Génération d'une base de courriers électroniques synthétiques par des grands modèles de langue dans le domaine de la relation client

Fatma-Zohra Hannou[1], Isabelle Renault[1], Florent Mely[2], Anne-Laure Guénet[3], Guillaume Dubuisson Duplessis[3], Sabrina Campano [1]

[1] EDF Lab Paris Saclay, SEQUOIA
[2] AI&Data
[3] EDF Commerce, Direction des Systèmes d'Information et du Numérique (DSIN)

fatma-zohra.hannou@edf.fr, isabelle.renault@edf.fr, anne-laure.guenet@edf.fr, guillaume.dubuisson-duplessis@edf.fr, sabrina.campano@edf.fr

Résumé

Dans le domaine de la relation client, exploiter les données textuelles offre un levier essentiel pour développer des systèmes d'IA performants, mais présente d'importants défis de confidentialité et de conformité réglementaire, et nécessite souvent l'annotation manuelle et coûteuse de données. Cet article décrit une approche d'utilisation des grands modèles de langue pour générer des e-mails synthétiques, sans intégrer de données client réelles. L'objectif est de permettre d'entrainer des systèmes d'IA sur ces données synthétiques, en offrant de meilleures garanties de protection de la vie privée, et en permettant de minimiser le volume de données manuellement annotées. Nous décrivons la chaîne de traitement et son implémentation notamment la phase de création de prompts, qui capture la diversité des sujets, des styles et les types de données personnelles. Au-delà de la génération, l'insertion d'entités fictives dans le texte permet de reformer un email automatiquement annoté similaire à un email réel. Les résultats d'évaluation sur un jeu de données de 1600 emails indiquent une piste prometteuse pour l'entraînement de systèmes d'IA tout en offrant de meilleures garanties du point de vue du respect de la vie privée et de la conformité réglementaire.

Mots-clés

Grands modèles de langue, génération de données synthétiques, ingénierie du prompt, RGPD, protection des données à caractère personnel, reconnaissance d'entités nommées.

Abstract

Training AI systems for customer relations tasks is hindered by the high costs of manual data annotation and stringent privacy regulations. This article describes an approach using large language models to generate synthetic emails without incorporating real customer data. The aim is to train AI systems on these synthetic data, providing better privacy guarantees and minimizing the volume of manually annotated data. We describe the processing pipeline and its implementation, particularly the prompt creation phase, which captures the diversity of topics, styles, and types of personal data. Beyond generation, inserting fictitious entities into the text allows for the automatic annotation of an email similar to a real one. Evaluation results on a dataset of 1,600 emails indicate a promising approach for training AI systems while offering better guarantees in terms of privacy and regulatory compliance.

Keywords

Large language models, synthetic data generation, prompt engineering, GDPR, protection of personal data named-entity recognition.

1 Introduction

La relation client d'un grand Groupe comme EDF produit un volume important de données textuelles. Ces données, issues d'e-mails, de commentaires de satisfaction ou encore de conversations téléphoniques retranscrites se caractérisent par leur diversité et leur variété. Elles peuvent être courtes et spontanées ou, au contraire, longues et structurées, présentant une hétérogénéité en termes de style, de ton et de niveau de langue, soulevant ainsi de véritables défis pour les explorer efficacement [7]. En particulier, les e-mails, souvent rédigés en français, sont utilisés pour répondre au mieux aux attentes de nos clients en suivant le cadre réglementaire du « règlement général sur la protection des données » (RGPD) [6]. L'analyse de ces données textuelles par des techniques d'intelligence artificielle (IA) permet d'améliorer la qualité du service client, de personnaliser les interactions et d'optimiser les processus. Par exemple, l'IA peut être utilisée pour automatiser le traitement des e-mails (par exemple le routage vers le bon service [8]) ou encore détecter les irritants et les motifs de satisfaction [25].

Cependant, l'utilisation de données réelles pour entraîner ces systèmes d'IA soulève des défis importants en matière de protection de la vie privée, dans le contexte du RGPD

et dans le cadre d'une démarche *privacy-by-design*, visant à intégrer la protection des données dès la conception des systèmes d'IA. Une approche classique pour répondre à ces enjeux consiste à désidentifier les données réelles en supprimant ou en masquant les informations identifiantes, notamment à l'aide de techniques de reconnaissance d'entités nommées (NER) [6]. Or, l'entraînement de ces modèles de désidentification nécessite lui-même l'utilisation de données réelles pour être performant et robuste.

Dans ce contexte, l'essor des grands modèles de langue (LLM) [1, 9] a ouvert de nouvelles perspectives pour la génération de données synthétiques textuelles, offrant une alternative à des situations où la disponibilité des données est un frein. En effet, les LLM peuvent générer des données textuelles réalistes et variées [3, 12], permettant ainsi d'éviter l'utilisation de données personnelles des clients présentes dans les systèmes d'information privés des entreprises.

Cependant, cette approche s'accompagne de plusieurs défis. Au-delà du risque d'hallucinations (génération de contenu incorrect ou impertinent), il est généralement difficile de contraindre les LLM à produire des données variées conformes à des caractéristiques observées dans les données réelles telles que les contextes relatifs à des sujets de mails, le style de rédaction, ou l'orthographe [18].

Cet article explore le potentiel des LLM pour la génération de données synthétiques anonymes dans le domaine de la relation client. La chaine de traitement présentée permet de créer une base d'e-mails synthétiques en utilisant les LLM, et sans accéder à des données client du Groupe EDF. Elle repose notamment sur une phase de conception du prompt (*prompt design*) qui guide la génération au moyen de caractéristiques descriptives des données réelles. La base générée se décline en deux versions : la première contient des textes de mails désidentifiés qui font uniquement référence à des types d'entité (comme nom, prénom, adresse), et la deuxième version instancie dans chaque e-mail les types d'entités qu'il mentionne par des données de clients fictifs générées automatiquement. Cette deuxième version de la base fournit un corpus dont les entités sont annotées par le LLM, ce qui augmente son utilité pour les tâches d'entraînement. Nous analysons ensuite la qualité des données générées (1600 mails), en évaluant leur coût, leur diversité et leur utilité pour des applications concrètes. Conscients qu'il est difficile de qualifier d'anonyme l'algorithme de génération d'une part, les données synthétiques d'autre part [22, 14], nous analysons ce processus de génération au travers du prisme des risques liés à l'utilité, à la réidentification, et à l'exactitude. Enfin, nous discutons du caractère anonyme ou non des jeux de données générées, en tenant compte des spécificités du domaine de la relation client et des exigences du RGPD.

Cet article s'articule autour de quatre sections principales. La Section 2 présente un état de l'art sur les principaux travaux d'anonymisation et l'usage des LLM pour la génération de données synthétiques. La Section 3 détaille la chaîne de traitement implémentée pour la génération des e-mails synthétiques. La Section 4 expose les résultats obtenus et présente l'évaluation de la qualité des données synthétiques et les coûts liés à leur génération. Ces résultats sont discutés en Section 5, puis la Section 6 dresse les principales conclusions et quelques perspectives prometteuses.

2 Travaux antérieurs

2.1 Techniques de désidentification

La désidentification utilisée pour pseudonymiser voire anonymiser les données personnelles est une question centrale. La pseudonymisation est définie comme « ...un traitement de données personnelles réalisé de manière à ce qu'on ne puisse plus attribuer les données relatives à une personne physique sans information supplémentaire. », l'anonymisation comme « ...un traitement qui consiste à utiliser un ensemble de techniques de manière à rendre impossible, en pratique, toute identification de la personne par quelque moyen que ce soit et de manière irréversible. » [4]. Différentes techniques de désidentification des données existent. Majeed et al. [15] distinguent les modèles de désidentification de texte basés sur le traitement du texte de ceux associés à la publication de données préservant la vie privée (PPDP). Les premiers détectent puis suppriment les entités sensibles à l'aide de règles, d'apprentissage automatique ou encore de réseaux neuronaux Ils nécessitent de disposer de données annotées dont le coût de production reste élevé. Les seconds introduisent du bruit pour masquer les données identifiantes et réduire le risque de divulgation. Zhao et Chen [27] abordent le compromis confidentialité - utilité des données des modèles PPDP, et soulignent la nécessité d'évaluer les risques de réidentification.

2.2 Génération de données synthétiques

Les LLM ouvrent de nouvelles perspectives avec la génération de données synthétiques à l'aide d'instructions (prompts). Le contrôle de la conformité des sorties des LLM à ces instructions fait l'objet de nombreux travaux de recherche. Sahoo et al. [20] proposent un état de l'art des techniques d'ingénierie de prompts en fonction des domaines d'application (réalisation de nouvelles tâches sans nouvel apprentissage, réduction des hallucinations...). Sahoo et al. [5] se focalisent sur l'apprentissage en contexte (ICL). Long et al. [12] proposent une vision unifiée des travaux menés sur la génération de textes, la curation et l'évaluation des textes générés. Ils soulignent que les principaux défis à relever restent de garantir la fidélité des contenus des textes générés et de s'assurer de leur diversité.

L'utilisation de données synthétiques fait l'objet de nombreux travaux, notamment dans le domaine médical. Différentes études montrent des gains à finetuner des modèles basés sur une architecture *transformer* comme BERT [1], RoBERTa [2], BioBERT [3] ou encore ClinicalBERT [4] pour des

1. Hugging Face - Modèle BERT : https://huggingface.co/docs/transformers/model_doc/bert
2. Hugging Face - Modèle RoBERTa : https://huggingface.co/docs/transformers/model_doc/roberta
3. BioBERT : https://github.com/dmis-lab/biobert
4. ClinicalBERT : medicalai/ClinicalBERTÂůHuggingFace

tâches aval telles que l'extraction d'information et la détection de symptômes à l'aide de jeux de données incluant des données générées par des LLM. Li et al. [10] explorent deux approches d'augmentation de données de dossiers médicaux électroniques à l'aide de LLM : l'annotation de jeux de données publics et la génération de dossiers médicaux fictifs. Tang et al. [23] utilisent des données synthétiques étiquetées générées par un LLM à l'aide de prompts sélectionnés lors d'un processus itératif intégrant une évaluation humaine. Chung et al. [3] adoptent une démarche alliant stratégie de paramétrage des LLM et interactions humaines pour la correction d'étiquettes afin d'accroître la diversité des textes générés. Ils montrent que cette correction augmente de 14,4% la précision absolue de modèles BERT entrainés avec leurs jeux de données diversifiés.

Plusieurs stratégies ont également été conçues pour contrôler les sorties des LLMs. Väth et al. [24] proposent une appproche de génération de données synthétiques par des LLMs basée sur le parcours d'un arbre de dialogue et constatent l'apport de cette structure pour l'entraînement d'agents d'apprentissage par renforcement. [26] proposent *KnowGPT* un système permettant de créer des prompts enrichis avec des informations issues de graphes de connaissances pour une meilleure contrôlabilité des LLMs.

L'utilisation de données créer des données textuelles synthétiques n'est cependant pas synonyme d'anonymisation. Ainsi, Staab et al. [21] montrent des risques de réidentification, par les LLMs, de données personnelles incluses dans leurs données d'entraînement y compris lors d'inférences réalisées à partir de textes anonymisés. Par ailleurs, Stadler et al. [22] montrent empiriquement que des données tabulaires synthétiques ("différentiellement privées") produites par des modèles génératifs n'offrent pas un meilleur compromis entre confidentialité et utilité que celles issues de techniques d'anonymisation traditionnelles.

L'utilisation des LLMs pour synthétiser des données, notamment dans le cadre de l'entraînement ou de l'amélioration d'autres modèles d'IA, pose plusieurs défis. Il est crucial de vérifier si la licence du LLM permet la réutilisation de ses sorties et de s'assurer que la propriété intellectuelle est respectée. Cela implique de garantir que le LLM n'a pas été entraîné sur des données soumises à des restrictions de propriété intellectuelle. La traçabilité des données utilisées pour l'entraînement est essentielle pour assurer la conformité, mais les documentations actuelles manquent souvent de transparence [11] (le site web "opening up ChatGPT" [5] étudie l'ouverture des LLMs). Conscients de ces manques, Longpre et al. [13] ont réalisé avec l'aide d'experts juridiques et d'experts en apprentissage automatique, un audit sur plus de 1800 jeux de données textuelles. Ils montrent des omissions de licence de plus de 70% et des taux d'erreurs de plus de 50% sur les sites d'hébergement de données largement utilisés, et proposent une interface interactive [6] pour permettre de retracer la provenance des jeux de données de finetuning open-source les plus populaires.

5. https://opening-up-chatgpt.github.io/

6. Data Provenance Explorer : https://www.dataprovenance.org/

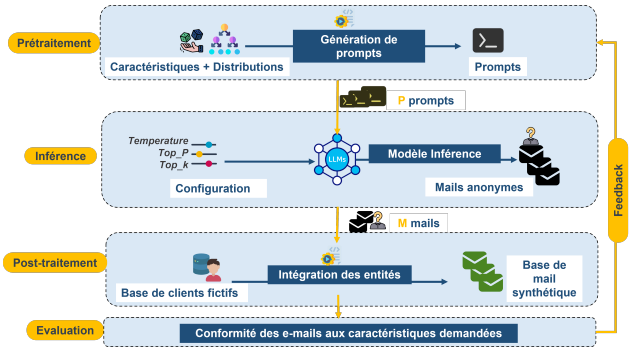

FIGURE 1 – Chaîne de traitement implémentée pour la création des e-mails synthétiques. Le processus comporte 4 phases : le **prétraitement** pour la conception de prompts diversifiés capturant les caractéristiques souhaitées, **l'inférence** utilisant des LLMs pour la génération de texte avec des types d'entités non instanciés, le **post-traitement** qui intègre des instances d'entités fictives issues de données ouvertes. L'**évaluation** qui mesure la qualité et l'utilité des données créées, et ces résultats permettent d'ajuster le prétraitement via une boucle de feedback.

Dans ces travaux, nous prenons le parti de générer des données synthétiques sans utiliser des données presonnelles des clients et nous étudions leur intérêt dans une approche « privacy by design ». L'utilisation des LLM pour la génération de données synthétiques semble d'autant plus pertinente dans le domaine de la relation client que ces outils sont de plus en plus utilisés par les clients eux-mêmes pour leurs communications écrites avec les entreprises.

3 Méthodologie

La méthodologie de génération des e-mails synthétiques repose sur une chaîne de traitement constituée de quatre étapes principales, illustrées dans la Figure 1.

Une première étape de *prétraitement* permet de créer des prompts diversifiés, représentatifs des caractéristiques des corpus client réels. Ces prompts fournis en entrée d'un LLM permettent de contrôler la génération des e-mails. Le paramétrage du LLM via ses hyperparamètres est un moyen de gérer les aspects de créativité, cohérence, longueur du texte en sortie... Les e-mails bruts générés par le LLM sont des e-mails incluant des libellés de type d'entités tel que *nom* ou *prénom*, et qui n'incluent pas d'exemple de ces types d'entités (tels que "*Dupont*" ou "*Marie*"). Cette spécificité a été explicitement demandée dans les prompts, afin de faciliter le processus d'annotation (détaillé dans la Section 4). Par la suite, une phase de *post-traitement* traite les textes des e-mails pour réintégrer des entités fictives issues de données ouvertes, augmentant ainsi le réalisme et l'utilité du corpus créé. La dernière phase consiste à évaluer la conformité du corpus généré aux caractéristiques encodées dans les prompts lors du prétraitement, et permet ainsi de réajuster le processus de génération en fonction des résultats obtenus. La Section 4 détaille l'application de cette méthodologie pour notre cas d'usage.

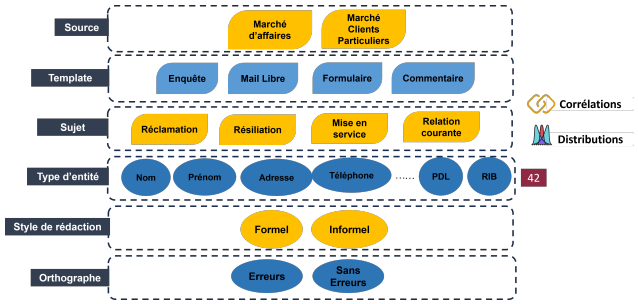

FIGURE 2 – Liste des caractéristiques considérées pour la création des prompts. Les 6 dimensions capturent la diversité observée dans un corpus de données clients réel : la source métier, le gabarit (template), le sujet, les entités à intégrer, le style de rédaction et la qualité orthographique.

3.1 Données d'entrée

Un ensemble de descriptions statistiques des e-mails a été collecté pour permettre d'enrichir le processus de génération, et de diversifier les résultats. Aucun extrait provenant d'e-mails réels n'a été utilisé dans la chaîne de traitement. Cette stratégie vise à garantir la conformité de la base de données créée aux réglementations européennes qui régissent la qualification des jeux de données anonymes [2]. Cela permet notamment d'éviter les risques d'individualisation, de corrélation ou d'inférence, entre autres, pouvant mener à la ré-identification.
Ces descriptions statistiques incluent :

1. La liste des types d'entités directement ou indirectement identifiantes couramment présentes dans les e-mails clients (nom, adresse, point de livraison,...).
2. Des métadonnées décrivant les caractéristiques des e-mails (sujets traités, types des canaux sources).
3. Des distributions probabilistes modélisant la fréquence des types d'entités, directement ou indirectement identifiantes, et des thématiques abordées.

3.2 Stratégies de création des prompts

La création des prompts représente une étape clé de la chaîne de traitement, car elle permet de guider le LLM vers caractéristiques attendues et de capturer la diversité des e-mails souhaitée.
La Figure 2 montre les caractéristiques considérées pour la création des prompts : la source des e-mails (marché d'affaire ou marché des clients particuliers), leur gabarit ou *template* (ex : enquêtes, e-mails libres), leur sujet (thématique, ex : réclamation, mise en service, résiliation), les types d'entités possibles pour les e-mails ainsi que le style (formel ou informel) et enfin le niveau la qualité orthographique de ces derniers.
Chaque prompt est constitué de deux parties :

— Un prompt système générique <syst> définissant le rôle du LLM pour cette tâche de génération des textes synthétiques.
— Un prompt d'instruction structuré </INST> pour garantir une variété linguistique et contextuelle, définissant notamment le sujet de l'e-mail, le style de rédaction et la qualité orthographique attendue.

Exemple 1. *Extrait de prompt pour générer un e-mail client avec pour thème :* réclamation.

```
<s> Tu es un générateur de courriers é
lectroniques. Le courrier électronique est
envoyé par un client à un fournisseur d'énergie.
Le courrier électronique doit être rédigé
uniquement en français.
<INST> Génère un courrier électronique de
réclamation. Le courrier électronique doit être
rédigé par un client qui décrit l'objet de sa
réclamation, en fournissant les informations
nécessaires pour expliquer le contexte.
Utilisez un style de rédaction décontracté et
simple.</INST></s>
```

Nous avons envisagé plusieurs approches pour l'intégration des types d'entités à l'instruction du prompt : approche basée sur la distribution des types d'entités dans le corpus réel, approche basée sur la distribution du nombre d'entités par e-mail, ... Pour ces travaux, la première approche est retenue.
L'objectif de cette stratégie est de contrôler la fréquence d'apparition de chaque type d'entité dans le corpus final, afin d'imiter un corpus réel. En effet, un corpus d'entraînement synthétique avec des entités plus fréquentes ou rares que dans le corpus réel pourrait affecter les performances du système d'IA. Cette approche prend en compte les corrélations entre les entités pour augmenter le réalisme des e-mails générés. Par exemple, lors d'une première mise en service dans une nouvelle maison, un client ne mentionnera probablement pas le point de livraison (référence du point de soutirage ou d'injection de l'électricité du client). La corrélation entre le thème *"mise en service"* et l'entité *"point de livraison"* est faible voire inexistante.

3.3 Génération

La phase de génération utilise un LLM pour la création des textes des e-mails. Plusieurs paramètres contrôlant la part d'aléatoire, tels que la *température, le top-p et le top-k*, ont été ajustés au cours des différents tests afin de déterminer la configuration optimale, permettant un meilleur compromis entre la diversité/créativité d'un côté et le respect des prompts sur les aspects de formalité et styles de rédaction (voir Section 4 pour les résultats et évaluation et la Section 5 pour la discussion associée). Plusieurs valeurs de $maxnumtokens$, contrôlant la longueur maximum de la séquence générée par le LLM, peuvent être spécifiées avec un potentiel impact sur les performances du modèle (temps de génération).
Par ailleurs, une stratégie d'optimisation par lot (*« batch strategy optimization »*) permet de réduire le temps de génération, en exécutant, à partir du même prompt, plusieurs inférences en parallèle sur différents GPU réduisant ainsi le coût total de génération.

Exemple 2. *Le prompt défini dans l'Exemple 1, permet la génération de l'e-mail synthétique suivant* [7] *:*

Madame, Monsieur,
Je me permets de vous écrire ce mail car j'ai rencontré un problème avec mon approvisionnement en électricité chez [Adresse], sous le point de comptage et d'estimation [pce].
En effet, ma maison est actuellement privée d'électricité. Je suis ... confort minimum.
Malgré mes tentatives de vous contacter par téléphone, je n'ai pu obtenir aucune information sur la durée de cette coupure et sur les raisons de celle-ci.
Par conséquent, je vous serais reconnaissant de bien vouloir m'informer ...
Dans l'attente de votre réponse, je vous prie d'agréer, Madame, Monsieur, mes salutations distinguées.
Cordialement,
[nom]
[email]"

3.4 Post-traitement

Les e-mails générés par le LLM sont des e-mails incluant des entités génériques (ex. *"[nom de famille]"*) et non pas des entités instanciées (ex. *"Durand"*) offrant ainsi deux avantages : minimiser le risque d'intégration d'entités instanciées utilisées lors de la phase de pré-entraînement du LLM et permettre l'annotation automatique du corpus d'e-mails généré, évitant une annotation humaine coûteuse.
L'instanciation de chaque e-mail comprend quatre étapes principales. La première étape consiste à détecter les entités génériques contenues dans l'e-mail sur la base d'une expression régulière matchant tous les patrons de chaînes de caractères délimitées par des crochets ouvrant et fermant (ex. : *"[nom de famille]"*). Le prompt demande explicitement à ce que ces types d'entités soient délimités par des crochets. L'identification du type de chacune de ces entités est ensuite effectuée en parcourant un dictionnaire associant à chaque type d'entité une expression régulière couvrant les différents patrons (les différentes formes prises par les entités génériques) de reconnaissance de cette entité (ex. *"nom"* : "nom", "nom de famille"). Puis, un dictionnaire d'entités instanciées est généré sur la base des types d'entités détectés par l'outil : les clés représentent les types d'entités et les valeurs les entités instanciées (ex. "nom" : *" Durand"*). Cette instanciation est faite, suivant les types d'entités, soit par tirage aléatoire dans des échantillons issus de bases de données ouvertes , soit sur la base d'expressions régulières conformes aux formats des types d'entités concernées, ou alors via le module python *Faker* [8]. La dernière étape consiste à remplacer chaque entité générique présente dans l'e-mail par sa valeur correspondante dans le dictionnaire des entités instanciées généré. Pour les entités citées plusieurs fois dans le même e-mail (co-référence), un traitement *naïf* a été effectué, en attribuant la même valeur pour ces entités du même type.

Exemple 3. *Le post-traitement de l'e-mail de l'Exemple 2 produit l'e-mail suivant :*

Madame, Monsieur,
Je me permets de vous écrire ce mail car j'ai rencontré un problème avec mon approvisionnement en électricité chez 13 rue Verdier Bagneux, sous le point de comptage et d'estimation 09992424547933.
En effet, ma maison est actuellement privée d'électricité. Je suis ... confort minimum.
Malgré mes tentatives de vous contacter par téléphone, je n'ai pu obtenir aucune information sur la durée de cette coupure et sur les raisons de celle-ci.
Par conséquent, je vous serais reconnaissant de bien vouloir m'informer ...
Dans l'attente de votre réponse, je vous prie d'agréer, Madame, Monsieur, mes salutations distinguées.
Cordialement,
DURAND
jean.durand@gmail.com"

Lors du processus de post-traitement, toutes les informations requises lors d'une annotation sont mises à jour de façon incrémentale (type d'entité et position de chaque entité) garantissant ainsi l'utilisation du corpus instancié et annoté pour les tâches d'entraînement futures de systèmes IA.

4 Résultats & Evaluation

Cette section présente l'application de la méthodologie décrite dans la précédente section pour la création d'une base de données synthétique d'e-mails clients. Le périmètre de la première base synthétique se focalise sur des e-mails de clients particuliers sous forme de texte libre.

4.1 Implémentation

L'implémentation de la chaîne de traitement décrite dans la Section 3 s'effectue sur un DGX A-100 [9], en utilisant 5 GPUs, de 40GB en mémoire chacun. Au-dessous de 4 GPUs, le chargement du modèle *Mixtral 8x7b* ne peut s'effectuer (voir la discussion dans la Section 5). Tous les tests ont été effectués avec deux LLMs : *Mistral-7B-Instruct-v0.2* [10] et *Mixtral-8x7B-Instruct-v0.1* [11]. Après plusieurs séries d'expérimentations, seuls les résultats issus de *Mixtral-8x7B* ont été jugés suffisamment qualitatifs et diversifiés pour être intégrés à la base finale. Avant l'usage de ces modèles, une vérification de licence a été effectuée. Certaines licences peuvent, en effet, restreindre l'utilisation des données générées, interdisant ainsi leur intégration dans des ensembles de données d'entraînement pour d'autres modèles.

7. Email raccourci pour des raisons d'espace.
8. https://pypi.org/project/Faker/
9. https://resources.nvidia.com/en-us-dgx-systems/dgxa100-system?xs=489761
10. https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2
11. https://huggingface.co/mistralai/Mixtral-8x7B-Instruct-v0.1

Les modèles utilisés (Mistral 7b et Mixtral 8x7b) sont sous licence Apache 2.0 permissive pour ce type d'utilisation.
Un ensemble de jeux de données ouverts a été employé pour réintégrer des entités réalistes dans les e-mails : les noms [12], prénoms [13], et les adresses [14]. Tous ces jeux de données sont publiés sous licence ouverte qui autorise leur usage.
Afin d'exploiter le potentiel créatif des LLMs, la création de la base de 1600 e-mails se fait en plusieurs générations (15) chacune contenant 100 ou 200 instances, et correspondant à un paramétrage spécifique (par exemple des valeurs de température différentes). Le tableau 1 résume les caractéristiques intégrées comme instructions des prompts et leurs distributions souhaitées. Le style de rédaction varie entre formel et informel, l'orthographe peut être sans erreurs gramaticales ou avec certaines erreurs. Quatre sujets d'e-mails ont été considérés : coupure d'électricité, mise en service, résiliation, ou réclamation. Douze types d'entités ont été retenus pour la première base d'e-mails. La prise en compte des distributions permet de contrôler la diversité de la génération des e-mails afin d'augmenter l'utilité de la base synthétique comparée à un corpus réel. A noter que les probabilités dans la Table 1 ne sont données qu'à titre indicatif, et ne reflètent pas la description d'un corpus réel, pour des raisons de confidentialité.

Caractéristique	Valeurs (Pourcentage %)
Style	Formel (70%) Informel (30%)
Sujet	Coupure d'électricité (30%) Mise en service (25%) Résiliation (25%) Réclamation (20%)
Orthographe	Correcte (90%) Avec fautes (10%)
Entités	Nom (90%), Prénom (90%), Adresse (60%), Téléphone (30%), Contrat (50%) Point de Livraison (50%), Point de comptage et d'estimation (50%), Numéro client (60%) RIB (30%), IBAN (20%), BIC (20%)

TABLE 1 – Distribution des caractéristiques des prompts utilisés pour la génération de la base d'e-mails synthétiques. Les probabilités capturent la diversité des e-mails en termes de style de rédaction (formel /informel), qualité orthographique (avec ou sans erreurs), les sujets des e-mails, et la fréquence d'occurence des types d'entités.

4.2 Performance de génération

L'évaluation des performances permet de mesurer le temps de création de la base d'e-mails, avec un focus sur le temps d'inférence. Nous analysons les paramètres qui impactent le temps d'inférence, ainsi que la répartition des différentes étapes de la chaîne de traitement dans le temps global.

Analyse du temps d'inférence La Figure 3 illustre le temps d'inférence individuel d'un e-mail, l'analyse est réalisée sur un corpus de 100 e-mails générés. On remarque une variabilité des temps d'inférence qui peut s'expliquer par les longueurs variables des prompts (car incluent plus d'entités par exemple), ou leur complexité comme des prompts relatifs à des contextes de thématiques ayant plus de détails, donc une longueur de texte générée supérieure.
La plupart des e-mails sont générés avec des temps d'inférence entre 10 et 60 secondes, mais certaines rares valeurs atteignent les 80 secondes. La moyenne des temps d'inférence observés est de $\tilde{4}6$ secondes et est équivalente à la moyenne générale ($\tilde{4}4$ secondes) sur tous les jeux de données constituant la base synthétique.

Analyse de l'impact du nombre de tokens sur le temps d'inférence Au regard de la variabilité des temps d'inférence observée, une mesure de l'impact de la longueur des prompts a été effectuée. Les résultats obtenus permettent d'abord de souligner une forte diversité de longueur de prompts (entre 200 et 400 tokens) qui peut s'expliquer par la phase de création de prompts suivant une approche d'intégration de caractéristiques détaillées (expliquée dans la Section 3). En dépit de ces longueurs différentes, les temps d'inférence mesurés ne sont pas corrélés avec les longueurs de prompts utilisés dans la générations des e-mails.

Nous effectuons donc une deuxième analyse qui porte cette fois-ci sur la longueur du texte généré, dont les résultats confirment une corrélation linéaire.

Nous observons à travers les résultats une corrélation linéaire entre la longueur du texte généré et le temps d'inférence. La complexité de certains prompts conduit à des textes plus longs avec plus de détails relatifs au contexte. Des e-mails au sujet de réclamations pour coupure d'électricié peuvent énumérer les dommages causés au client, quand parfois une simple mention du problème est incluse.

Analyse de la décomposition du temps de génération global Pour illustrer le temps nécessaire à la création d'un jeu de données, la Figure 4 détaille les temps requis pour chaque phase de la chaîne de traitement. La phase d'inférence requiert 98 % du temps global du création du jeu de données, quand la phase de post-traitement (instanciation des types d'entités par des données ouvertes) s'effectue en 20 secondes, fournissant ainsi un corpus annoté. Le temps du chargement du modèle Mixstral_8x7b est de 67 secondes. Le chargement du modèle s'effectue une seule fois quel que soit le nombre d'e-mails générés et peut être mutualisé pour créer la base en un seul chargement.

12. https://www.data.gouv.fr/fr/datasets/liste-de-prenoms-et-patronymes/
13. https://www.data.gouv.fr/fr/datasets/fichier-des-prenoms-depuis-1900/
14. https://adresse.data.gouv.fr/donnees-nationales

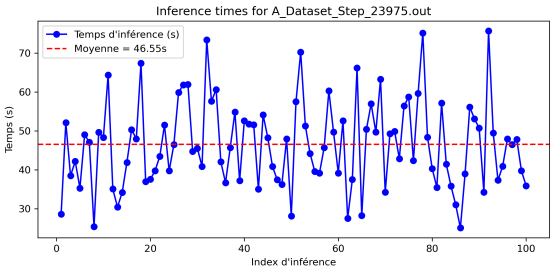

FIGURE 3 – Répartition du temps d'inférence des e-mails individuels. La variabilité des temps correspond à la diversité des prompts, leur longueur et la longueur des e-mails générés. Ce jeu de données contient 100 e-mails générés avec *Mixtral-8x7b*, température= 1.2, max_tokens_length= 4000, top_p = 0.95, top_k = 40

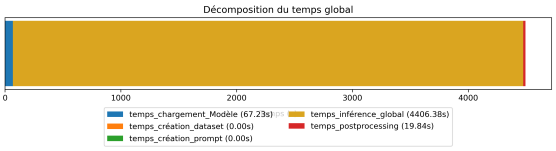

FIGURE 4 – Décomposition du temps de la création d'un jeu de données de 100 e-mails sur les différentes étapes de la chaîne de traitement. Les temps de création du jeu de prompts et son paramétrage étant insignifiants ils apparaissent comme nuls. La majeure partie du temps est consommée par la phase d'inférence (98 %).

4.3 Conformité qualitative

La conformité des e-mails générés à la langue, au sujet, au style de rédaction et à la qualité orthographique requis dans les prompts est évaluée de manière qualitative. Un échantillon de 410 e-mails a été constitué à partir des 1600 e-mails générés en se restreignant aux e-mails générés avec des températures égales à 0.9, 1.2 et 1.4. Une vérification est faite sur les distributions des caractéristiques croisées "Style" et "Qualité orthographique" de l'échantillon qui restent proches de celles du corpus.

Pour chaque caractéristique de chaque e-mail généré , la conformité aux consignes de prompt est évaluée en comparant la modalité de la caractéristique présente dans l'e-mail généré à celle du prompt de génération de l'e-mail. Une non-conformité est déclarée lorsqu'un écart entre les modalités requise et constatée, écart auquel peut être associé un seuil de tolérance, est observé. Le protocole d'annotation est le suivant pour les caractéristiques considérées :

1. Langue : la présence de plus d'un mot non français dans l'e-mail généré implique une non-conformité, tous les e-mails devant être en français.
2. Sujet : lorsque le sujet de l'e-mail généré diffère de la consigne donnée dans le prompt, une non-conformité est déclarée (ex. le prompt requiert une demande de mise en service alors que l'e-mail généré porte sur une réclamation suite à une coupure).
3. Qualité orthographique : la présence d'une faute d'orthographe ou de grammaire implique que l'e-mail généré est "avec erreurs", dans le cas contraire l'e-mail généré est "sans erreurs".
4. Style : les caractéristiques du style formel suivantes ont été considérées : 1- l'utilisation des formules de politesse, de salutations, de titres. 2- emploi du registre soutenu, 3-absence d'expressions verbales, 4-structure claire et organisée. Par exemple, lorsque l'e-mail comporte des formulations non formelles (ex. "Bonjour, J'espère que vous allez bien."), son style est considéré comme "informel", dans le cas contraire comme "formel".

La campagne d'annotation a été réalisée par deux annotateurs différents sur le même échantillon avec les mêmes règles d'annotation hormis pour le Style. En effet, cette caractéristique reste subjective, et d'autant plus difficile à appréhender qu'elle porte sur l'entièreté de chaque e-mail qui peut inclure plusieurs phrases dont certaines peuvent être formelles et d'autres informelles [19].

Les résultats de l'annotation montrent que la conformité des e-mails à la langue et au sujet est quasi toujours respectée, à la différence du style et de la qualité orthographique pour lesquels une variabilité est observée, en fonction de la température considérée. La température de 1.4 permet de gagner en conformité des e-mails générés aux prescriptions des prompts pour toutes les modalités. Une analyse plus approfondie révèle des disparités en fonction des modalités des caractéristiques de style et de qualité orthographique : la modalité "style formel" (resp. "sans erreurs") atteint un taux de conformité plus élevé que celui associé à la modalité "style informel" (resp. "avec erreurs", hors temp 1.4). Un écart est observé entre les deux annotations, plus marqué pour le style informel, où la qualification d'un texte entier (plusieurs lignes) diffère. Certains mails comportent à la fois des formules formelles et des tournures informelles, ce qui peut expliquer cet écart d'annotation. Cependant, les discussions restent quasi similaires. Le LLM a beaucoup plus de difficultés à créer des e-mails formels avec erreurs, mais ceci s'améliore avec des températures plus élevées.

La génération d'e-mails avec un style informel ou une qualité orthographique dégradée reste plus difficile à atteindre pour les températures les plus faibles (cf. Table 2 avec les résultats de l'annotation croisée des caractéristiques de style et de qualité orthographique).

4.4 Conformité de l'intégration des entités

Une analyse des e-mails générés a été effectuée pour observer la fréquence d'intégration des entités, comparée à la distribution de ces entités demandée dans le prompt de génération. Un fichier de configuration a été utilisé pour générer des prompts capturant cette variabilité (voir Section 3). La Figure 5 illustre la comparaison entre la distribution demandée, et la distribution observée dans le corpus des e-mails synthétiques. Globalement, les fréquences demandées dans les prompts ont été respectées par le LLM durant la génération, à l'exception des informations bancaires, où

Style	Orthographe	Emails conformes aux attendus		
		Temp. 0.9 A 1 / A 2	Temp 1.2 A 1 / A 2	Temp. 1.4 A 1 / A 2
Formel	Sans erreurs	97,1% / 96,4 %	98,5% / 97.2 %	98,4% / 100 %
Formel	Avec erreurs	25,0% / 40 %	52,6% / 50 %	100,0% / 100 %
Informel	Sans erreurs	86,2% / 85,7 %	36,2% / 94,7 %	88,9% / 81,81 %
Informel	Avec erreurs	50,0% / 83,3 %	100,0% / 100 %	100,0% / 100 %

TABLE 2 – Conformité des e-mails générés par modalité croisée Style × Qualité orthographique en fonction de la température. Chaque valeur représente le pourcentage de conformité pour l'annotateur A1 ou l'annotateur A2.

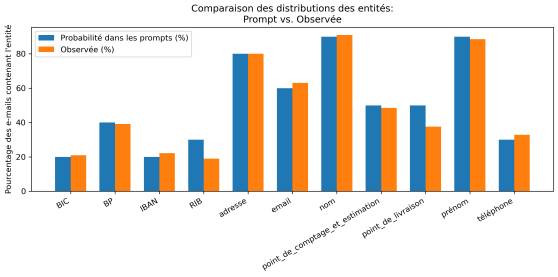

FIGURE 5 – Comparaison des distributions des entités dans les prompts et dans la base des e-mails synthétiques générée. La majorité des distributions sont respectées à part un léger déséquilibre sur les informations bancaires.

on observe moins d'entités *RIB* que demandées mais compensées par plus d'*IBAN* et de BIC.

5 Discussion

5.1 Contrôle de la génération des LLM

L'un des principaux défis observés durant la création de la base de données synthétiques était de contraindre les modèles LLM à se conformer aux instructions de prompts. La qualité des e-mails générés a donc été évaluée sur le respect de ces instructions et non sur l'adéquation du corpus généré pour des tâches aval comme la reconnaissance d'entités nommées ou autres. Lors des tous premiers tests effectués sur le modèle *Mistral 7b*, les résultats contenaient encore beaucoup de parties de textes en anglais, malgré les instructions spécifiques de se limiter à la langue française. Sur la conformité de la langue, l'usage de *Mixtral 8x7b* a nettement amélioré l'adhérence à la langue. Cela a également été le cas en ce qui concerne la formalité et la qualité orthographique. Les LLM sont souvent entraînés pour produire un contenu *propre, correctement écrit*. La génération d'e-mails qui reproduisent les comportements clients observés dans un corpus réel implique la présence de fautes d'orthographe et un usage fréquent du style informel. Pour forcer le LLM à mieux respecter ces dernières consignes, différentes températures ont été testées pour la génération, avec un impact sur la qualité des e-mails générés. Des hallucinations (peu fréquentes) ont été observées : insertion de code Python, reprise de parties d'instructions dans les corps d'e-mails (malgré les consignes de prompts), succession d'e-mails dans un même corps d'e-mail, répétition en fin d'e-mail de phrases du corps de l'e-mail sous forme de "Note : une phrase de l'e-mail". Pour pallier une partie de ces hallucinations, une étape de curation des e-mails générés a été intégrée au post-traitement. La génération de plusieurs types d'entités en lieu et place d'une unique entité générée (*[Nom Prénom]* au lieu de *[Nom]*, *[Prénom]*) par le LLM a aussi été constatée. Enfin, bien que peu fréquente, la réinjection d'entités instanciées issues du pré-entraînement du LLM dans les e-mails générés a aussi été observée.

L'analyse de la conformité met en exergue le compromis à faire entre créativité du LLM et respect des instructions de prompts par le LLM. Des améliorations de la chaîne de traitement proposée restent à faire et devront porter sur un affinement conjoint des prompts et des post-traitements (expressions régulières, curation des données). Par exemple, pour le style et l'orthographe, l'ajout d'une instruction de conformité à des standards tel que *Common European Framework of Reference for Languages* ou encore la mise en place d'une génération de texte multi-étapes avec, dans un premier temps, une demande faite au LLM de caractériser un e-mail formel vs un e-mail informel, puis la génération d'e-mails sur cette base. La diversité des e-mails générés reste à évaluer en termes de vocabulaire utilisé.

5.2 Performance et Coûts

L'utilisation d'un LLM tel que *Mixtral-8x7B* pour la génération d'e-mails synthétiques implique des coûts associés à l'infrastructure GPU. Dans cette étude, les expérimentations ont été réalisées sur une infrastructure serveur DGX-A100 exploitant 5 GPUs de 40GB chacun, et un modèle chargé en local. Aucun coût supplémentaire n'a été engendré. Pour estimer le coût d'utilisation d'une telle infrastructure, on peut comparer les coûts directs d'infrastructure locale à ceux d'une solution cloud. Sur une plateforme cloud, le coût horaire d'utilisation d'un GPU s'élève en moyenne à 3,7$ [15]. Ainsi, l'utilisation de 5 GPUs simultanément pour un cycle de génération de 15 heures entraîne un coût approximatif de 278$. A noter que ces résultats peuvent s'optimiser via *vLLM* (vs. pipeline Hugging Face [16]).

A titre de comparaison, l'annotation de 1600 e-mails réels nécessiterait une équipe d'annotateurs et un temps d'annotation cumulé de 2 à 3 minutes par e-mail, donc pouvant atteindre 80 h pour l'entièreté du cycle d'annotation (incluant les annotations croisées et leur validation). A ce titre, la génération d'e-mails s'avère compétitive (facteur × 4).

L'utilisation d'un LLM permet une économie de coût tout en offrant un corpus annoté scalable, avec une qualité homogène, et qui capture la diversité d'un corpus réel (distributions, cas particuliers,..). De plus, l'usage d'une base d'e-mails synthétiques présente un avantage de conformité réglementaire. Contrairement à l'annotation des données réelles, qui implique la manipulation d'entités identifiantes et des risques de divulgation, la génération de données synthétiques ne requiert aucun accès à des données réelles.

15. https://aws.amazon.com/fr/ec2/pricing/on-demand/

5.3 Utilité des données synthétiques

L'utilisation de données synthétiques soulève la question de leur utilité, i.e. de savoir si ces données synthétiques sont une bonne approximation de la réalité pour construire des modèles d'IA sur des données texte de la relation client. Des tests préliminaires sur la reconnaissance d'entités nommées dans des e-mails client à partir d'un modèle d'IA entraîné exclusivement sur les données synthétiques générées dans ces travaux indiquent une performance honorable se situant à 15 points en deça des modèles opérationnels les plus performants. Ces résultats encourageants nécessitent des travaux supplémentaires. Ils indiquent la probable nécessité d'un mix de données réelles et de données de synthèse pour l'entraînement de modèles opérationnels performants, permettant ainsi la minimisation de l'utilisation de données réelles. En outre, ces travaux devront veiller à ce que l'usage de données de synthèse ne renforce pas les biais existants ou ne crée pas de nouvelles discriminations [14].

5.4 Pseudonymisation ou anonymisation

Le caractère pseudonyme des données synthétiques créées dans ces travaux semble incontestable car elles sont synthétisées de manière désidentifiées avec insertion d'entités factices. Notamment, ces données synthétiques ne peuvent pas être considérées comme exactes et étant associées à des individus. En outre, le processus décrit ne repose pas sur l'utilisation de données réelles présentes dans les systèmes d'information de notre entreprise. Néanmoins, il reste des incertitudes sur le caractère anonyme des données synthétiques [14]. En effet, le LLM lui-même est pré-entraîné sur des données réelles (e.g., les données extraites de l'internet public) et, dans certaines conditions, peut générer des données identiques aux données de pré-entraînement du LLM [17]. Même si cela peut sembler improbable, il n'est pas impossible que des données suffisamment proches des données de pré-entraînement du LLM soient générées permettant une réidentification d'un individu. Ce dernier point nécessite des approfondissements pour quantifier la probabilité de réidentification dans ce type de situation.

6 Conclusion et perspectives

Nous avons présenté dans cet article une méthode permettant de produire des e-mails synthétiques de relation client répondant à deux enjeux : (i) offrir la meilleure garantie possible de la protection de la vie privée, et (ii) minimiser le temps d'annotation de données requis pour l'entraînement de systèmes d'IA. Pour cela, nous avons proposé une chaîne de traitement en plusieurs étapes générant un corpus d'e-mails annoté en entités et sans aucune référence directe à des e-mails réels. Les e-mails synthétiques sont générés au moyen de LLMs par des instructions reposant sur des caractéristiques descriptives générales des e-mails réels (ex : thèmes, distributions des types d'entités), et contiennent des entités fictives issues de bases ouvertes, ou générées artificiellement par des règles. Nos expérimentations ont mis en lumière des défis propres à la génération de données synthétiques avec des LLMs, en particulier le contrôle de l'adhérence du texte généré avec les instructions fournies dans le prompt. Nous avons montré qu'il est possible d'obtenir un corpus d'e-mails synthétiques respectant le format et la distribution des types d'entités souhaitées, la langue demandée, ou le thème. Il subsiste des limites dans les capacités du modèle à adopter un style d'expression informel et à faire des fautes d'orthographe, qui peuvent être compensées en augmentant la température. L'augmentation de la température produit une diversité qui peut engendrer un post-traitement plus coûteux.

En perspective, nous prévoyons d'apporter des compléments aux méthodes proposées, pour atteindre un niveau de performance permettant d'utiliser ce corpus de données synthétiques afin d'entrainer un système d'IA interne EDF de détection d'entités nommées. Des tests préliminaires à partir d'un modèle d'IA entraîné exclusivement sur les données synthétiques générées dans ces travaux ont montré une performance se situant à 15 points en deça des modèles opérationnels les plus performants Pour affiner davantage le contenu généré et aller plus loin dans le degré d'adhérence au prompt, nous prévoyons d'étendre la méthodologie par l'exploration arborescente pour la génération guidée de prompts et l'usage des graphes [26]. Si les résultats sont satisfaisants, ces travaux seront étendus à d'autres périmètres (segments de marché différents, et formats différents comme des formulaires ou des fils de discussion client conseiller). Un traitement plus précis des co-références est également envisagé, pour augmenter l'utilité des données pour d'autres tâches. Des guides fournis aux conseillers, donnant par exemple des informations sur le "ton de voix" (*tone of voice*) à adopter avec un client, pourraient être intégrés au prompt au moyen d'une architecture RAG, pour permettre de traiter les textes rédigés par un conseiller.

Remerciements

Nous remercions chaleureusement toutes les personnes qui sont intervenues de près ou de loin sur ce projet : Sofiane Kerroua, Anne Gayet, Laetitia Leroux, Sonia Audheon, Dominique Manzoni-Quantin, François Raynaud.

Références

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv :2303.08774*, 2023.

[2] Article 29 Data Protection Working Party. Opinion 05/2014 on Anonymisation Techniques (WP216). Technical Report WP216, European Commission, April 2014.

[3] John Joon Young Chung, Ece Kamar, and Saleema Amershi. Increasing diversity while maintaining accuracy : Text data generation with large language models and human interventions. *arXiv preprint arXiv :2306.04140*, 2023.

[4] CNIL. L'anonymisation de données personnelles, May 2020. Accessed : 2025-02-25.

[5] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Tianyu Liu, et al. A survey on in-context learning. *arXiv preprint arXiv :2301.00234*, 2022.

[6] Guillaume Dubuisson Duplessis, Elliot Bartholme, Sofiane Kerroua, Mathilde Poulain, Ahès Roulier, and Anne-Laure Guénet. Désidentification de données texte produites dans un cadre de relation client. In *Conférence Traitement Automatique des Langues Naturelles (TALN) – démonstrations*, pages 10–13, 2020.

[7] Guillaume Dubuisson Duplessis, François Bullier, and Anne-Laure Guénet. Démonstration : exploration sémantique de données texte de la relation client. In *Conférence Nationale sur les Applications Pratiques de l'Intelligence Artificielle*, pages 103–106, 2023.

[8] Guillaume Dubuisson Duplessis, Sofiane Kerroua, Ludivine Kuznik, and Anne-Laure Guénet. Cameli@ : analyses automatiques d'e-mails pour améliorer la relation client. In *Conférence sur le Traitement Automatique des Langues Naturelles (TALN). Volume IV : Démonstrations*, pages 623–626, 2019.

[9] Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. Mixtral of experts. *arXiv preprint arXiv :2401.04088*, 2024.

[10] Rumeng Li, Xun Wang, and Hong Yu. Two directions for clinical data generation with large language models : data-to-label and label-to-data. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2023, page 7129, 2023.

[11] Andreas Liesenfeld, Alianda Lopez, and Mark Dingemanse. Opening up chatgpt : Tracking openness, transparency, and accountability in instruction-tuned text generators. In *Proceedings of the 5th International Conference on Conversational User Interfaces*, CUI '23, New York, NY, USA, 2023. Association for Computing Machinery.

[12] Lin Long, Rui Wang, Ruixuan Xiao, Junbo Zhao, Xiao Ding, Gang Chen, and Haobo Wang. On llms-driven synthetic data generation, curation, and evaluation : A survey. *arXiv :2406.15126*, 2024.

[13] Shayne Longpre, Robert Mahari, Anthony Chen, Naana Obeng-Marnu, Damien Sileo, William Brannon, Niklas Muennighoff, Nathan Khazam, Jad Kabbara, Kartik Perisetla, et al. A large-scale audit of dataset licensing and attribution in ai. *Nature Machine Intelligence*, 6(8) :975–987, 2024.

[14] Alexis Léautier. [Données synthétiques] - Et l'Homme créa les données à son image 2/2. *Laboratoire d'Innovation Numérique de la CNIL (LINC)*, août 2022. Consulté le 24 janvier 2025.

[15] Abdul Majeed and Sungchang Lee. Anonymization techniques for privacy preserving data publishing : A comprehensive survey. *IEEE access*, 9 :8512–8545, 2020.

[16] Matias Martinez. The impact of hyperparameters on large language model inference performance : An evaluation of vllm and huggingface pipelines. *arXiv preprint arXiv :2408.01050*, 2024.

[17] Milad Nasr, Nicholas Carlini, Jonathan Hayase, Matthew Jagielski, A. Feder Cooper, Daphne Ippolito, Christopher A. Choquette-Choo, Eric Wallace, Florian Tramèr, and Katherine Lee. Scalable extraction of training data from (production) language models, 2023.

[18] Rajvardhan Patil and Venkat Gudivada. A review of current trends, techniques, and challenges in large language models. *Applied Sciences*, 14(5) :2074, 2024.

[19] Ellie Pavlick and Joel Tetreault. An empirical analysis of formality in online communication. *Transactions of the association for computational linguistics*, 4 :61–74, 2016.

[20] Pranab Sahoo, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal, and Aman Chadha. A systematic survey of prompt engineering in large language models : Techniques and applications. *arXiv preprint arXiv :2402.07927*, 2024.

[21] Robin Staab, Mark Vero, Mislav Balunović, and Martin Vechev. Beyond memorization : Violating privacy via inference with large language models. *arXiv preprint arXiv :2310.07298*, 2023.

[22] Theresa Stadler, Bristena Oprisanu, and Carmela Troncoso. Synthetic data–anonymisation groundhog day. In *31st USENIX Security Symposium (USENIX Security 22)*, pages 1451–1468, 2022.

[23] Ruixiang Tang, Xiaotian Han, Xiaoqian Jiang, and Xia Hu. Does synthetic data generation of llms help clinical text mining ? *arXiv preprint arXiv :2303.04360*, 2023.

[24] Dirk Väth, Lindsey Vanderlyn, and Ngoc Thang Vu. Towards a zero-data, controllable, adaptive dialog system. *arXiv preprint arXiv :2403.17582*, 2024.

[25] Nicolas Vautier, Marc Héry, Mourad Miled, Irène Truche, François Bullier, Anne-Laure Guénet, Guillaume Dubuisson Duplessis, Sabrina Campano, and Philippe Suignard. Utilisation de llms pour la classification d'avis client et comparaison avec une approche classique basée sur camembert. In *Conférence Nationale sur les Applications Pratiques de l'Intelligence Artificielle*, 2024.

[26] Qinggang Zhang, Junnan Dong, Hao Chen, Daochen Zha, Zailiang Yu, and Xiao Huang. Knowgpt : Knowledge graph based prompting for large language models. *Advances in Neural Information Processing Systems*, 37 :6052–6080, 2025.

[27] Ying Zhao and Jinjun Chen. A survey on differential privacy for unstructured data content. *ACM Computing Surveys (CSUR)*, 54(10s) :1–28, 2022.