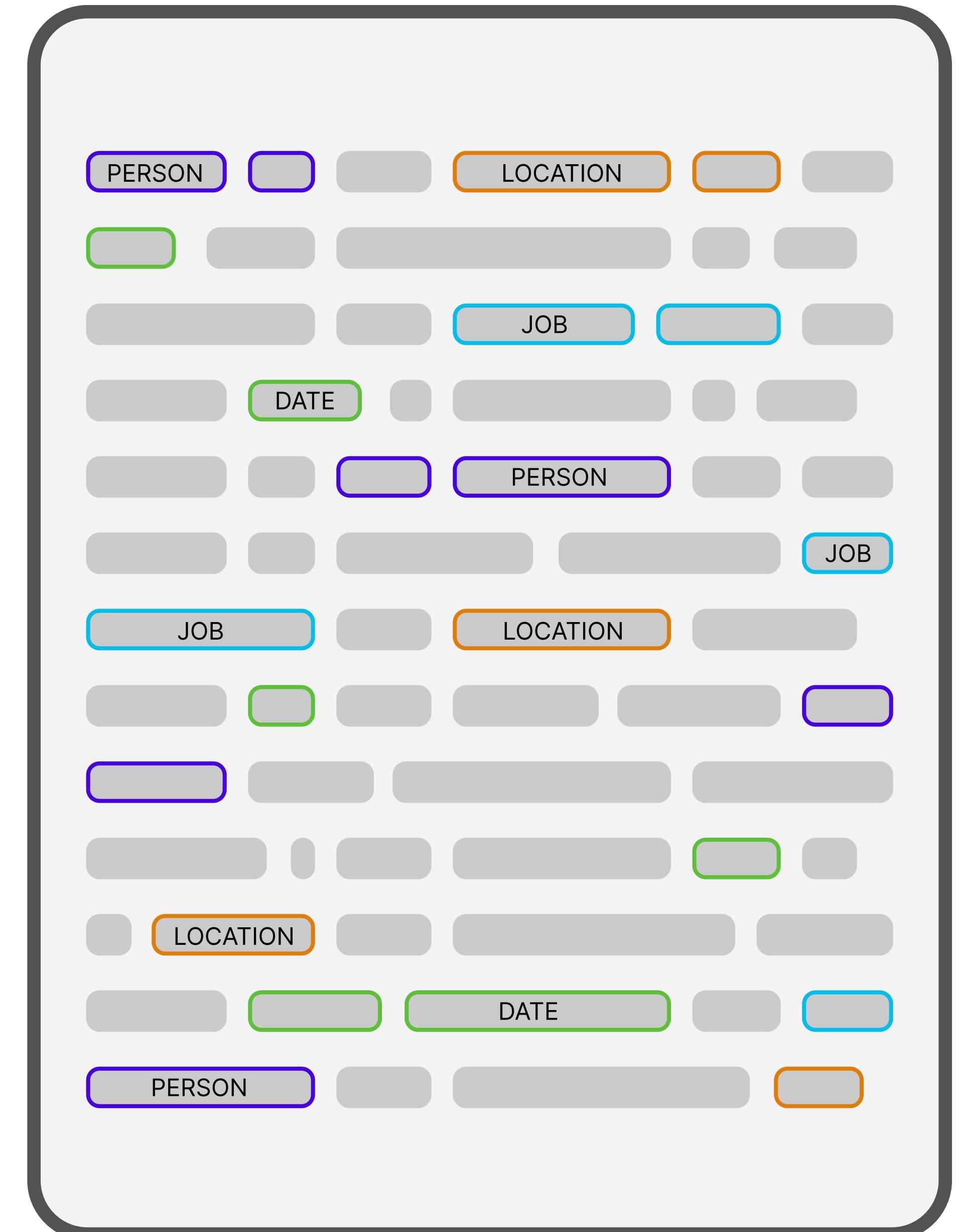


Extraction d'entités nommées dans le cadre du projet EXO-POPP

Valentin BONSI - M1 SID - 2022 2023

02/06/23



Sommaire

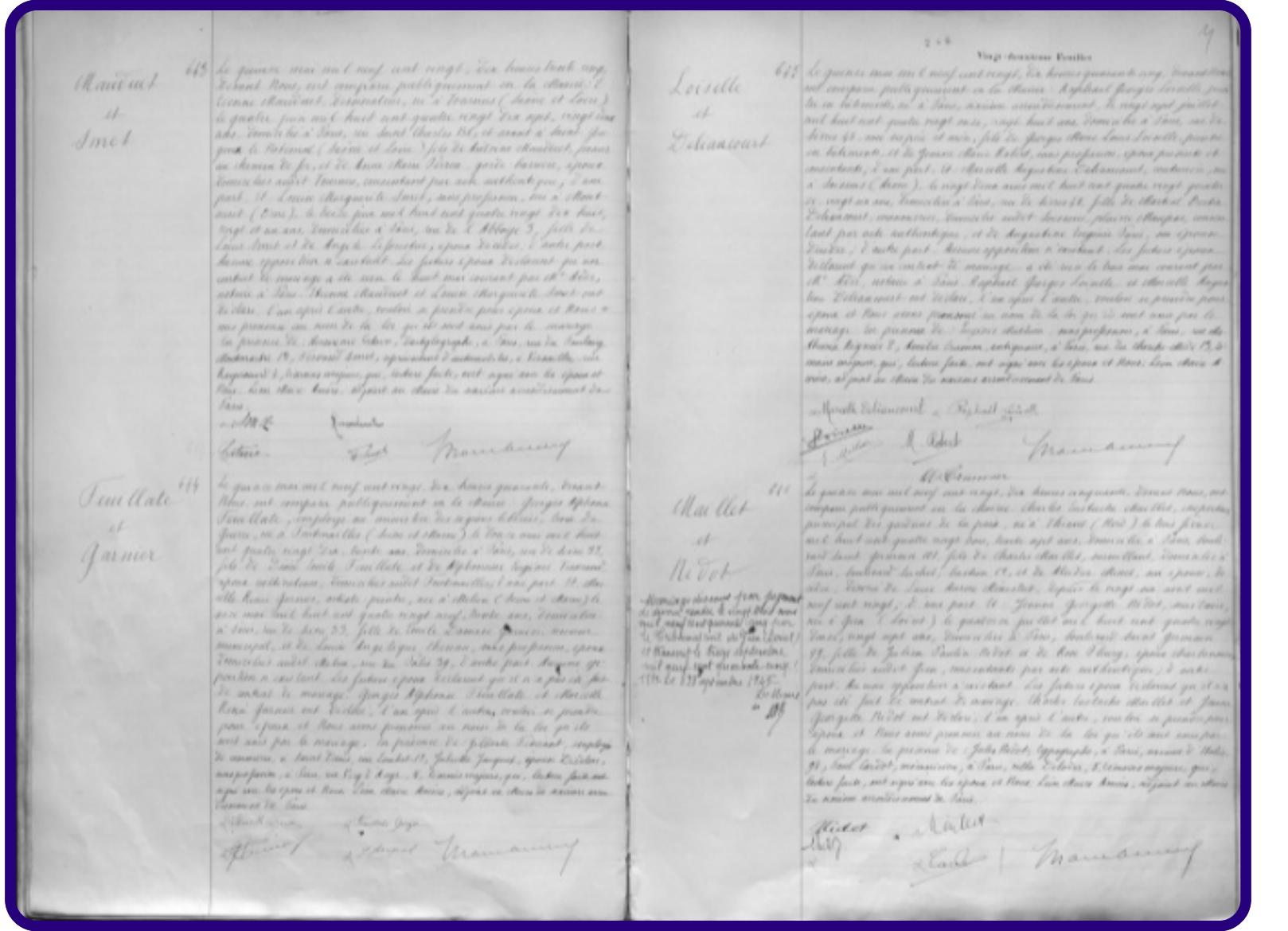
Le projet **EXO-POPP**

La librairie **FlairNLP**

Les limites d'une approche classique

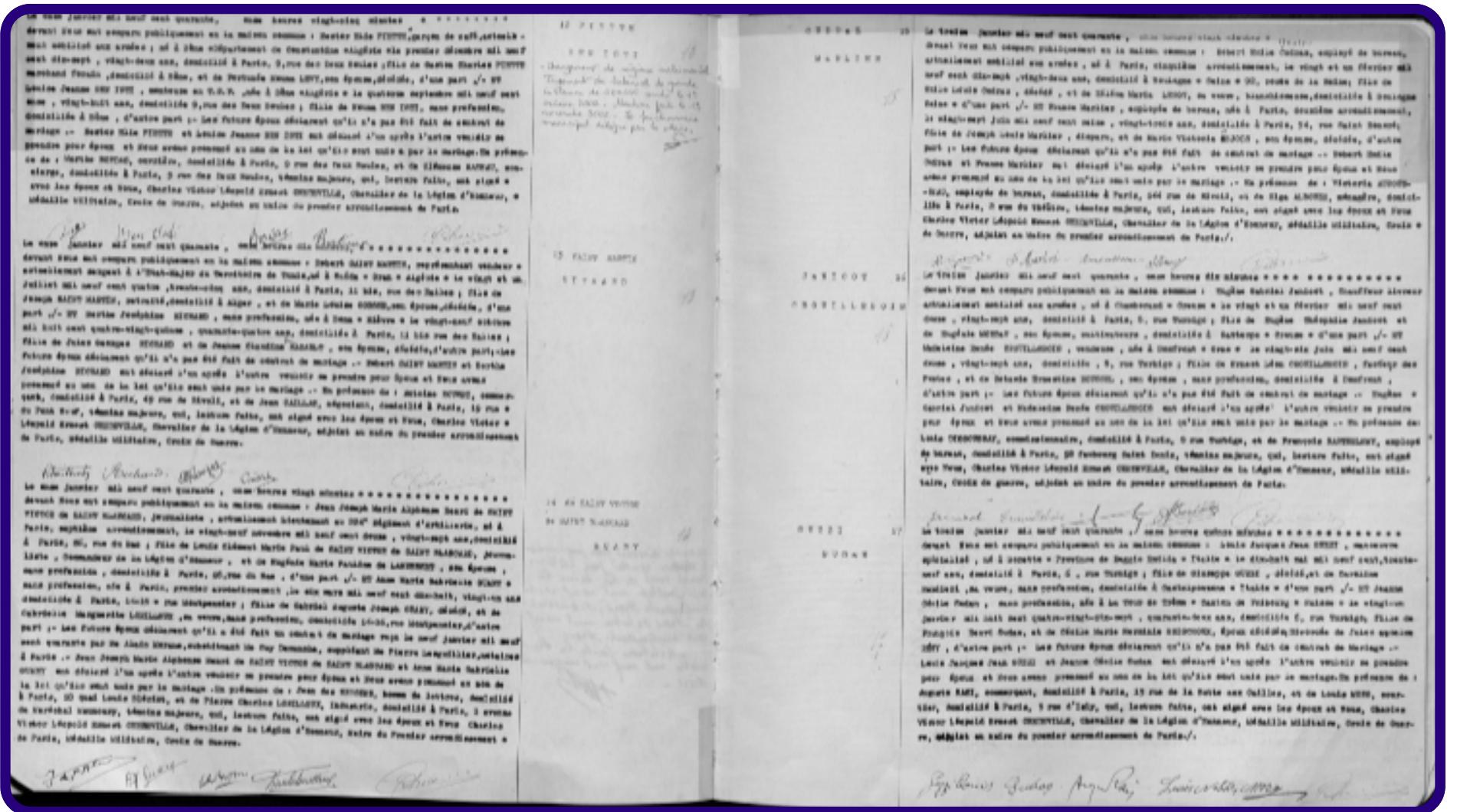
GPT3.5 pour de l'extraction d'entités nommées

Le projet EXO-POPP



Objectif : Passer des registres d'état-civil papier à une grande base de données

Contenu : Les actes de mariage de Paris et sa banlieue de 1880 à 1940

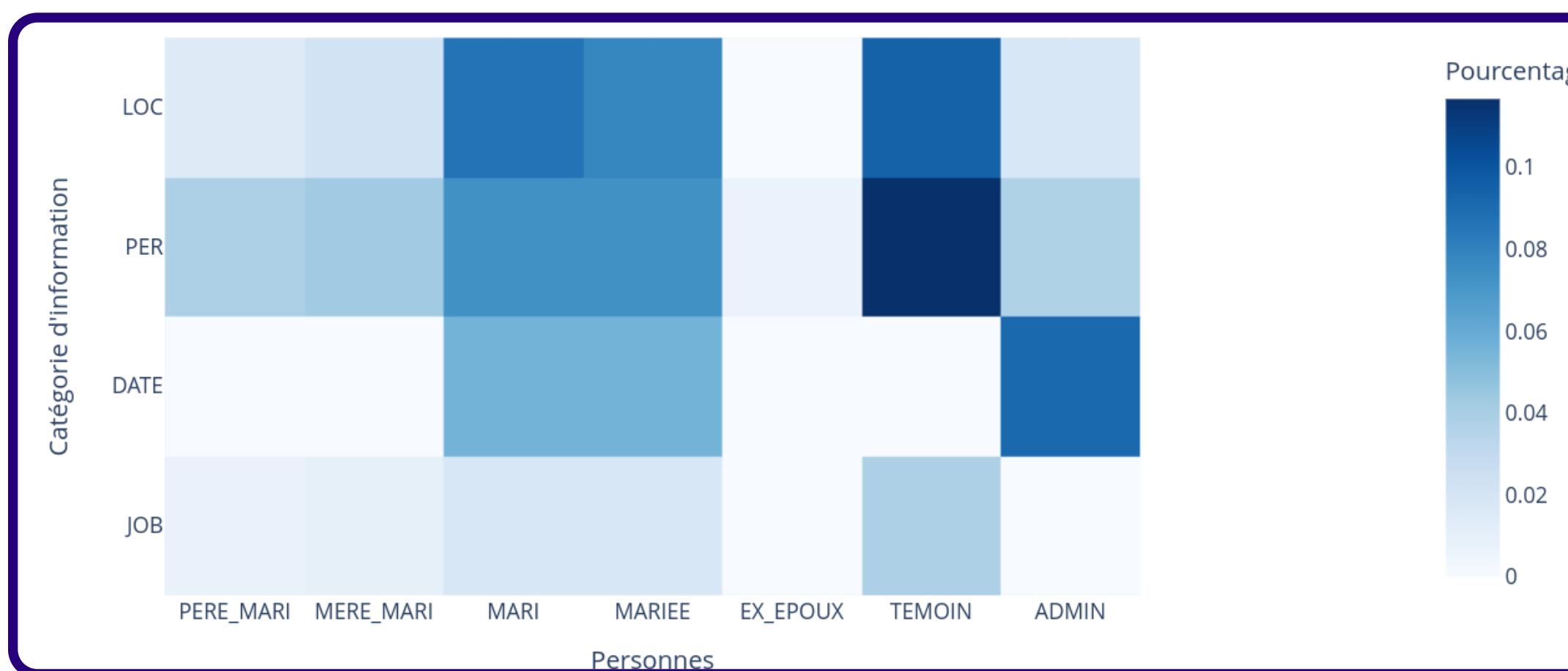


Taille : environ 300 000 scans d'actes manuscrits et tapuscrits

Tâche : Reconnaissance manuscrite puis comprehension du texte

Le projet EXO-POPP

Niveau	Tags				
1	Administratif	Mari	Epouse	Témoin	
1.1	Père	Mère	Ex-Epoux		
2	Naissance	Residence	Décès	Disparu	Age Veuf
3		Profession	Prenom	Nom	
4	Pays	Département	Ville	Numéro voie	Type voie Nom voie
	Année	Mois	Jour	Heure	Minute



→ Les Tags sont classés par niveaux : *Nom → ex-epouse → mari*

→ Il peut y avoir des changements dans les formulations, comme lors d'un décès, d'un scripteur différent ou des cas de pluriels :

...Denis, résidant 8 rue du palais, avec ses pere et mere...

...Homer et Marguerite, marchands de bestiaux...

...fils de Simon, decede, et Ophelie, résidant au 7 place de l'arche...

...fille de Bertrand et de Julie, sa veuve, résidant au 4 boulevard Pascal...

→ Les informations à relever sont nombreuses, souvent complexes et apparaissent parfois rarement.

neuf cent douze, vingt-sept ans, domicilié à Paris, 16, rue Hérole: fils de Ernest CARDOT et
de Julie GANIVET, son épouse, retraités, domiciliés à Saint Amand de Puisaye * Nièvre *

Étiquettes utilisées dans : retraités,

< MARI ✕ PÈRE ✕ PROFESSION ✕ EPOUSE ✕

La librairie flair

Embedding : représentation vectorielle d'un mot capturant des informations synthaxiques, sémantiques, contextuelles ou encore spatiales sur celui-ci.

„Mr. Powell finds it easier to take it out of mothers, childrens and sick people than to take on this vast industrie,“ Mr. Brown commented icily. „let us have a

Un exemple du dataset IAM utilisé pour la prochaine experience

Quelques embeddings importants accessibles via la librairie :

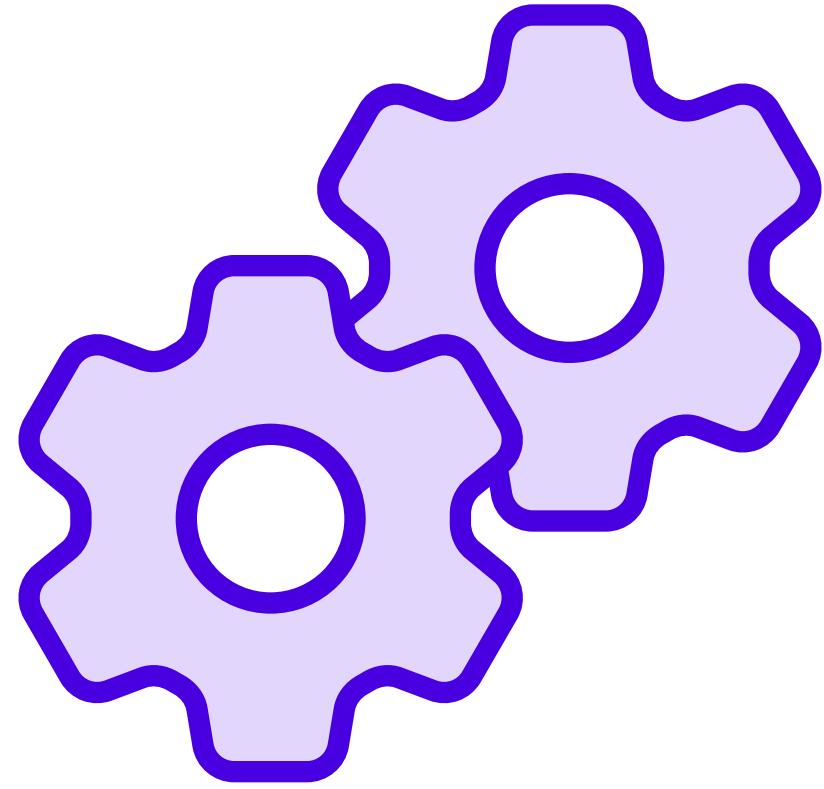
- **Flair Embeddings (2018)** : Un modèle BiLSTM basé sur des embeddings encodant des sous-mots dans leur contexte.
- **BERT (2018)** : Un modèle d'embeddings basé sur les Transformers.
- **XLM-RoBERTa (2020)** : Une grosse évolution de BERT, multilingue, avec plus de paramètres, et entraînée sur plus de données.

La librairie flair

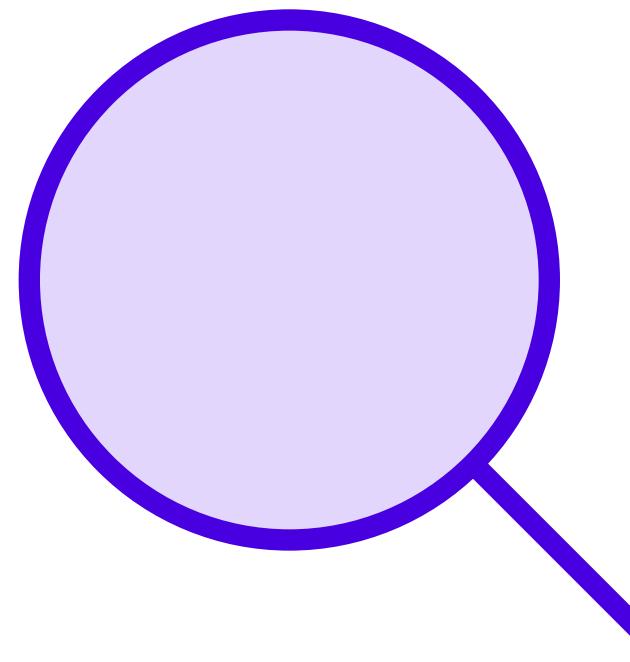
Modèle	FLAIR embeddings			BERT			RoBERTa			
Métrique	precision	recall	F1	precision	recall	F1	precision	recall	F1	support
PERSON	0.84	0.87	0.85	0.90	0.92	0.91	0.96	0.95	0.95	448
LOCATION	0.80	0.79	0.80	0.86	0.85	0.85	0.89	0.93	0.91	326
CARDINAL	0.79	0.77	0.78	0.84	0.85	0.84	0.88	0.87	0.88	237
TIME	0.63	0.66	0.64	0.68	0.78	0.73	0.78	0.86	0.82	191
ORG	0.51	0.54	0.52	0.55	0.65	0.60	0.76	0.78	0.77	166
NORP	0.75	0.81	0.78	0.79	0.88	0.83	0.88	0.93	0.90	114
micro avg	0.75	0.76	0.76	0.80	0.84	0.82	0.88	0.90	0.89	1489
macro avg	0.62	0.64	0.63	0.66	0.70	0.68	0.83	0.80	0.80	1489
weighted avg	0.75	0.77	0.76	0.80	0.84	0.82	0.88	0.90	0.89	1489

- Les modèles à base de Transformers sont largement supérieurs.
- RoBERTa est bien plus performant, mais bien plus lourd (330M params)

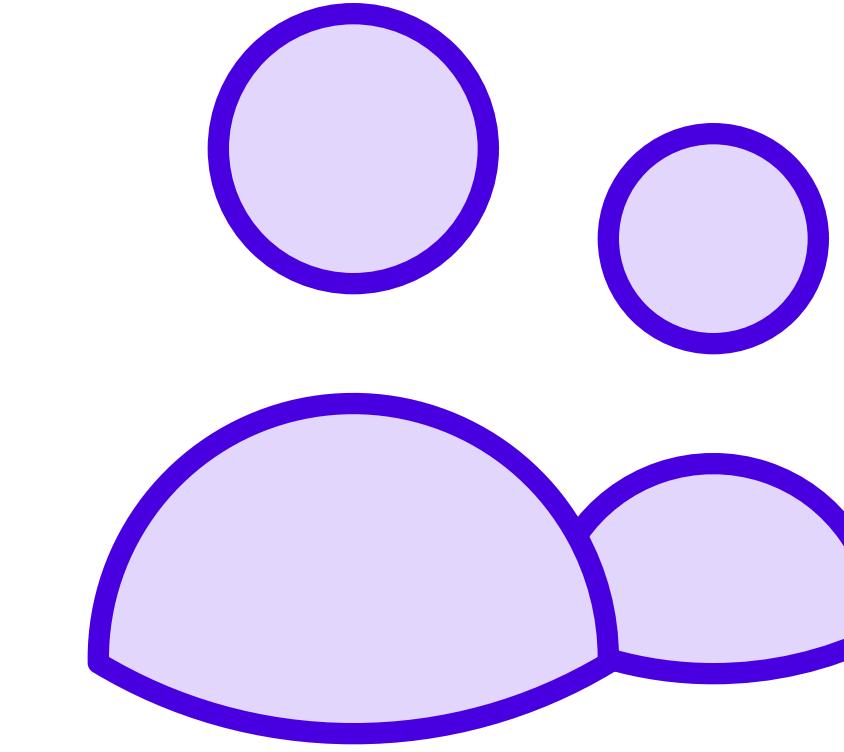
Les limites d'une approche classique



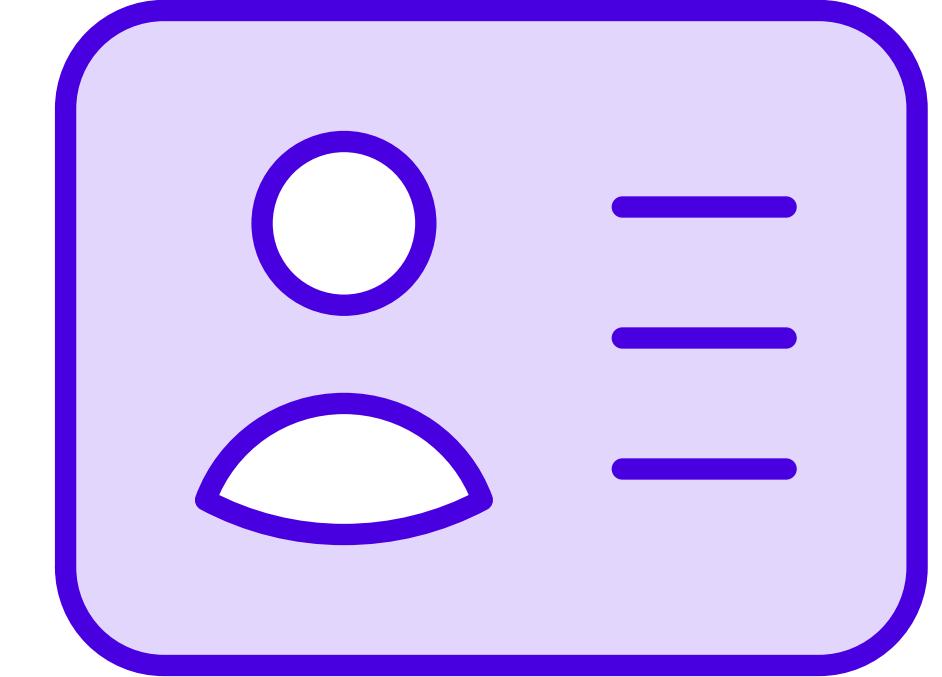
Demandeur en ressources



Les entités rares posent problème



Les données doivent être annotées

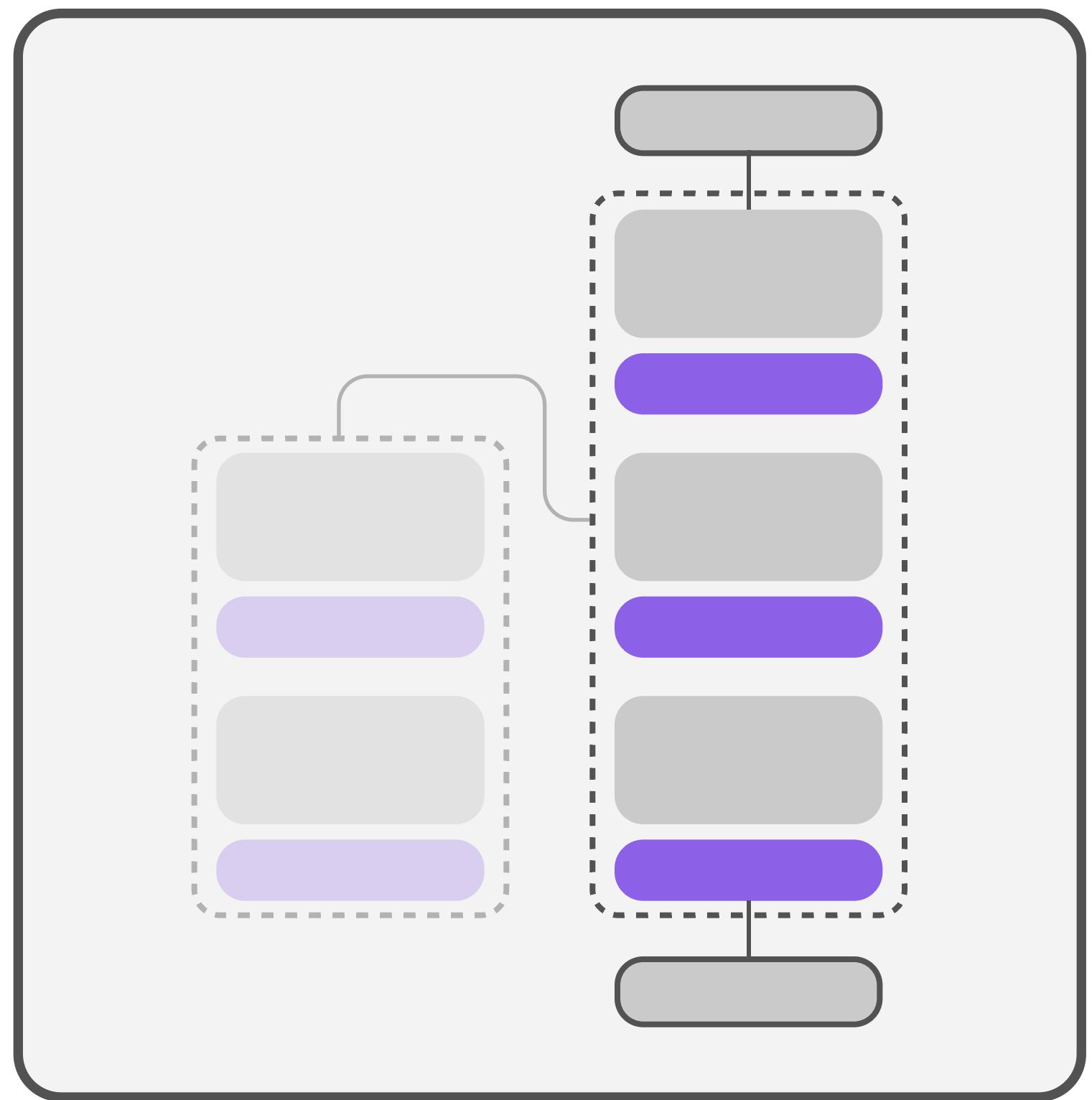


Les tags complexes posent problème

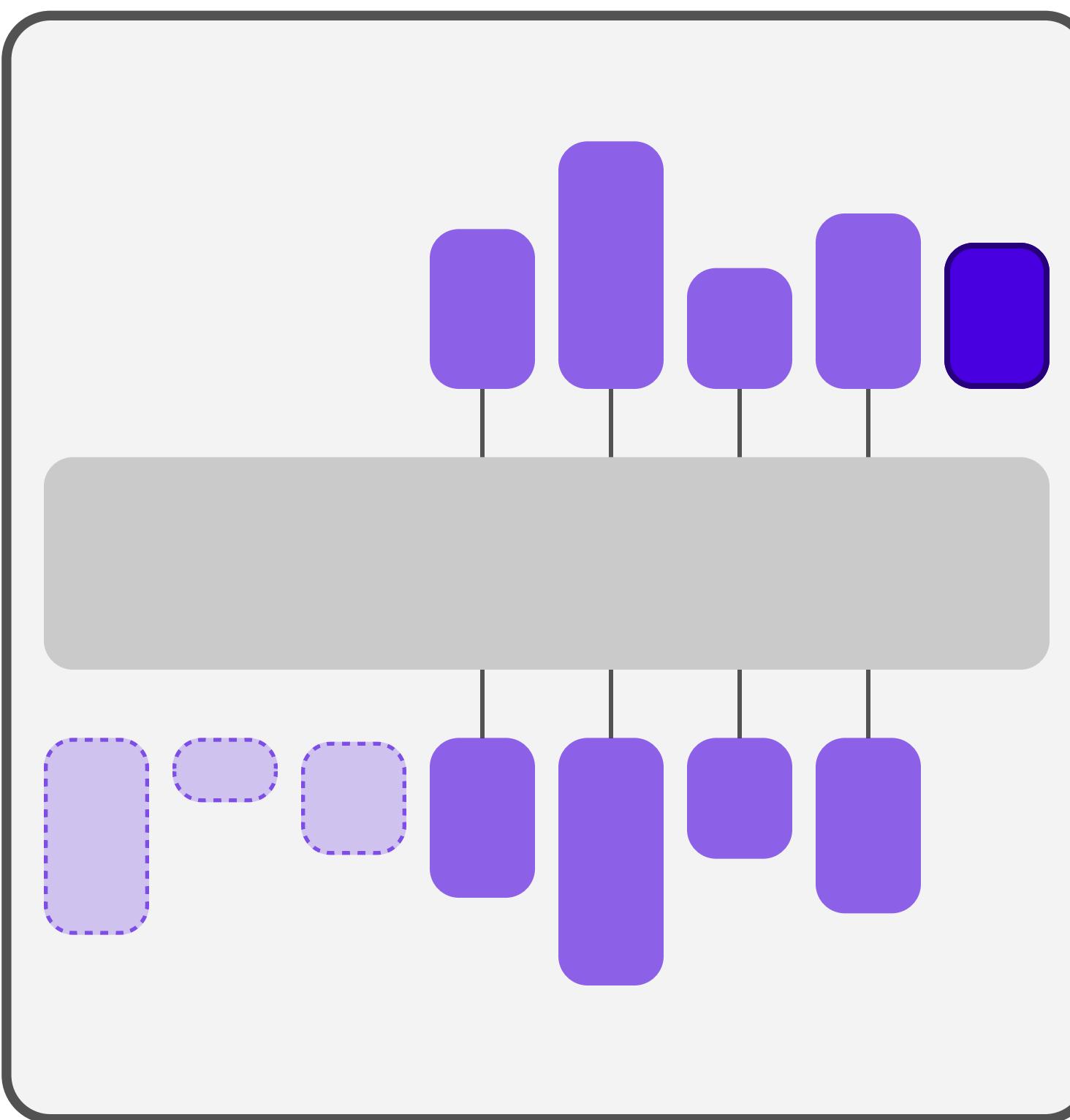
→ Le projet EXO-POPP pose problème dans ces quatres domaines

→ Les Large Language Models sont une piste à explorer

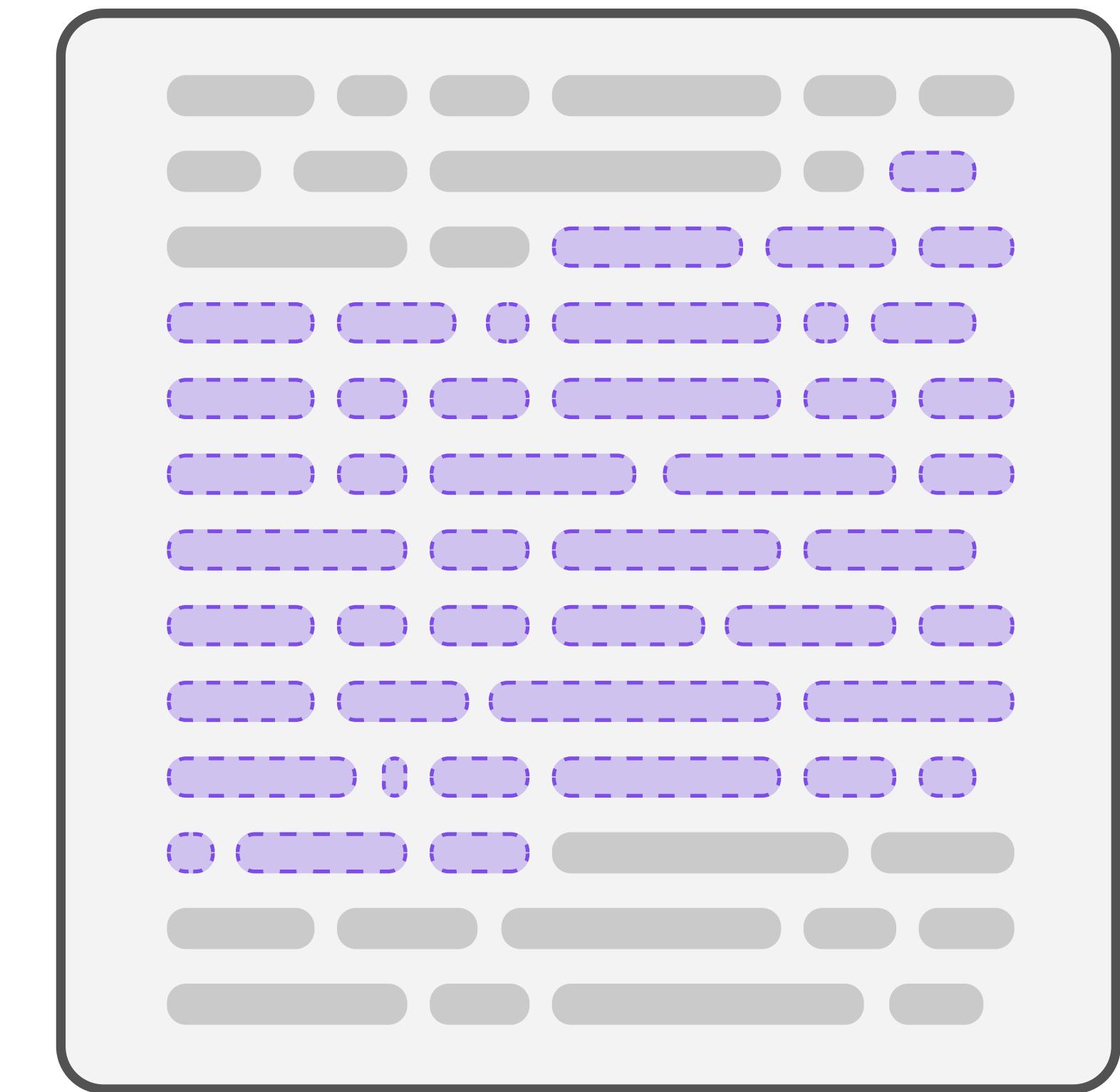
GPT, C'EST QUOI ?



Transformers
Decoder



Generation
Autorégressive

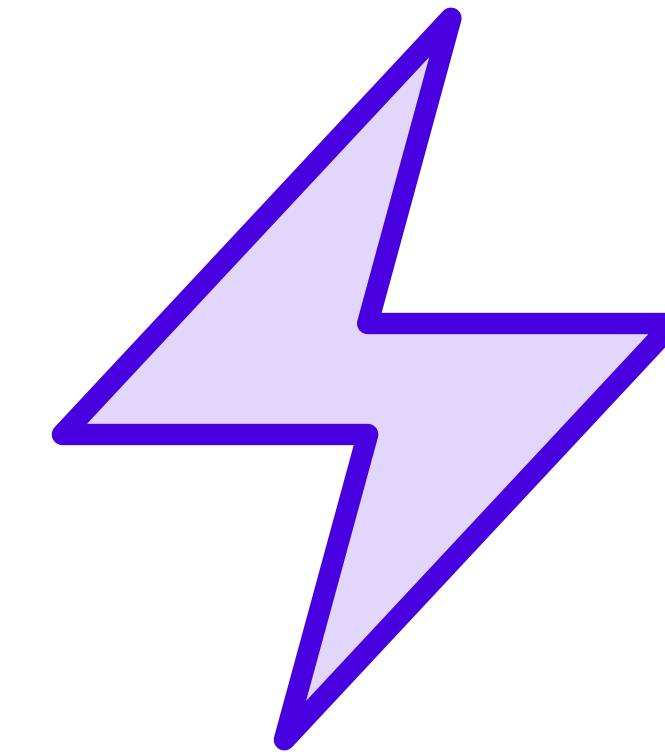


Grande fenêtre
de contexte

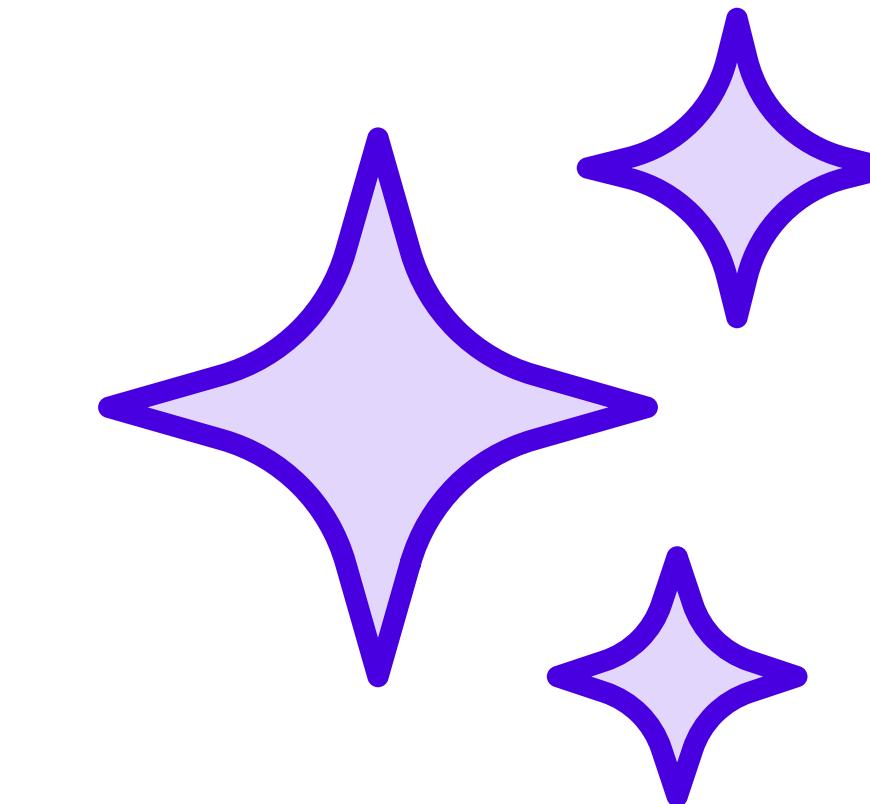
GPT3.5, EN QUOI C'EST INTERRESSANT ?



**Excellent adaptation
à une tâche**



**Besoin de très peu
d'exemples**



**Détournement des
capacités génératives**

⇒ Plutôt que de taguer des éléments, pourquoi ne pas générer directement les réponses à ce que l'on cherche ?

NER AVEC GPT3.5 - One-Shot Learning

Initialisation :

Tâche : extraire les entités comme dans l'exemple ci-dessous.

Le neuf janvier mil neuf cent quarante, onze heure...

Edmond Henri Celestin JOSEPH,
Chauffeur livreur, actuellement
mobilisé aux armées, né à Manerbe
*Calvados * le vingt- cinq avril mil neuf
cent onze, vingt-huit ans...

...Officier de l'état-civil du premier
arrondissement de Paris, Chevalier de
la Légion d'Honneur.

Labels :

Jour-Mariage : neuf

Mois-Mariage : janvier

...

Ville-naissance-mari : Manerbe

Utilisation :

Le vingt-sept mars mil neuf cent
quarante dix heures quarante-cinq
minutes...

Pierre DESTREGUIL, menuisier,
actuellement sapeur-Pompier, né à
Thenon * Dordogne * le onze décembre
mil neuf cent dix-neuf, vingt ans...
...Officier de l'état-civil du premier
arrondissement de Paris, Chevalier de
la Légion d'Honneur.

Labels :

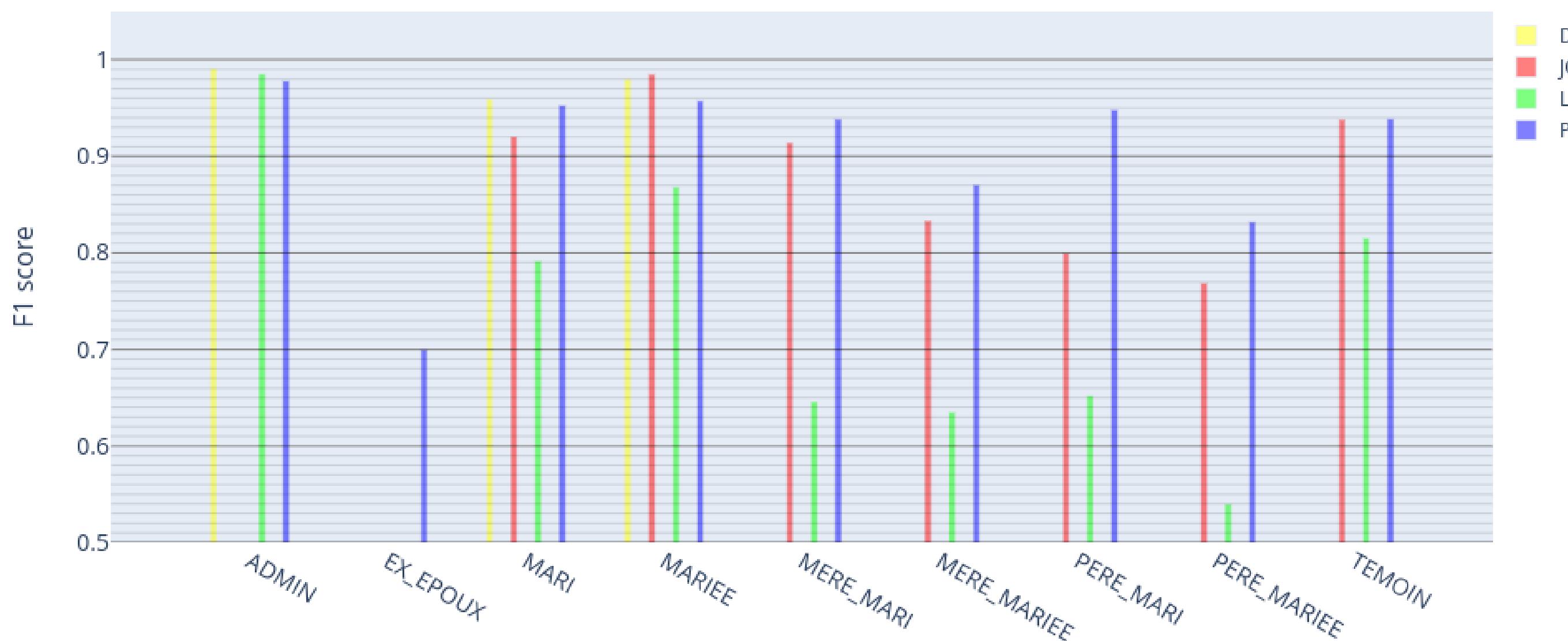
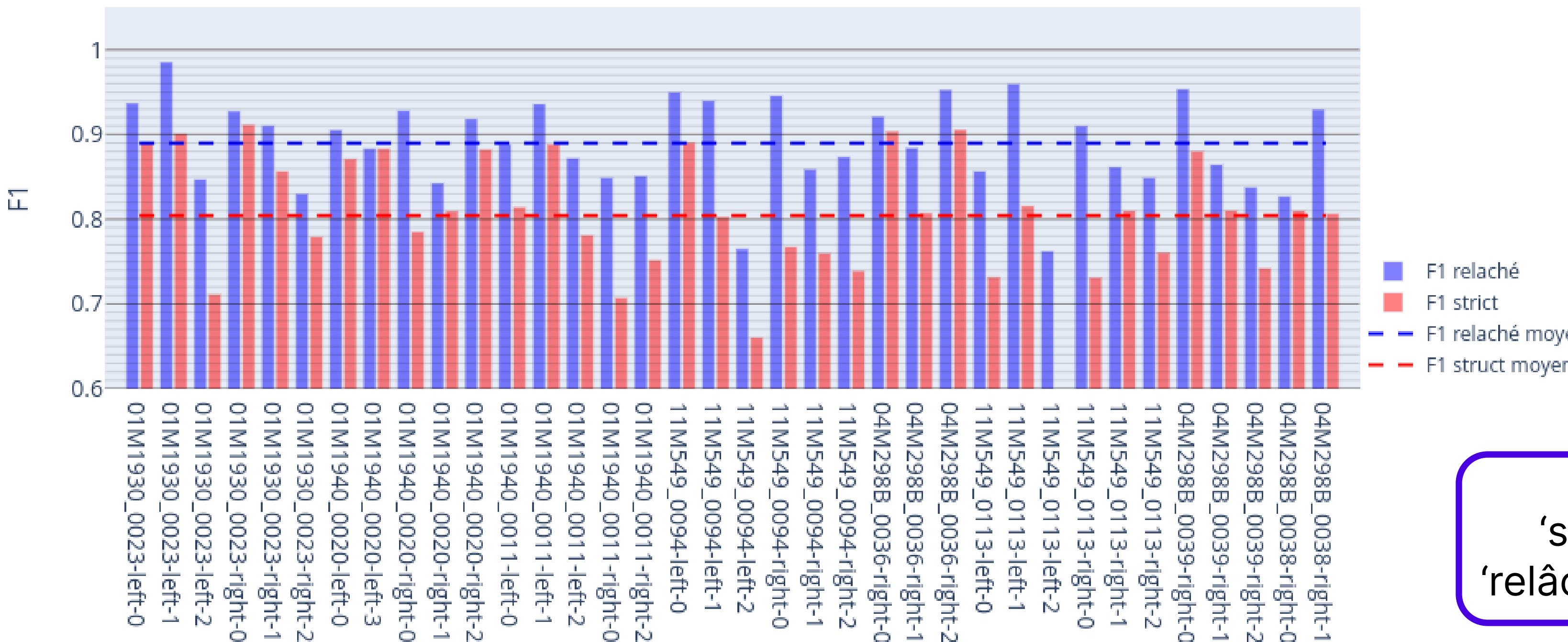
Jour-Mariage : vingt-sept

Mois-Mariage : mars

...

Ville-naissance-mari : Thenon

NER AVEC GPT3.5 - One-Shot Learning



Mesures :
'stricte' → aucune différence autorisée
'relâchée' → Distance d'édition max. de 30%

NER AVEC GPT3.5 - One-Shot Learning

Problèmes rencontrés :

Certaines réponses dépassent la limite de 4000 Tokens : Les réponses sont incomplètes.

On remarque des difficultés (Score F1 faible) sur les Localisations : il y a beaucoup d'erreurs sur des pluriels. Le modèle ne peut pas connaître toutes les formulations.

Marcel et Jeanne, marchands : metier-mari : "" / metier-mariee : "marchands"

Comment faire mieux ?

En donnant plusieurs exemples variés.

Comment ne pas dépasser la longueur maximale ?

En réduisant la taille des exemples.

NER AVEC GPT3.5 - Few-Shot Learning

Les actes tapuscrits (entre 1930 et 1940) sont toujours organisés avec la même structure :

Paragraphe 1 5 labels

Le jour, la date et l'heure du mariage

Paragraphe 2 37 labels

Toutes les infos concernant le mari et sa famille

Paragraphe 3 37 labels

Toutes les infos concernant la mariée et sa famille

Paragraphe 4 0 label

La présence d'un contrat de mariage et les noms des mariés

Paragraphe 5 21 labels

Toutes les infos concernant les témoins et l'adjoint au maire

→ En séparant les actes selon celle-ci, on peut fragmenter la collecte des informations et ainsi faire rentrer plusieurs exemples dans une requête

NER AVEC GPT3.5 - Few-Shot Learning

One-shot Learning

Acte
exemple
Complet

Labels
exemple
Complets

Acte
question
Complet

Labels
réponse

Few-shot Learning

Exemple
Paragraphe 1

Labels §1

Question
Paragraphe 1

Réponse §1

Exemple
Paragraphe 2

Labels §2

Question
Paragraphe 2

Réponse §2

Exemple
Paragraphe 3

Labels §3

Question
Paragraphe 3

Réponse §3

Exemple
Paragraphe 4

Labels §4

Question
Paragraphe 4

Réponse §4

Exemple
Paragraphe 5

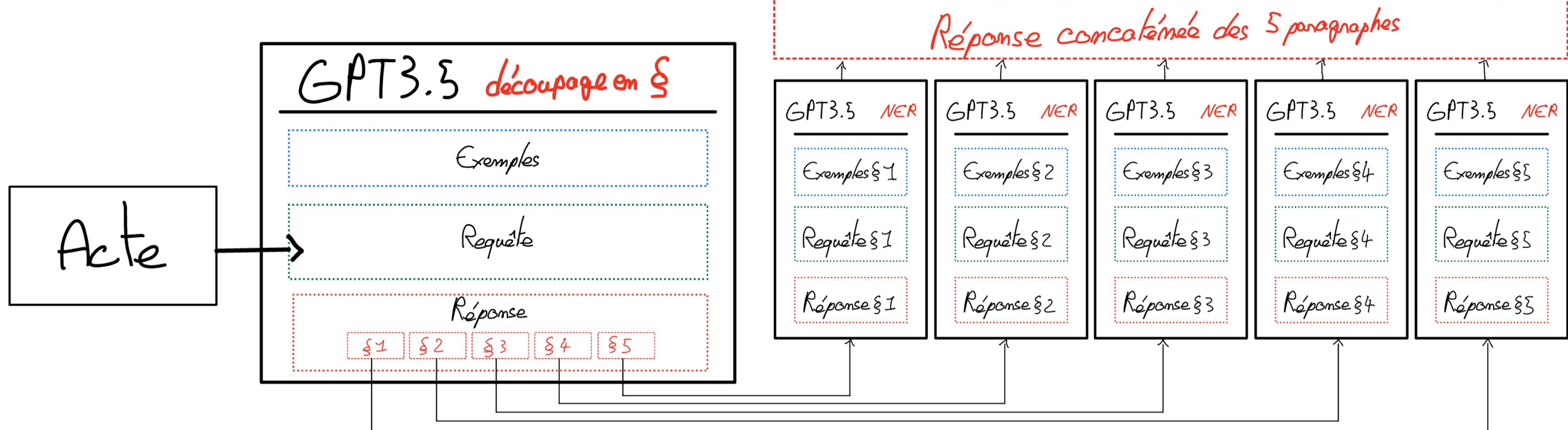
Labels §5

Question
Paragraphe 5

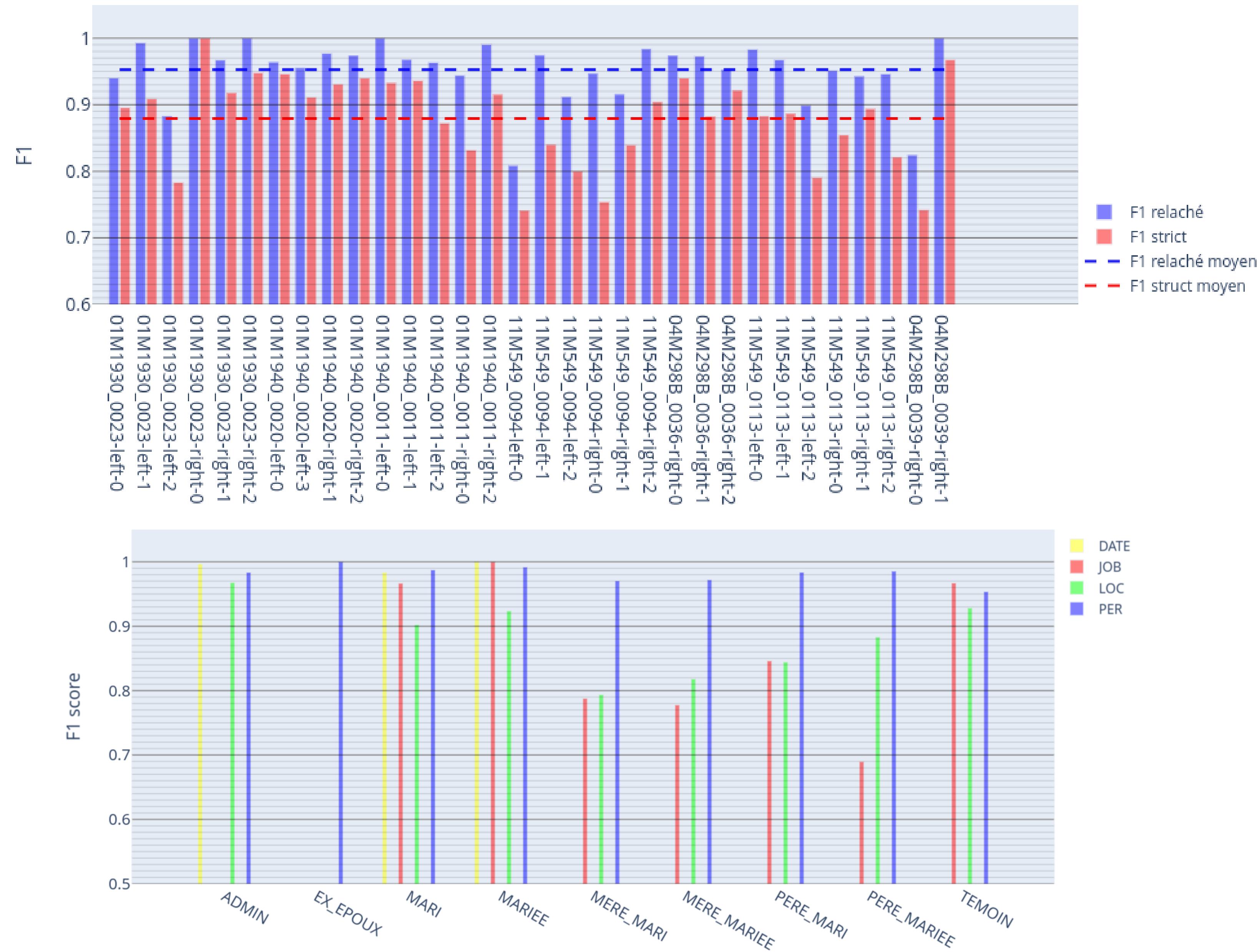
Réponse §5

NER AVEC GPT3.5 - Few-Shot Learning

⇒ On peut utiliser une autre instance de GPT3.5 pour découper les actes.



NER AVEC GPT3.5 - Few-Shot Learning



NER AVEC GPT3.5 - Correction des réponses

Belle amélioration des performances générales, mais il semble y avoir maintenant une sorte de suradaptation syntaxique : le modèle colle trop à la structure des exemples donnés

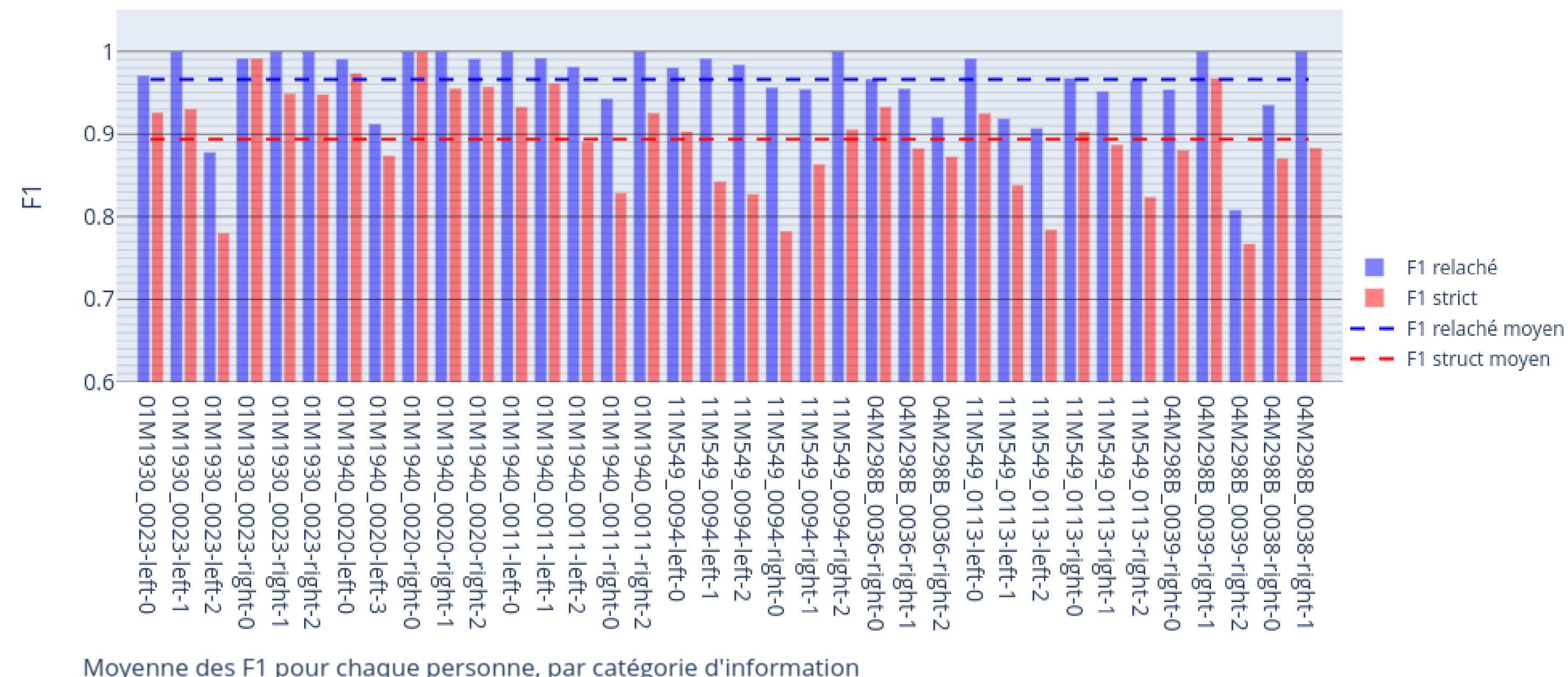
→ On peut corriger ces erreurs automatiquement après avoir reçu la réponse de l'API :

“Hernest, décédé” / “Hernest, charcutier”
“Né à Orly (Seine)” / “Né à Barcelone (Espagne)”
“Jeanne, sa veuve” / “Jeanne, horlogère”

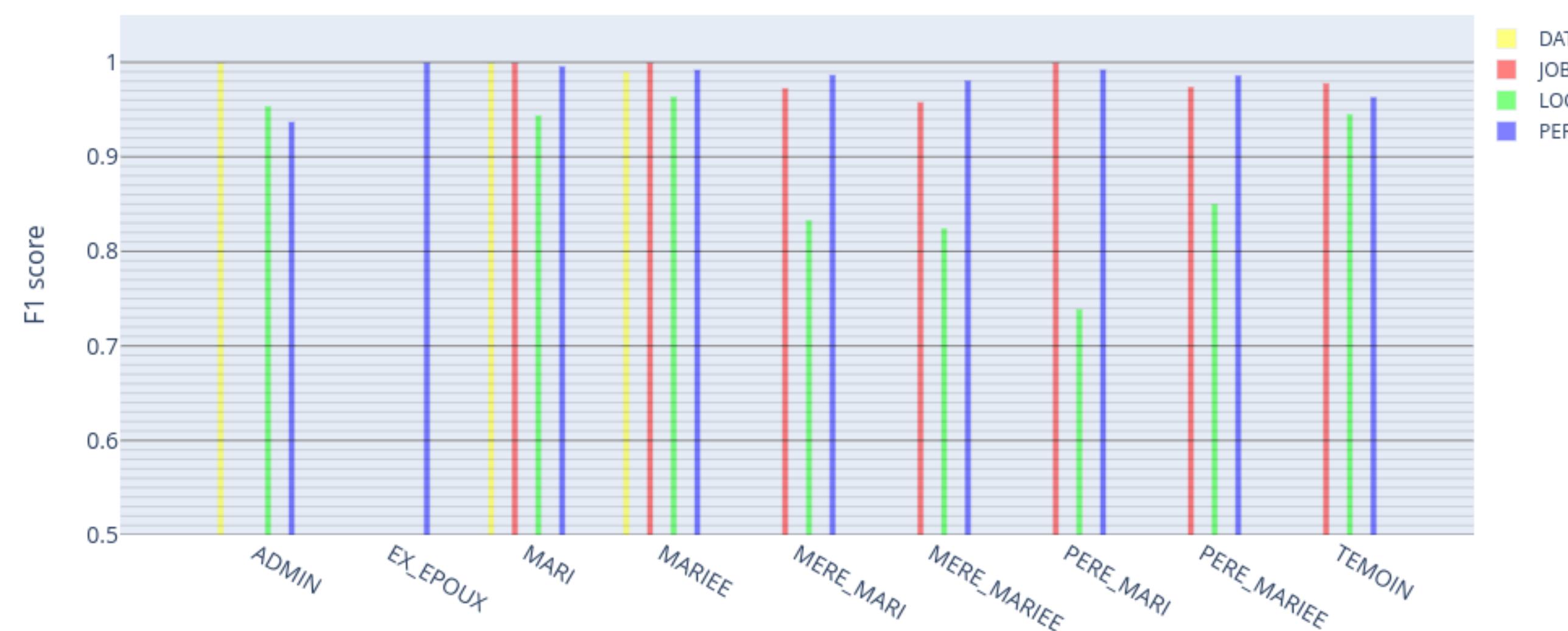
Label	Réponse API	Correction
Minute-mariage	Non mentionné, None, Aucune...	—
Profession-X	Disparu, décédé	Sans profession
Profession-mère-X	Sa veuve, X	X
Département Pays	Espagne —	— Espagne
Ville Pays	Alcira (Espagne) —	Alcira Espagne

NER AVEC GPT3.5 - Correction des réponses

Scores F1 pour chaque archive



Moyenne des F1 pour chaque personne, par catégorie d'information



Conclusion

L'utilisation de LLM permet de palier à certains aspects problématiques en NER :

- Manque de données annotées
- Grand nombre d'entités à extraire
- Temps et ressources nécessaires conséquents

⇒ Un coût très réduit et de bonnes performances peuvent représenter une alternative préliminaire à une annotation manuelle.

Travail à suivre dans la continuité du stage