

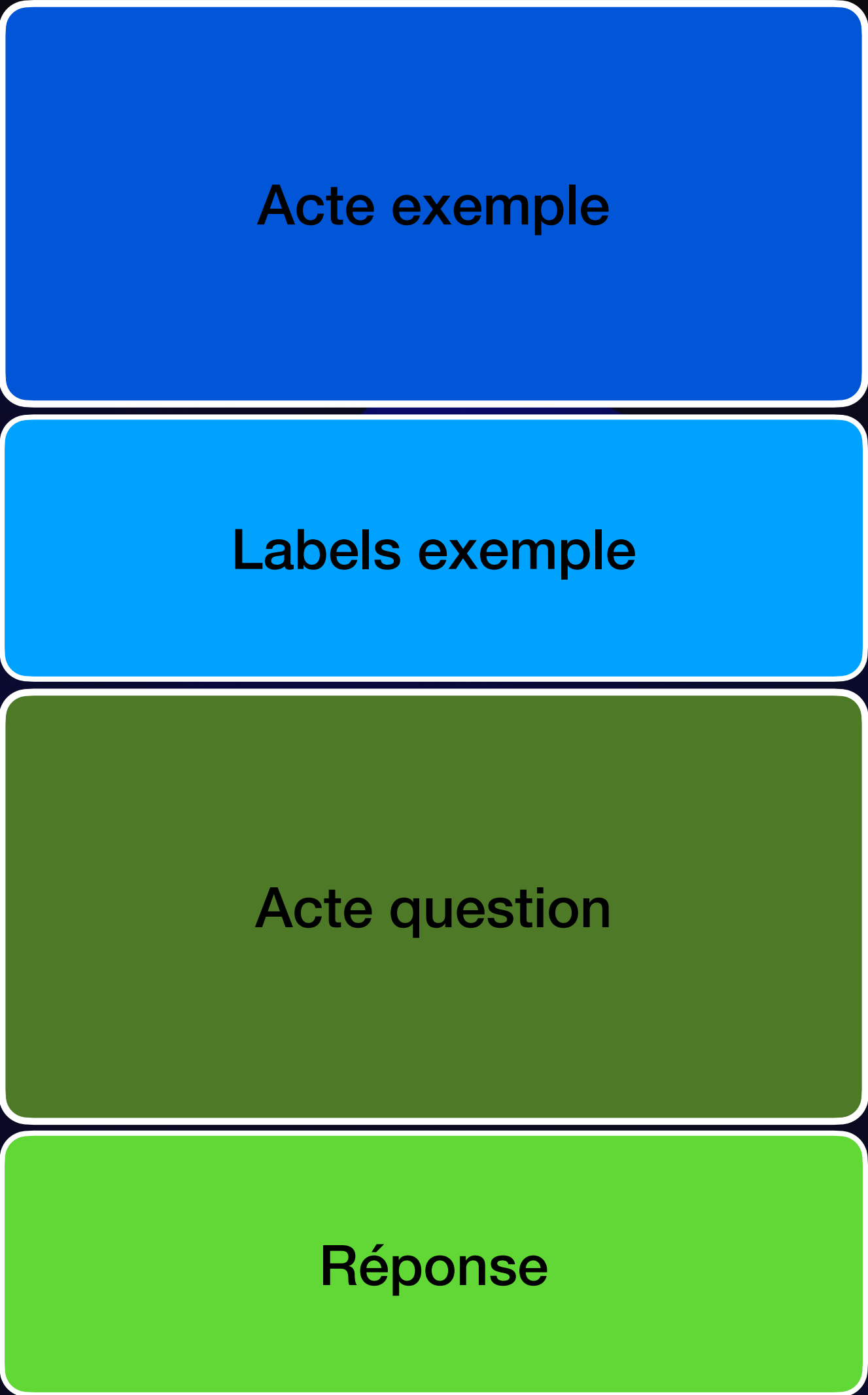
Réunion 09/05/2023

NER avec GPT3.5

Valentin BONSI

NER avec GPT3.5

One-shot learning



Few-shot learning



NER avec GPT3.5

Simplification des Labels par mot-clés :

- pas de caractères spéciaux
- pas de prépositions
- Cohérence des dénominations

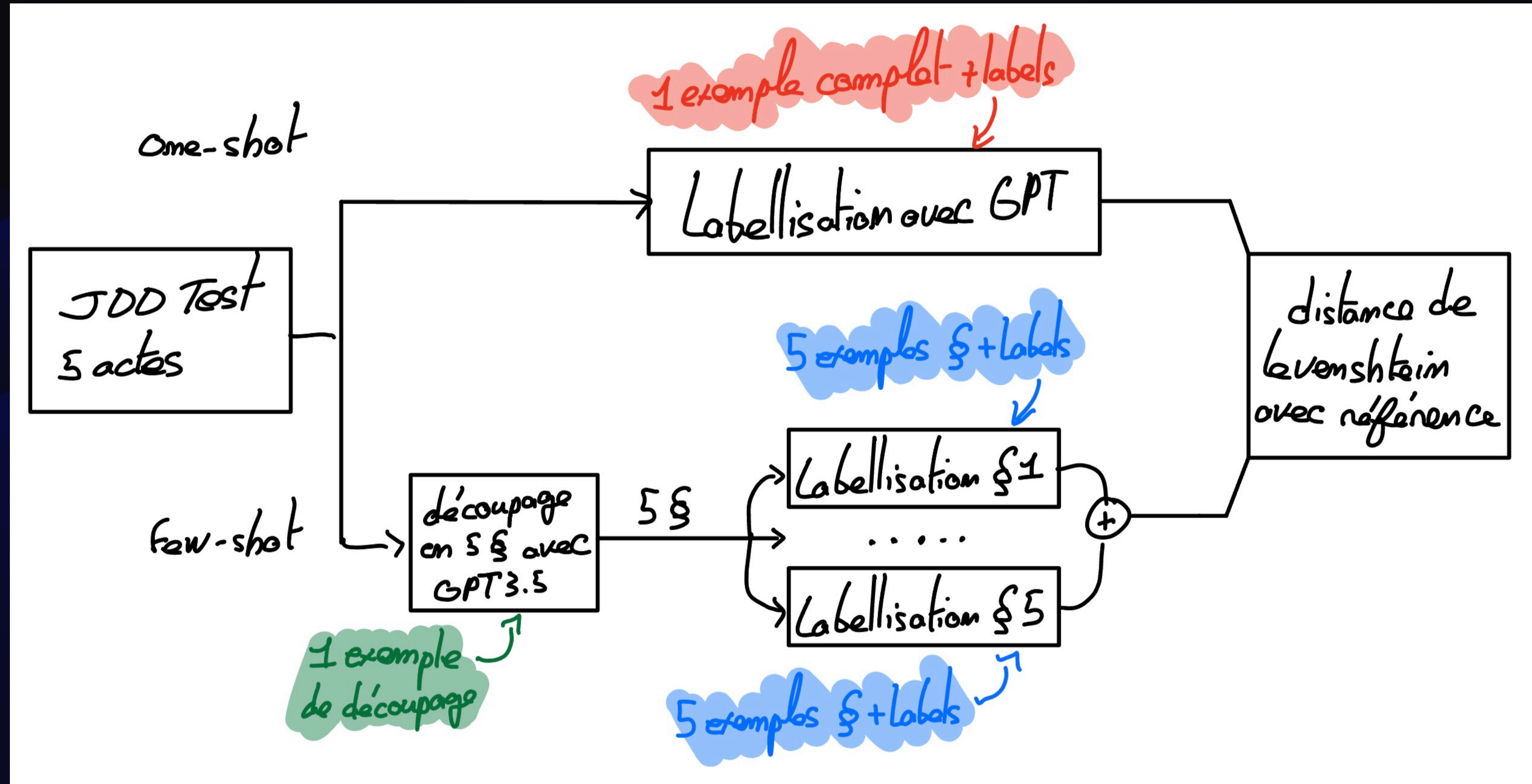
Référence	Renommage
Métier -de-la-mariée Profession -du-mari	Profession -mariee Profession -mari
Nom-de-rue-de-résidence-de-la-mère-de-la-mariée	Nom-rue-residence-mere-mariee
Prénom-de-l'adjoint-au-maire	Prenom-adjoint-maire

NER avec GPT3.5

Plus d'exemples et réponse à fournir moins longue =

- 1. Meilleur apprentissage**
- 2. Moins d'erreurs sur formulation complexes**
- 3. Moins d'erreur sur les clés des labels**
- 4. Moins d'hallucinations**
- 5. Pas besoin de consignes**

NER avec GPT3.5



NER avec GPT3.5

	One-shot Long labels	One-shot Compact labels	Few-shot Long labels	Few-shot Compact labels
Acte 1 (1940)	Error	Levenshtein : 19 Nb erreur : 2	Levenshtein : 54 Nb erreur : 4	Levenshtein : 4 Nb erreur : 0
Acte 2 (1940) Résidents Étrangers	Error	Levenshtein : 40 Nb erreur : 4	Levenshtein : 45 Nb erreur : 5	Levenshtein : 14 Nb erreur : 1
Acte 3 (1935)	Error	Levenshtein : 87 Nb erreur : 7	Levenshtein : 18 Nb erreur : 2	Levenshtein : 0 Nb erreur : 0
Acte 4 (1935) Bien rempli	Error	Levenshtein : 131 Nb erreur : 8	Levenshtein : 1 Nb erreur : 0	Levenshtein : 1 Nb erreur : 0
Acte 5 (1935) Plusieurs morts	Error	Levenshtein : 77 Nb erreur : 7	Levenshtein : 34 Nb erreur : 3	Levenshtein : 8 Nb erreur : 1
Acte 6 (1930)	Error	Levenshtein : 122 Nb erreur : 5	Levenshtein : 46 Nb erreur : 4	Levenshtein : 3 Nb erreur : 0

NER avec GPT3.5

Remarques :

One-shot Learning :

Variabilité des résultats entre les 3 runs, même en jouant avec la température
Erreurs principales : Non détection ou mauvaise association -> correction difficile

Few-shot Learning :

Très faible variabilité au cours des 3 runs avec une température à 0.7
Erreurs principales : Inversion département/pays -> correction facile, on connaît la liste des pays

Exemples	GPT	Référence
Departement_naissance_mari	Allemagne	/
Pays_naissance_mari	/	Allemagne
Departement_residence_pere_mari	Allemagne	/
Departement_residence_mere_mari	Allemagne	/
Pays_residence_pere_mari	/	Allemagne
Pays_residence_mere_mari	/	Allemagne

NER avec GPT3.5 - Correction des erreurs

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 - Dataset Étendu

Nombre d'actes : 22 actes (13x 1940, 4x 1935, 5x 1930)

Nombre de labels :

- §1 : 5 labels
- §2 : 37 labels
- §3 : 37 labels
- §4 : aucun labels (ignoré)
- §5 : 21 labels

-> TOTAL : 100 labels

NER avec GPT3.5 - Dataset Étendu

Méthode	Nb erreur max	Nb erreur total	Taux d'erreur pour 2200 labels :
One-shot	13	127	5.77%
Few-shot	7	35	1.5%
Few-shot-corrigé	5	15	0.68%

A suivre :

- Tester sur actes structures différentes (avant 1920, avant 1900)
- Créer un jeu d'exemple **de synthèse** cernant au mieux les spécificités des différents actes
- - Evaluer les métriques F1 en regroupant les labels par catégories