de développement (Open Source ou propriétaire)

Par nature l'open source permet de maîtriser son modèle de bout en bout et réduit le risque de dépendance à une solution. Le temps, les compétences et les coûts nécessaires à l'entraînement du modèle sont à prendre en compte très en amont dans le projet. À l'inverse l'utilisation d'un modèle propriétaire sous forme d'API sera plus rapide mais nécessitera un encadrement contractuel et réglementaire plus important.

Vérifier la conformité aux exigences de « confidentialité des données »

Selon le Laboratoire d'Innovation Numérique de la CNIL, OpenAI a lancé, depuis le mois d'avril 2023, neuf mesures qui s'appliquent à tous les pays de l'Union européenne. « Une information détaillant les traitements mis en œuvre, la possibilité de s'opposer au traitement

de ses données pour les utilisateurs et non-utilisateurs (via un formulaire ou par email), l'introduction de mesures d'effacement des données inexactes (sachant que la correction des données apparaît impossible aujourd'hui), la clarification de la façon dont les données des utilisateurs peuvent être réutilisées à des fins d'amélioration de l'algorithme sans préjudice de pouvoir s'y opposer, la mise en œuvre d'un mécanisme de déclaration de l'âge (avec interdiction pour les utilisateurs de moins de 18 ans sauf pour les mineurs âgés entre 13 et 18 ans bénéficiant du consentement de leurs parents). »

Acculturer les équipes et la direction
La sensibilisation de la Direction Générale, et plus largement des salariés, aux enjeux et aux limites des IA génératives, est nécessaire. Elle peut prendre la forme de modules d'e-learning ou d'ateliers de tests.

Solutions

Avant d'aborder quelques solutions rappelons qu'il est important de définir, à chaque phase, l'outil le plus pertinent en fonction des choix technologiques.

L'exemple suivant, proposé par une entreprise interviewée, fait apparaître à chaque étape, une solution :

Microsoft Open AI → Playground de test

Google AI → Permet de comprendre les différences entre les LLM

CEC Generative AI → Coordination transversale de GenAI (Corporate Expertise center)

De façon générale, les IA génératives les plus utilisées sont : ChatGPT, Character.ai, Google Bard, Poe et Quillbot.

Dans la gestion de la réclamation client on retrouve le plus souvent :

ChatGPT, un LLM spécialisé dans la génération de dialogues. Il est basé sur le modèle GPT-3, mais il a été entraîné sur des données spécifiques de conversations en ligne. ChatGPT peut répondre aux questions des clients, les orienter vers les bonnes solutions, les divertir avec de l'humour ou des anecdotes, etc. ChatGPT peut être utilisé comme un agent virtuel, un chatbot, un assistant vocal, etc. Il s'adapte au ton, au style et à l'humeur du client ce qui peut augmenter sa satisfaction et sa fidélisation.

Midjourney est un LLM spécialisé dans la génération d'images à partir de texte (prompts). Cette technologie est utilisée pour créer des œuvres d'art, des illustrations et des designs. Contrairement à Dall-e (lire ci-après), Midjourney n'est pas entraîné à viser le réalisme, mais plutôt à étendre les pouvoirs de l'imagination humaine. Midjourney a également créé une communauté en ligne où les utilisateurs peuvent partager leurs œuvres d'art et discuter de l'utilisation de l'IA pour la créativité.

Dall-e, un LLM spécialisé dans la génération d'images. Il est basé sur le modèle GPT-3, mais il a été entraîné sur des données spécifiques d'images et de textes. Dall-e peut créer des images originales, réalistes ou fantastiques, en fonction d'une description textuelle. Dall-e peut être utilisé pour illustrer des produits, des services, des concepts, des idées, etc. Dall-e peut aider à enrichir, personnaliser et différencier la communication visuelle, en augmentant l'attention, l'émotion et la mémorisation des clients.

Autres LLMs utilisés : Claude, Deepmind Gopher, ILLUIN Technologies, LightOn et les startups Batvoice AI, Konverso, Seedext, TALKR.ai, AlloBrain.

Sources :

Le Journal du Net – Déployer une stratégie d'IA générative en 6 étapes
http://tinyurl.com/2c2ty7n3

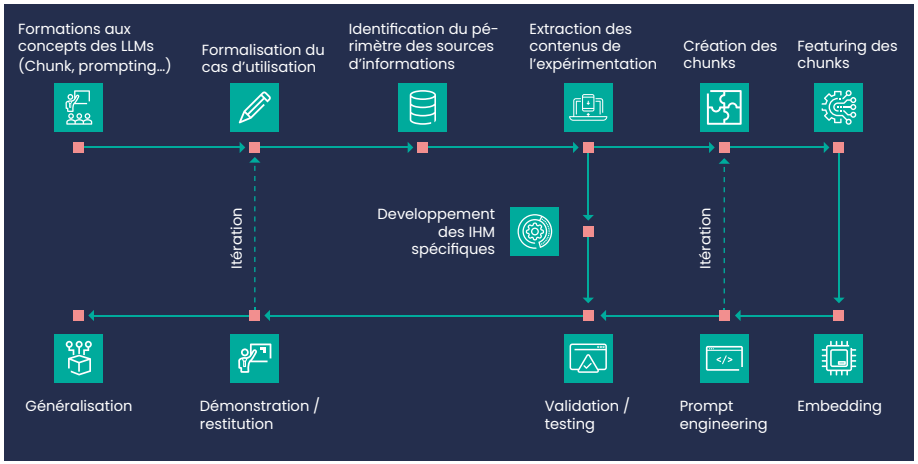
LINC – Quelles régulations pour la conception des IA génératives ?
http://tinyurl.com/ycdmw464

BDM – Étude : les 50 outils d'IA générative les plus utilisés en 2023
http://tinyurl.com/jh3zx9t6

Méthodologie d'implémentation

5

Méthodologie de passage à l'échelle



Profils

Les postes dépendent de la nature, de la taille et de l'objectif de votre projet. Mais voici quelques exemples de métiers qui peuvent être utiles :

Le prompt engineer rédigera les requêtes efficaces pour les modèles d'IA génératives et testera et améliorera les productions des IA génératives utilisées.

Le prompt engineer pourra être accompagné d'un **Expert LLM** et d'un **Expert MLOps/LLMOps.** Leurs missions sont multiples. Au-delà du choix du LLM le plus pertinent, ils interconnectent les LLM avec les APIs et bases de données internes.

Un chef de projet IA. C'est un professionnel qui coordonne et supervise le déroulement du projet d'IA générative. Il définit les besoins, les objectifs, les ressources, les délais, les livrables, etc. Il anime et gère l'équipe projet, en assurant la communication, la collaboration et la résolution des problèmes. Il suit et contrôle la qualité, le coût et les risques du projet. Enfin, Il définit les métriques et assure le reporting.

Un data engineer. Dans le cadre d'un projet d'IA générative ce profil :

- Conçoit, construit et maintient les infrastructures de données nécessaires au projet.
- Collecte, stocke, traite, nettoie, enrichit et sécurise les données provenant de sources internes ou externes.
- Il met en place des pipelines de données, des bases de données, utilise les solutions cloud de l'entreprise, conçoit des outils d'analyse, etc.
- Il assure la disponibilité, la fiabilité, la scalabilité et la performance des données.

Un data scientist. Ses missions sont diverses : il analyse, modélise et interprète les données pour le projet d'IA générative. Il utilise des méthodes statistiques, mathématiques, informatiques et de machine learning pour extraire des informations, des tendances, des patterns, des prédictions, etc. Il développe, entraîne, teste, évalue et améliore les modèles d'IA générative, en utilisant des frameworks comme TensorFlow, PyTorch, etc. Il restitue ses résultats sous forme de rapports et/ou dashbords.

Un UX designer. Ces missions principales sont de :

- Concevoir et optimiser l'expérience utilisateur du projet d'IA générative.
- Étudier les besoins, les attentes, les comportements, les émotions, etc., des utilisateurs cibles.
- Créer des persona, des scénarios, des parcours utilisateurs, des wireframes, des prototypes et tester, valider, mesurer et améliorer l'ergonomie, l'utilisabilité et l'accessibilité.

Sources :

Studyrama - Devenir Prompt Engineer - Fiche métier, formations et salaire
http://tinyurl.com/euhjbnk6

OpenClassrooms for Business - Objectif IA : le défi de l'intelligence artificielle générative en entreprise
http://tinyurl.com/5a6pxa4s

Comment bien dimensionner son projet ?

Comme tout projet utilisant de l'intelligence artificielle, un « projet LLM » va embarquer un grand nombre d'incertitudes.

La toute première étant évidemment « Est-ce que ce type de technologie peut m'apporter tout ou partie d'une solution pour mon problème (le cas d'utilisation envisagé) ? », suivi de très près par « Est-ce que cela fonctionne dans mon contexte, avec mes données ? »

La bonne nouvelle avec les LLM, c'est qu'ils sont pré-entrainés, on peut donc directement prototyper les sujets retenus, sans avoir à passer toutes les étapes d'entrainements comme dans un cas d'IA « classique ».

Le revers de la médaille, c'est qu'ils sont si volumineux qu'ils ne peuvent être « adaptés » qu'avec des très fortes contraintes techniques et budgétaires… On peut donc considérer (en faisant un rapide raccourci) que si le prototypage ne donne pas des résultats encourageants immédiatement, il vaut mieux, à part cas exceptionnels, passer à un autre cas d'utilisation.

Pour cette première phase, que l'on peut considérer comme un cadrage fonctionnel/prototypage, le dimensionnement est faible, de l'ordre d'une semaine ou deux avec un investissement très limité.

Si cette phase donne des bons résultats, alors l'étude de la « généralisation » peut être envisagée.
Décision quant à l'architecture à retenir (modèles, infrastructure, etc.), projection de l'intégration avec les éléments en interaction au niveau du SI (quelques prompts seuls n'ont que peu de valeur, une fois « l'effet whaou » passé…), étude des risques (impact quand le LLM se « trompera », analyse/risques contractuels, politique de communication aux utilisateurs), et surtout étude détaillée du ROI, vont constituer le cœur de l'étude de généralisation.

Ce cadrage projet peut s'avérer être assez long et peut durer plus d'un mois en fonction de la complexité d'intégration notamment.

Si le ROI projeté est au rendez-vous, alors le projet va démarrer, toujours de manière itérative, car il va être nécessaire de prévoir des cycles d'amélioration pour l'utilisation du LLM et il sera nécessaire de l'opérer dans des conditions le plus proche possible de la « production ».
Une approche en pilote en somme, avec des équipes spécifiques, formées et bien au fait des objectifs de la phase, une fois un MVP développé avec l'ensemble des API et interactions avec les applications et services impliqués dans le processus.

L'unité de temps du pilote sera plutôt au mois qu'à la semaine, et devra de fait durer plusieurs mois.

En parallèle de cette phase, le plan de conduite du changement devra être établi (en prenant en compte tous les impacts qu'aura le projet), et la politique d'exploitation définie (en s'assurant que le service de production sera bien en mesure d'assurer le monitoring du futur projet).

Pour opérer et orchestrer toutes ces étapes, il sera nécessaire, comme nous l'avons vu précédemment, d'avoir de nouvelles compétences… Architecte IA, Prompt Engineer, Expert LLM, Expert MLOps/LLMOps, vont être les nouveaux profils nécessaires pour mener à bien ce type de projet. Et que ce soit avec une équipe externe ou interne, une explication du fonctionnement des LLMs et du déroulement particulier du projet sera absolument obligatoire pour l'ensemble de l'équipe projet (métier et technique) afin que toute l'équipe comprenne les contraintes inhérentes à ce type de technologie.

Retours d'expérience

6

3 types d'écueils mentionnés par les contributeurs

Tests réalisés sur des enregistrements de conversation :

- « Le client était dans sa voiture et parlait fort. L'IA a interprété le ton de sa voix comme de la colère. »
- « Nous cherchions à différencier le locuteur du récepteur. Les résultats de l'IA étaient faibles. Ce cas d'usage était inutile car techniquement nous pouvions séparer les canaux de communication et identifier automatiquement les individus qui s'exprimaient. »

Traitement de FAQ

- « Les réponses proposées par le bot étaient insuffisantes car non personnalisées. De plus certaines fonctions de recherche devenaient trop compliquées et coûteuses en comparaison du service apporté (ex : afficher le solde du client). Il y a un juste équilibre à trouver entre l'effort consenti et la valeur générée. »

Diversité des produits et des configurations

- Dans les couacs, il peut être intéressant de partager les tests faits sur un configurateur en ligne. « Un client posait à 23 h chez lui une demande de configuration de véhicule. La diversité des modèles et des demandes faites au modèle, entrainaient le client dans une variété de réponses possibles et finissaient par le perdre. »

Bibliographie

Enseigner avec l'IA : Le guide d'OpenAI en français pour les enseignants
http://tinyurl.com/59b83pm5

OpenIA - FAQ éducation
https://urlr.me/vTLt4

Objectif IA : le défi de l'intelligence artificielle générative en entreprise
http://tinyurl.com/5a6pxa4s

LINC - Quelles régulations pour la conception des IA génératives ?
http://tinyurl.com/ycdmw464

Maîtriser l'IA générative | Top 8 des outils et compréhension des concepts
http://tinyurl.com/5ffpxyet

Étude META - "Sustainable AI: Environmental Implications, Challenges and Opportunities" | Arxiv
http://tinyurl.com/2kj48yt3

inetum.
Positive digital flow

Ce guide a pour vocation d'aider les directions de la relation et réclamation client à conduire un projet d'**IA générative**. C'est un support réalisé avec les directions métiers réunies au sein de l'**Amarc** (Association pour le Management de la Réclamation Client) auxquelles **Inetum** a apporté son savoir-faire technologique.

inetum.com