APIA 2025- Conférence Nationale sur les Applications Pratiques de l'Intelligence Artificielle

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 (externe)** (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

N°1

Contexte

Besoins et périmètre

Contexte

Génération des données synthétiques

La génération de données synthétiques implique la création des données artificielles (textuelles, séries temporelles, visuelles, ou multimodales). Plusieurs techniques sont employées telles que les LLMs, ou les GANS et modèles de diffusion.

Qualité de données: création de jeux de données de haute qualité

Privacy-by-design pipelines: politiques d'usage/ conservation des données sécurisées (ex. RGPD)

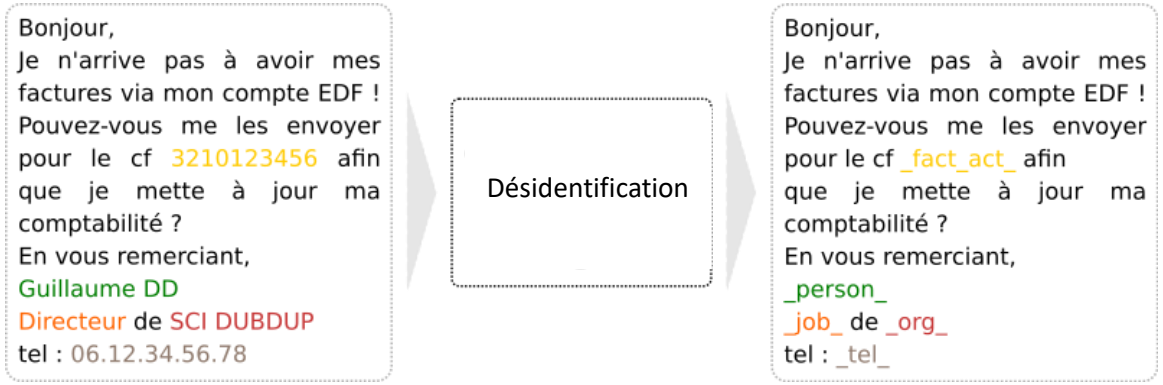
Prototypage/simulations: création de jeux de données qui couvrent des cas rares

Scalabilité: entrainement des modèles ML

La génération des données synthétiques est en forte augmentation dans plusieurs domaines: Médical (génération des données de patients, imagerie), industries (simulations de scénarios de production), …

Contexte

La désidentification permet de minimiser les entités dans des données texte de la relation client



PyTorch — Reconnaissance d'entités nommées (Named-Entity Recognition) par des techniques de « deep learning » (apprentissage automatique)

.[RegEx]* — Une séquence de caractères qui agit comme un pattern pour identifier et manipuler des chaînes de caractères

Motivation

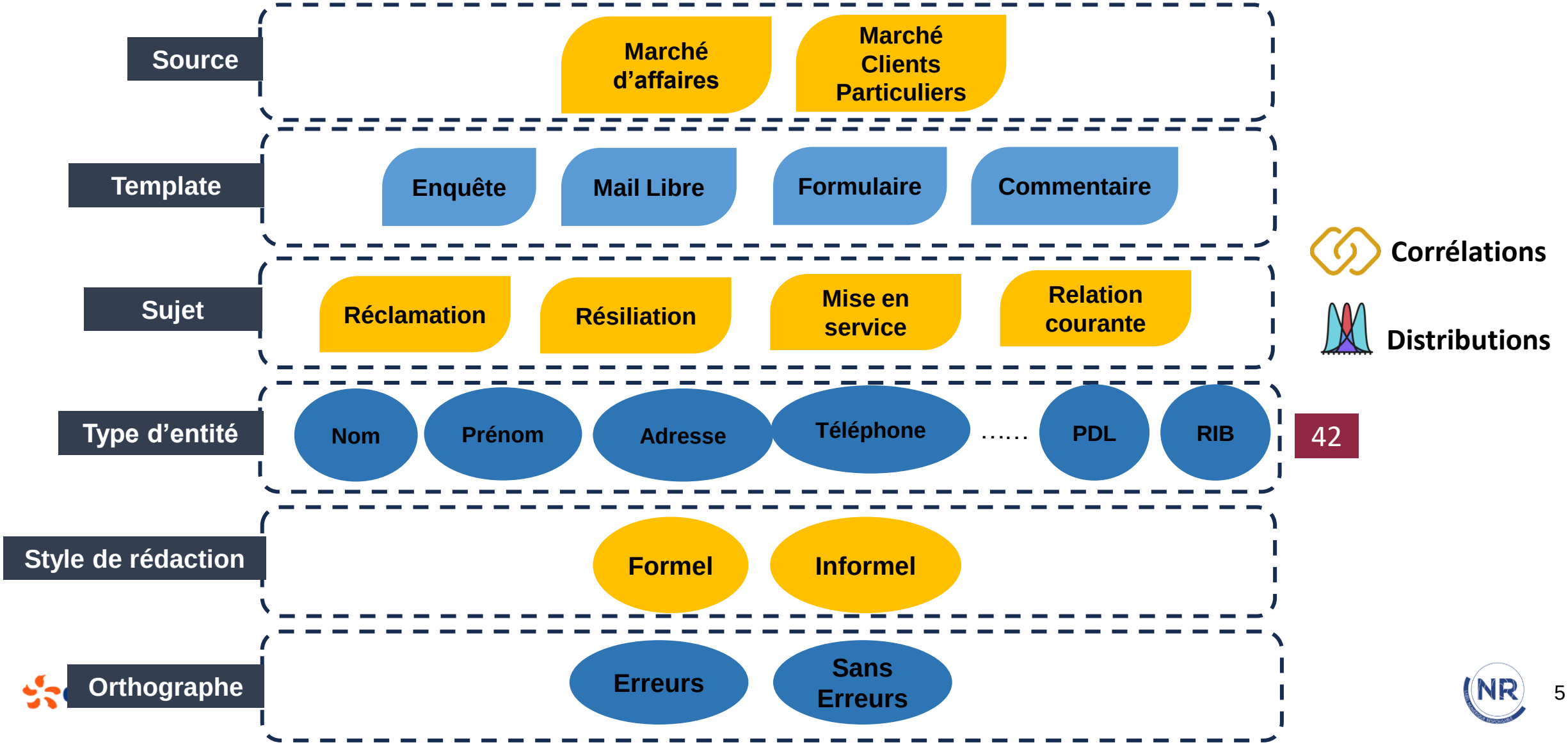
Application du RGPD : purge régulière du corpus d'apprentissage contenant des données à caractère personnel

- **Impact**: Ré-annotation à la main d'un corpus de données à chaque expiration du corpus.
- **Objectif**: Dans une démarche de privacy-by-design et de recherche d'efficience, étudier la possibilité de retirer tout ou partie des données réelles du corpus d'apprentissage en les substituant par des données synthétiques.

Contexte

2024

Caractéristiques des mails

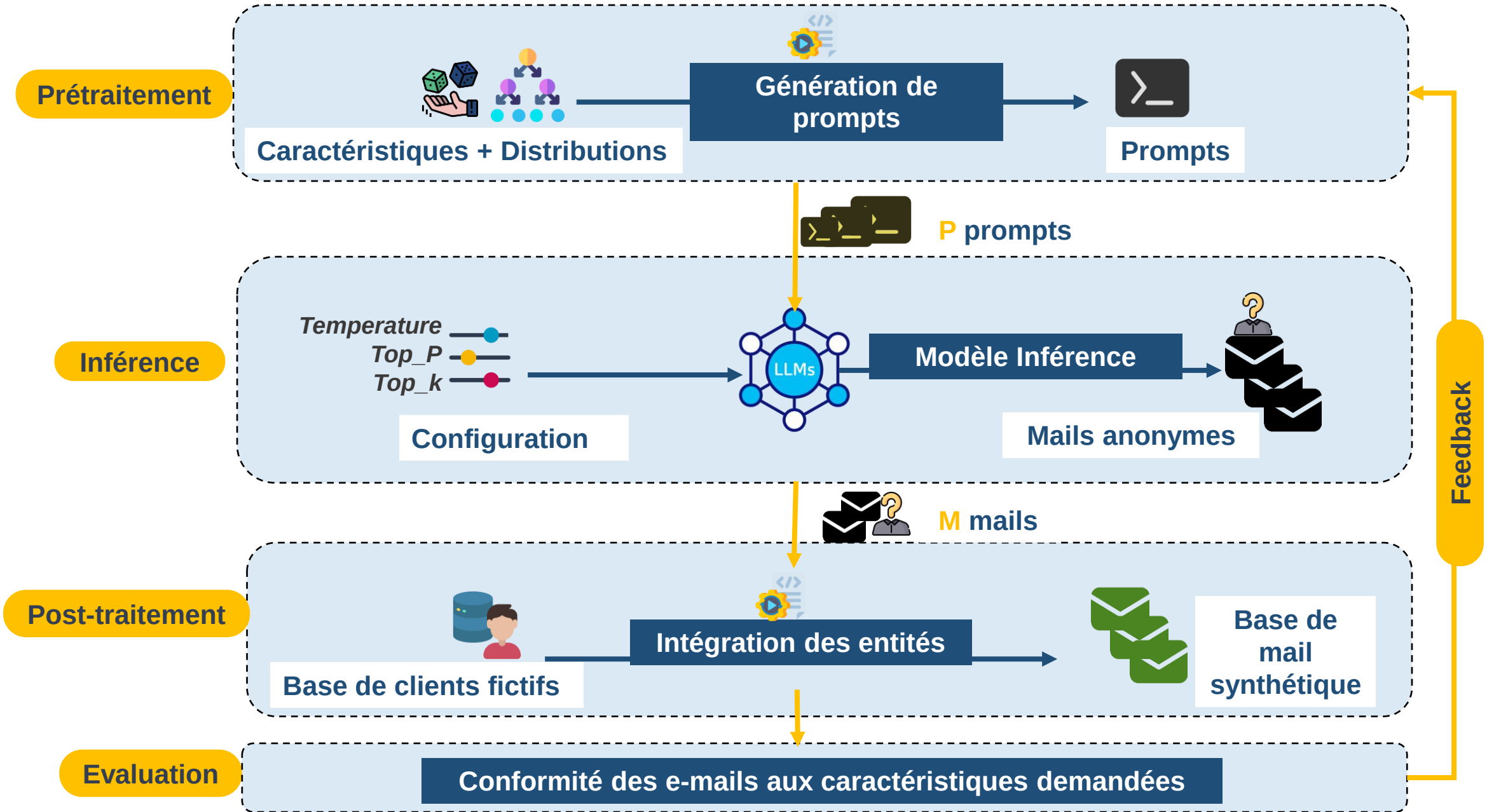


N°2

Approche

Chaîne de traitements

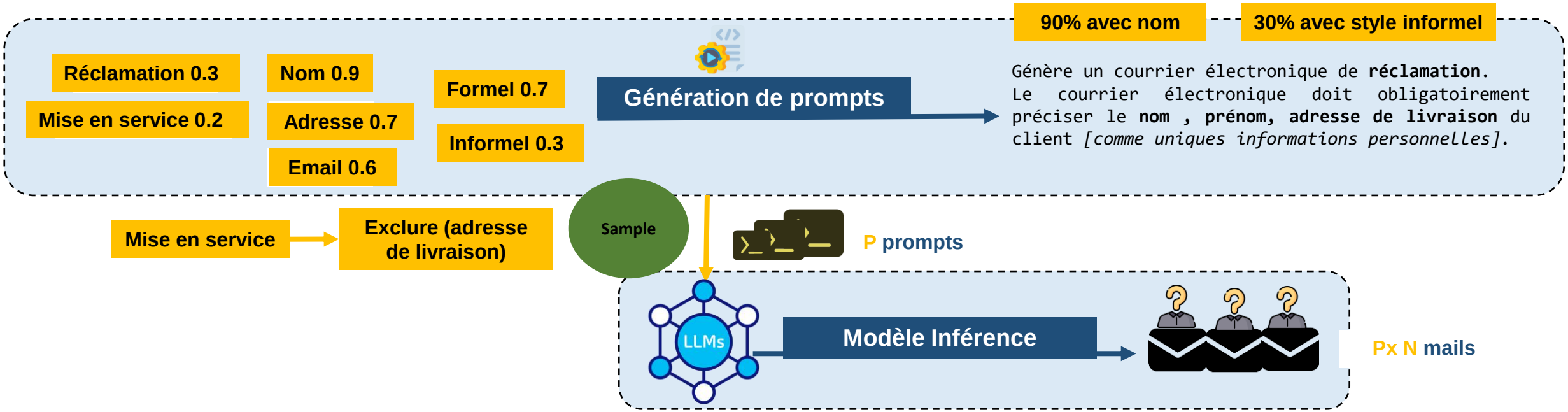
Chaîne de Traitements



Génération de prompts

Approche par distributions et corrélations

Contrôler le taux d'occurrence de chaque valeur caractéristique dans la totalité du corpus. Les corrélations sont prises en compte.



- Distributions réalistes
- Corrélations prises en compte
- Coût plus raisonnable

- Coût associé au nombre de prompts

N°3

Implémentation

Outils

Prétraitement

```python
class Email:
    def __init__(self, source, sujet, dcp, style, orthographe): ...

    def generate_instruction(self) -> str: ...

    def get_parameters(self):
        return (self.sujet, self.style, self.source, tuple(sorted(self.dcp)), self.orthographe)
```

```python
class Dataset:
    def __init__(self, config_path): ...

    def sample_parameters(self, param_name): ...

    def generate_emails(self, num_emails): ...
```

Inférence

MISTRAL AI_

MISTRAL AI_ 8X7 Mixtral of experts

8x7b de paramètres

~112 go de mémoire nécessaire

3 cartes GPU 40 go nécéssaires

Post-traitement

data.gouv.fr

MISTRAL AI_ 8X7 Mixtral of experts

DGX A100 : 8 GPUs de 40 go.

Constitution des prompts

Default prompt system

```
Tu es un générateur de courriers électroniques. Le courrier électronique est envoyé par un client à un fournisseur d'énergie.
	Le courrier électronique doit être rédigé uniquement en français.
	Toutes les informations personnelles du client que le mail peut contenir doivent être anonymisées et contenues entre crochets, par exemple [Nom], [Prénom], [Adresse].
	Produit uniquement le courrier et évite l'explication ou toute instruction supplémentaire.
```

Instruction par Thème

```
 Génère un courrier électronique rédigé par un client qui demande une mise en service, pour bénéficier du raccordement électricité ou gaz à son domicile. Dans un même courrier, l’énergie est spécifiée pour indiquer soit électricité, soit gaz, soit les deux.
```

Instruction par Style de rédaction

```
Utilisez un style de rédaction décontracté et simple.
```

Instruction par Qualité de rédaction

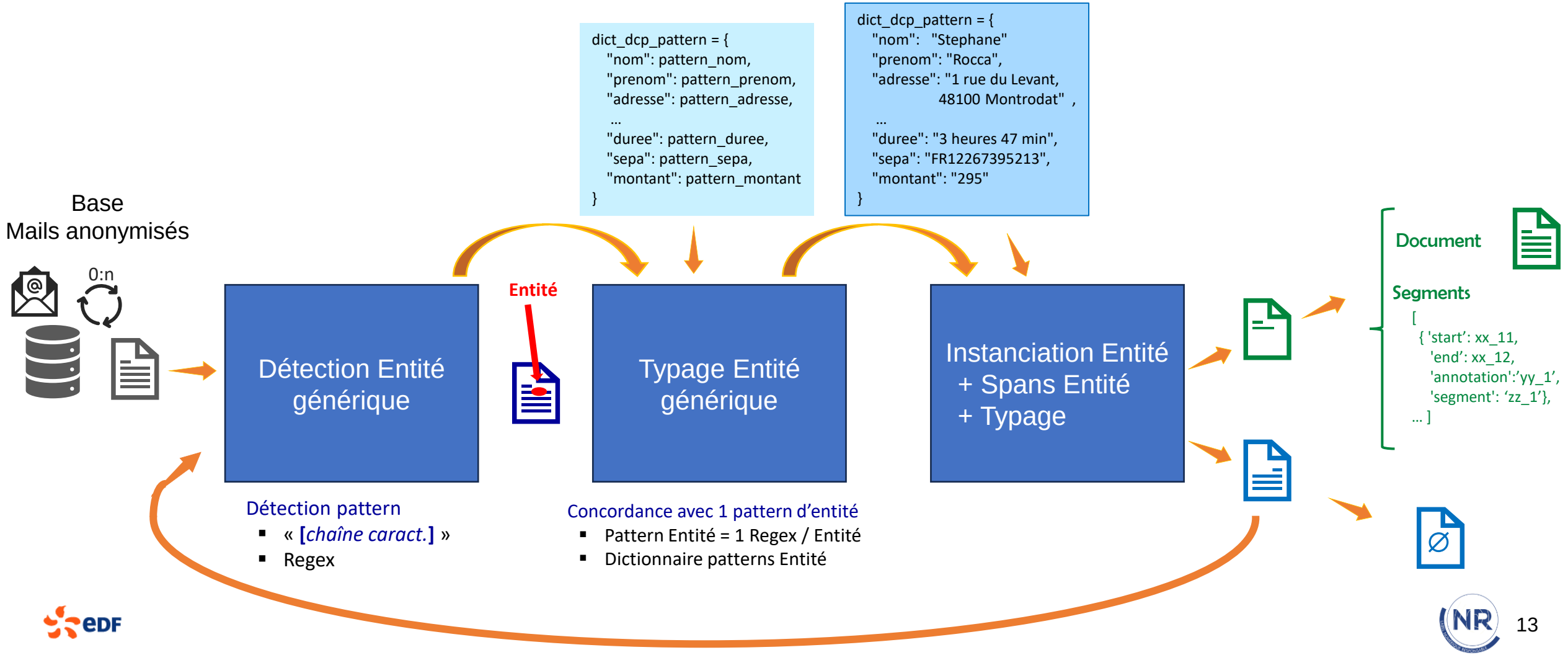
```
Ce courrier doit contenir des fautes d’orthographe.
```

N°4

Post-Traitement

Pipeline de post-traitement

Objectif : Instancier chaque mail anonymisé généré par LLM pour le réentrainement de modèles de désidentification



Base de clients fictifs

Des **données Open Source** pour instancier les entités: Nom, Prénom, Adresse

Base de données d'adresses
(https://adresse.data.gouv.fr/donnees-nationales)

- Numéro
- Nom de voie
- Code postal
- Nom commune
- Code insee
- …

Base de données de prénoms
(https://www.data.gouv.fr/fr/datasets/fichier-des-prenoms-depuis-1900/)

- Prénom : Effectif / an

Base de données de noms
(Fichier des noms | Insee)

- Noms de 1891 à 2000
- Noms attribués au moins 30 fois

Création d'un échantillon d'adresses

- Disposer d'une BD d'adresses réduite
- Restriction à 34 départements répartis sur les différentes régions
- Tirage aléatoire uniforme de 60 adresses / département

Echantillon d'adresses : **2040** adresses

Instance de nom et prénom
Tirage aléatoire pondéré

Instance d'adresse
Tirage aléatoire uniforme

Instances d'entités
Tirage aléatoire uniforme

Génération de données via regex

Principe

- Sur la base d'un ou plusieurs formats connus

- Numéro de téléphone
- Adresse e-mail (basée sur prénom et nom)
- SEPA : Single Euro Payments Area

- Numéro de client
- Numéro de contrat
- PDL : Point de Livraison (Electricité)
- PCE : Point de Comptage et Estimation (Gaz)

- Date
- Heure
- Durée

- Montant

Génération via Faker* (package Python)

- IBAN : International Bank Account Number
- RIB : Relevé d'Identité Bancaire
- BIC : Bank Identifier Code

* Welcome to Faker's documentation! — Faker 33.1.0 documentation

N°5

Résultats & Evaluation

Aperçu des résultats

Base de mails synthétiques de 1600 mails de clients particuliers.
Thématiques: Résiliation, mise en service, réclamation
Température LLM: 0.9-1.4 , Max tokens: 4000

coupure d'électricité,erreurs,informel,"nom, point_de_livraison, téléphone, contrat, RIB, numéro client BP, URL"

Sujet : Réclamation pour coupure d'électricité imprévue

Bonjour,
Je m'apelle [prénom] [nom] et j'ai eu un gros soucis avec mon elecetricité ces dernier jours. J'ai pas pu utiliser mon frigo, ma télé, mon ordi... Bref, tout ce qui est important pour un quotidien normal.

J'ai cru comprendre que vous aviez des problemes de coupure generalise sur ma zone et que mon point de livraison serait impacté. Vous trouverez ci-joint mon RIB ainsi que mon numéro de client BP: [BP]. J'espère qu'on pourra trouver une solution rapidement.

Voici mon téléphone : [Téléphone], n'hesitez pas a me contacter pour plus d'info. Et voici mon contrat: [contrat]

J'aimerais aussi recevoir une compensation financière pour cette gêne, je vous laisse donc mon RIB ci-dessus.

Cordialement,
PS: J'ai aussi un doute sur le URL de ma facture, est-ce que ca devrait etre : [URL] ?
[prénom] [nom]

Mille merci pour votre aide,

Sujet : Réclamation pour coupure d'électricité imprévue

Bonjour,
Je m'apelle **Sheima BORTLA CART** et j'ai eu un gros soucis avec mon elecetricité ces dernier jours. J'ai pas pu utiliser mon frigo, ma télé, mon ordi... Bref, tout ce qui est important pour un quotidien normal.

J'ai cru comprendre que vous aviez des problemes de coupure generalise sur ma zone et que mon point de livraison serait impacté. Vous trouverez ci-joint mon RIB ainsi que mon numéro de client BP: **6347 091 032**. J'espère qu'on pourra trouver une solution rapidement.

Voici mon téléphone : **06 / 82 / 97 / 63 / 05**
 n'hesitez pas a me contacter pour plus d'info. Et voici mon contrat: **007832294376**

J'aimerais aussi recevoir une compensation financière pour cette gêne, je vous laisse donc mon RIB ci-dessus.

Cordialement,
PS: J'ai aussi un doute sur le URL de ma facture, est-ce que ca devrait etre : **https://smqxr.edf.org**?

Mille merci pour votre aide,

Métriques d'évaluation

Conformité qualitative

Méthode

Evaluation humaine réalisée par 2 annotateurs sur un échantillon de 410 mails (~25% du corpus), tirés de manière aléatoire du corpus, en considérant 3 températures différentes du LLM (0.9, 1.2, 1.4).

l'évaluation consiste à comparer les consignes du prompt et le résultat obtenu sur 4 critères: sujet, langue, orthographe et style.

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.

Résultats

Langue et sujet: conformité quasi-parfaite

Style informel: meilleurs résultats avec des températures élevées

Orthographe: difficultés à générer des erreurs pour le style formel, et globalement pour des températures basses.

Conclusion

Les températures élevées du LLM permettent de mieux diversifier les générations, améliorer l'adhérence à des styles informels et réalisme.

Métriques d’évaluation

Conformité de l’intégration des entités

Méthode

- Vérifier le respect de la distribution des entités dans le corpus généré.
- La distribution des entités à un impact sur l’entraînement des systèmes IA.
- Comparaison entre les probabilités en entrée (fichier de configuration) et la fréquence d’apparition des entités dans le corpus généré.

Résultats

- Bonne conformité globale
- Léger déséquilibre dans l’inclusion des informations bancaires (RIB, BIC, IBAN)

Conclusion

La phase de prétraitement a permis une génération de prompts diversifiée qui reflète les distributions d’un corpus réel

Le LLM respecte généralement les consignes d’inclusion d’entité dans la génération

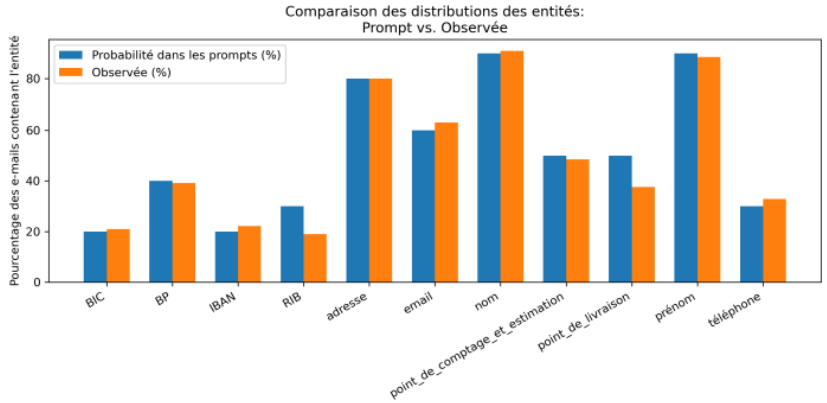


FIGURE 5 – Comparaison des distributions des entités dans les prompts et dans la base des e-mails synthétiques générée. La majorité des distributions sont respectées à part un léger déséquilibre sur les informations bancaires.

Métriques d'évaluation

Performances de génération

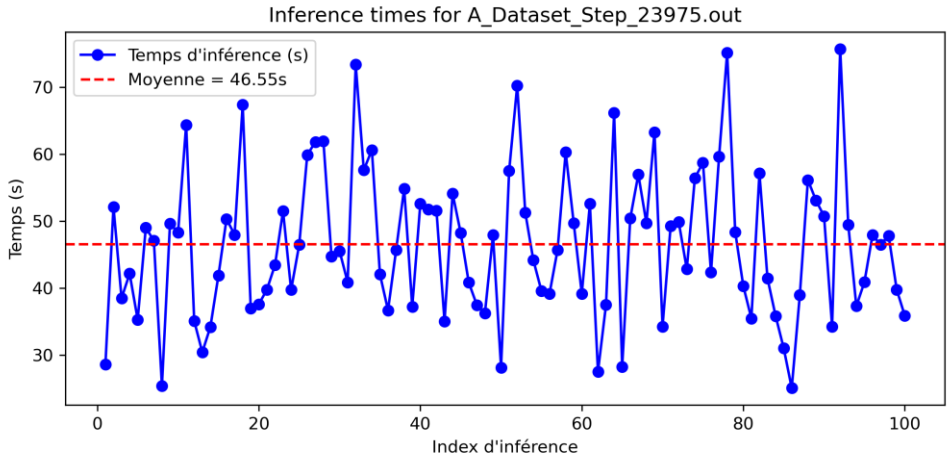
Méthode

Utilisation de la librairie **MLFlow** pour le *tracking* du temps d'exécution.

Usage de 4-5 GPUs pour paralléliser la phase de génération (la plus couteuse).

Résultats

- Moyenne d'inférence (34 sec).
- Corrélation entre la longueur de texte généré et le temps d'inférence
- Le temps d'inférence = **98%** du temps total
- Chargement du modèle **67 sec** (one shot, 4 GPUs)
- Post-traitement: instanciation des entités **~2min pour 1600** mails



Coûts

- Travaux sur infrastructure locale. Estimation effectuée des coûts directs sur une solution cloud équivalente.
- Coût horaire d'utilisation d'un GPU estimé en moyenne à 3,7$ ⇨ Coût approximatif de l'utilisation de 5 GPU pour un cycle de génération de 15h ≈ 278$

Annotation manuelle

- Annotation : équipe d'annotateurs pour une annotation croisée et une étape de validation
- Temps annotation cumulé / e-mail ≈ 2 à 3 min ⇨ 1600 e-mails pouvant atteindre 80 heures

⇨ Génération d'e-mails compétitive

N°6

Discussion & Perspectives

Discussion

Utilité du dataset de données synthétiques

Les modèles de désidentification ont été réentraînés en utilisant les données synthétiques, avec une perte de performance de 15 points comparativement **aux meilleurs modèles à disposition**.

→ L'usage d'un **dataset mixte données réelles/synthétiques** est envisagé pour obtenir de meilleures performances tout en améliorant l'application du principe de minimisation du RGPD.

Statut des données synthétiques

Le statut **pseudonyme** des données synthétiques créées dans ces travaux **semble incontestable** car elles sont synthétisées de manière désidentifiées avec insertion d'entités factices.

Les LLMs sont entraînés sur des données réelles et le risque de réidentification des individus issus de ces données lors de la génération est existant.

- **Le statut « anonyme » des données synthétiques est discutable** : cela nécessite des approfondissements pour quantifier la probabilité de réidentification dans ce type de situation.

Conclusion

- Mise au point d'une chaîne de génération d'e-mails synthétiques de bout en bout
- Axée sur le marché des clients particuliers, pour un format de mails libres, intégrant de nombreuses entités **à désidentifier** avec un enrichissement dans la suite des travaux.
- Diversité des e-mails en termes de thématiques et de formulation (formel / informel, avec ou sans fautes d'orthographes)
- Compromis entre créativité (température) du LLM et complexité des post-traitements
- Les résultats de l'évaluation qualitative et qualitatives confirment l'efficacité de la méthodologie implémentée pour contrôler la génération des LLMs.

Perspectives

Amélioration de l'adhérence aux prompts

Intégration d'un feedback **suite à analyse des forces et faiblesses des modèles entraînés sur les données synthétiques.**

Intégration de nouveaux formats de données

Intégrer les échanges entre client et conseiller (conversations)
Intégrer d'autres canaux de contact : commentaires, enquêtes satisfaction,…

Extension des périmètres couverts

Intégrer le marché d'affaire (clients business) et ses entités

Renforcement de la diversité du corpus

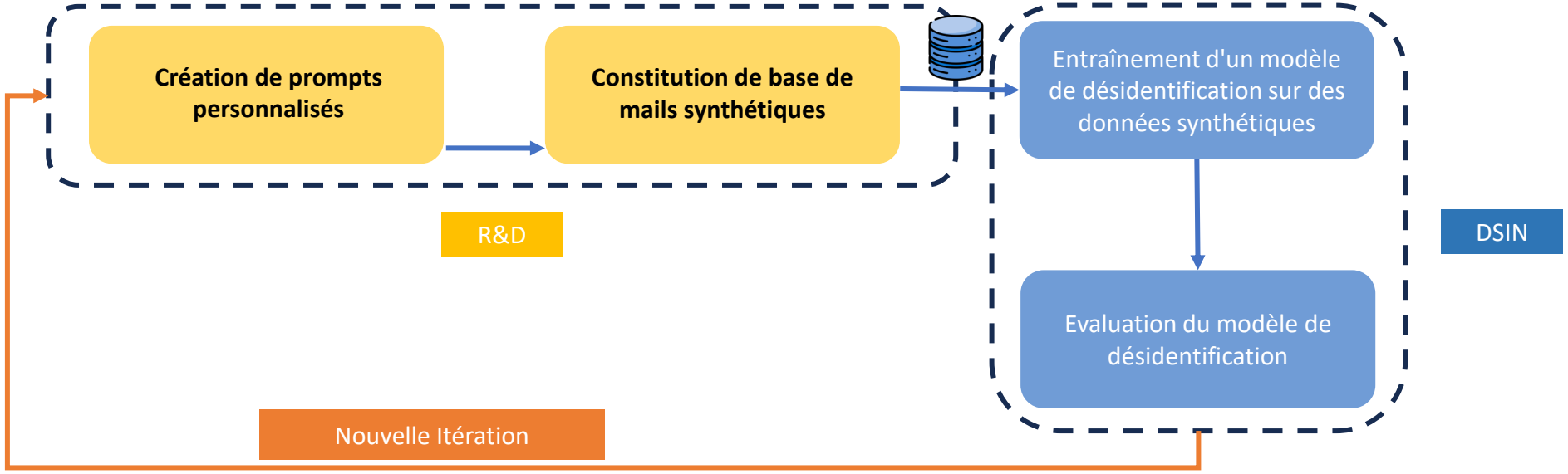
Etendre le traitement des co-références pour augmenter l'utilité de la base générée pour l'entraînement d'autres systèmes IA (généralisation)

Intégrer un guide métier pour le vocabulaire employé (tone of voice) via RAG pour augmenter la conformité/réalisme


Merci

Contexte

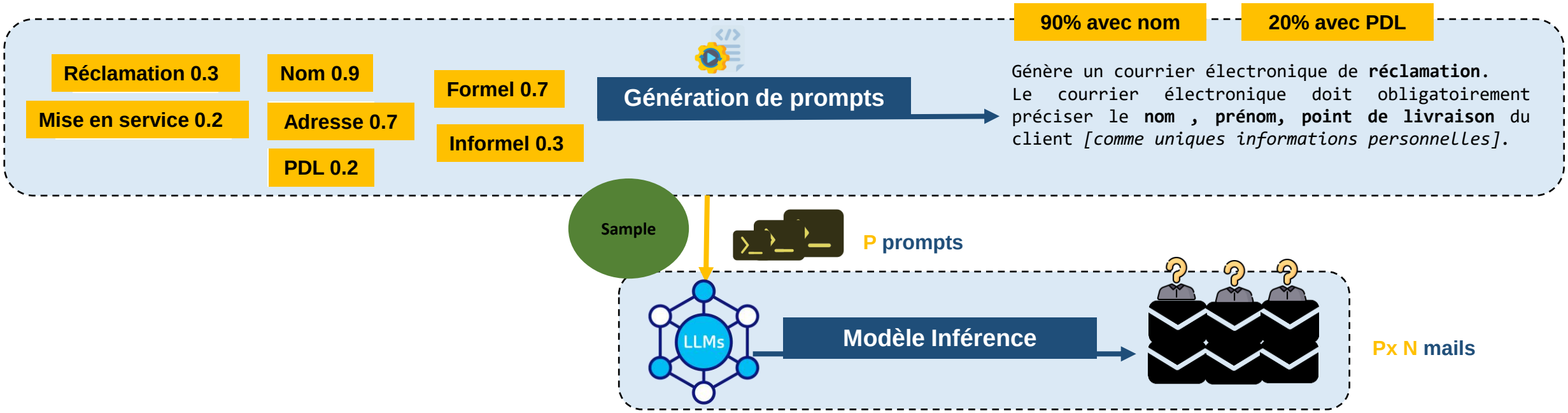
Approche



Génération de prompts

Optimisation du nombre de prompts

Contrôler le taux d'occurrence de chaque valeur caractéristique dans la totalité du corpus. Les corrélations sont prises en compte.



- Réduction du temps de tokenization
- Respect des distributions

- Garder une taille d'échantillon permettant une bonne couverture des caractéristiques et le respect de la distribution
- Être vigilant aux considérations de KV cache (compromis processing-mémoire)

Constitution des prompts

Configuration de la distribution (entrée)

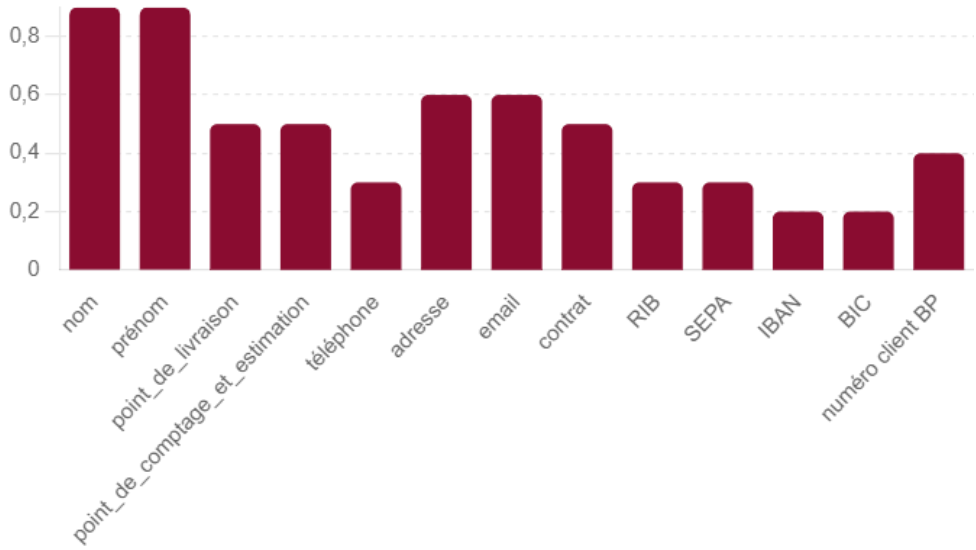
La prise en compte des corrélations implique de considérer des co-occurrences des entités avec des thèmes ou des styles de rédaction spécifiques

donnees-synthetiques-mails > src > inference > conf > ! Distribution.yaml

```yaml
dcp:
  - name: "adresse"
    probability: 0.6
  - name: "email"
    probability: 0.6
  - name: "contrat"
    probability: 0.5
  - name: "RIB"
    probability: 0.3
  - name: "SEPA"
    probability: 0.3
  - name: "IBAN"
    probability: 0.2
  - name: "BIC"
    probability: 0.2
  - name: "numéro client BP"
    probability: 0.4

styles:
  - name: "formel"
    probability: 0.7
  - name: "informel"
    probability: 0.3

orthographe:
  - name: "erreurs"
    probability: 0.1
  - name: "sans_erreurs"
    probability: 0.9
```




```yaml
correlations:
  - sujet: "mise en service"
    exclude_dcp: ["point_de_livraison", "Point_de_comptage_et_estimation"]
  - style: "informel"
    exclude: ["contrat"]
```

Prise en compte des corrélations lors de la création des prompts

Discussion

Contrôle de la génération des LLM

- Une première évaluation sur la base **du respect des instructions de prompts** par le LLM et non sur l'adéquation du LLM à des tâches aval telles que la reconnaissance d'entités nommées ou autres

Langue

- Adhérence marquée à l'anglais de **Mistral 7b**. Une nette amélioration avec **Mixtral 8x7b** pour la génération en français.

Formalité/ orthographe

- Des LLM le plus souvent entrainés pour produire des contenus formels sans erreurs . Des tests de différentes températures avec un impact sur la qualité des textes générés

Hallucinations

Peu fréquentes ; insertion de codes Python, reprise de parties d'instructions de prompts dans les e-mails, succession d'e-mails dans un même e-mail, répétition en fin d'e-mail de phrases du corps d'e-mail sous forme de « Note : *une phrase de l'e-mail* » ⇨ Mise en place d'une étape de curation des e-mails générés lors du post-traitement

Performance et coûts

Coûts

- Travaux sur infrastructure locale. Estimation effectuée des coûts directs sur une solution cloud équivalente.
- Coût horaire d'utilisation d'un GPU estimé en moyenne à 3,7$ ⇨ Coût approximatif de l'utilisation de 5 GPU pour un cycle de génération de 15h ≈ 278$

Annotation manuelle

- Annotation : équipe d'annotateurs pour une annotation croisée et une étape de validation
- Temps annotation cumulé / e-mail ≈ 2 à 3 min ⇨ 1600 e-mails pouvant atteindre 80 heures

⇨ Génération d'e-mails compétitive