



HAL
open science

Évaluation des modèles génératifs dans un contexte few-shot ciblant la reconnaissance des entités nommées

Mustapha Benbarka

► **To cite this version:**

Mustapha Benbarka. Évaluation des modèles génératifs dans un contexte few-shot ciblant la reconnaissance des entités nommées. Atelier sur l'évaluation des modèles génératifs (LLM) et challenge d'extraction d'information few-shot, Institut des sciences informatiques et de leurs interactions - CNRS Sciences informatiques [INS2I-CNRS], Jul 2024, Toulouse, France. hal-04678386

HAL Id: hal-04678386

<https://hal.science/hal-04678386v1>

Submitted on 27 Aug 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Évaluation des modèles génératifs dans un contexte *few-shot* ciblant la reconnaissance des entités nommées

Mustapha Benbarka¹

(1) MINDIMENSIONS, Paris, France

mustapha.benbarka@mindimensions.fr

RÉSUMÉ

Ce papier représente une synthèse de l'ensemble des expériences réalisées dans le cadre du challenge *EvalLLM2024* visant à évaluer certains LLMs dans un contexte *Few-Shot* pour l'extraction des entités nommées.

ABSTRACT

Evaluation of Large Language Models with Few-shot for Named Entity Recognition

This paper describes the experiments conducted during the challenge *EvalLLM2024* which targets the evaluation of Large Language Models in Few-shot context for Named Entity Recognition.

MOTS-CLÉS : Évaluation, Modèles Génératifs, Reconnaissance des Entités Nommées, Few-Shot.

KEYWORDS: Evaluation, LLMs, Named Entity Recognition, Few-Shot.

1 Contexte

La reconnaissance des entités nommées a atteint un niveau de maturité élevé. Cela est dû principalement aux dernières avancées dans le traitement automatique du langage (TAL). Si on prend par exemple le célèbre challenge lancé par la conférence *CoNLL* (Tjong Kim Sang & De Meulder, 2003) et qui prend comme objectif la reconnaissance de quatre entités : Personne (PER), Lieu (LOC), Organisation (ORG) et Divers (MISC). Et si on commence à suivre les contributions basées sur les jeux de données de ce challenge, on trouve une littérature abondante dominée par l'utilisation des réseaux de neurones profonds et les plongements lexicaux. L'ensemble de ceux travaux aboutissent à des scores très proches (entre 0.90 et 0.95 en terme de F_1)¹ ; montrant que cette problématique est un sujet bien exploré.

Néanmoins, bien qu'un système générant un taux d'erreur de 5%, peut être vu comme un système à l'état-de-l'art, les entités impliquées représentent moins de complexité si on les compare avec d'autres entités d'intérêt relevant d'un corpus métier ou manifestant un niveau élevé d'ambiguïté. En outre, ces scores ont été obtenus sur des corpus anglais ; ce qui exige de mettre ces systèmes aux défis par rapport à d'autres langues comme le français ou l'arabe. De surcroit, l'émergence des modèles génératifs (Large Language Models LLMs) nous oblige aussi à étudier leurs performances dans des tâches dominées par l'utilisation d'une modélisation discriminative.

Ce papier représente une synthèse de l'ensemble des expériences réalisées dans le cadre du challenge

1. https://nlpprogress.com/english/named_entity_recognition.html, consulté le 10/06/2024

*EvalLLM2024*² visant à évaluer certains LLMs dans un contexte *Few-Shot* pour l'extraction des entités nommées.

A ce niveau, il est important de noter que conduire des plans d'expérimentation sur des LLMs est une tâche délicate vu la volumétrie qu'ils manifestent, le coût de calcul énorme qu'ils génèrent et la rareté des modèles entraînés sur des corpus français.

Afin de surmonter ces contraintes, on a opté pour trois types d'expériences :

- Utiliser un modèle génératif de taille réduite
- Compresser le LLM (i.e. *Quantization*)
- Utiliser un modèle adapté à la reconnaissance des entités nommées.

Ces types d'expériences vont nous permettre d'étudier la capacité d'un LLM de taille réduite à extraire des entités dans un contexte *Few-Shot*, de tester les performances d'un LLM après une compression et finalement de les comparer avec un modèle adapté à la reconnaissance des entités nommées.

2 Expériences

2.1 Modèle génératif de taille réduite

Le modèle choisit est *CroissantLLM* (Faysse *et al.*, 2024). Le choix de ce dernier est basé sur :

- sa taille réduite
- entraîné sur un corpus français traitant plusieurs domaines (e.g. légal, administratif, scientifique)
- sa capacité à surpasser certains modèles larges dans un contexte *Few-Shot*, au moins pour des tâches génératives comme la traduction.

Afin de tester sa capacité à comprendre des instructions et produire des réponses adéquates, le modèle a été interrogé par des *prompts* donnant en premier lieu le contexte de la tâche, ensuite le texte sur lequel l'extraction doit être faite.

Les résultats de ces expériences indiquent que le modèle a bien compris la tâche (voir *prompt* ci-dessous) mais quand il s'agit d'exécuter la tâche, le modèle manifeste un degré d'hallucination fort en produisant souvent des sorties vides.

Une étude approfondie doit être faite afin d'isoler l'origine de ces prédictions.

Ci-dessous un exemple d'un *prompt*³ ainsi que la réponse du modèle. L'instruction demande au modèle de mettre les entités nommées entre des balises XML. Si le modèle prédit ces *tags*, on peut détecter à la fois l'entité et sa position dans le texte :

Instructions : Vous recevez un message d'un utilisateur qui peut contenir des informations d'intérêts. On vous donne également une liste de ces entités d'intérêts qu'on appelle des entités nommées. Votre tâche consiste à détecter et à identifier toutes les entités nommées qui existent dans le texte fourni par l'utilisateur. La sortie doit avoir le même contenu que l'entrée. Seuls les mots correspondant aux entités nommées de la liste doivent être inclus dans des balises XML. Une balise XML représente une entité nommée décrite dans la liste ci-dessous. Par exemple, le nom d'une organisation doit être inclus dans les balises <ORGANIZATION></ORGANIZATION>. Veillez à ce que toutes les entités soient identifiables. Ne procédez pas à de fausses identifications.

Liste des entités nommées :

2. <https://evalllm2024.sciencesconf.org/resource/page/id/2>, consulté le 10/06/2024

3. Ce dernier est inspiré du blog [fine-tuning-large-language-models-for-named-entity-recognition](#), consulté le 10/06/2024

PERSON : Nom propre d'une personne ou d'un avatar

FUNCTION : Fonction ou titre de personnes

ORGANIZATION : Toute structure organisée ayant une reconnaissance juridique

MILITARYUNIT : Toute organisation relevant du militaire GROUP : Groupe d'individus qui ne sont pas des organisations ou des unités militaires

LOCATION : Nom d'un lieu géographique ou administratif

EQUIPMENT : Ensemble du matériel appartenant à une organisation ou à une personne

SITE : Lieu occupé pour un événement ou une construction

RESOURCE : Moyens et produits dont une organisation peut disposer

EVENT : Évènement comme les événements historiques TIME : date

ID : Toute information une personne dans un système d'information

Est ce que ces instructions sont claires ?

Généralisations :

Oui, ces instructions sont claires.

1. Commencez par lire le message de l'utilisateur.
2. Identifiez les entités nommées dans le message de l'utilisateur.
3. Identifiez les entités nommées qui existent dans le texte de l'utilisateur.
4. Détectez et identifiez toutes les entités nommées dans le texte de l'utilisateur.
5. Compilez la sortie XML en utilisant les balises XML fournies.

2.2 Un LLM compressé et *fine-tuné*

Pour des contraintes liées aux ressources de calcul et afin de tester plusieurs configurations, le choix de compresser le modèle a été fait.

Le LLM en question est *Claire-Mistral-7B-0.1*⁴ qui est un variant adapté à partir de *Mistral-7B-v0.1* (Jiang *et al.*, 2023).

Les configurations principales testées sont :

2.2.1 Un *fine-tuning* basé sur des *prompts* similaires à ceux testés sur *CroissantLLM*

Les sorties fournies par *Claire-Mistral-7B-0.1* indiquent que ce dernier rencontre des difficultés à mettre les entités nommées entre des balises XML.

2.2.2 Un *fine-tuning* basé sur des *prompts* demandant au modèle de produire des sorties en format JSON (*"Entité nommée" : "Type de l'entité"*)

Cette configuration indique que le modèle tolère ce format et arrive parfois à détecter certaines entités nommées. Cependant les sorties sont très bruitées ce qui rend l'extraction de ces résultats

4. <https://huggingface.co/OpenLLM-France/Claire-Mistral-7B-0.1>

inexploitable. La génération obtenue durant ce test a été soumise en tant que RUN1 pour le challenge. La Figure 1 illustre un exemple de prédiction fourni par le modèle. En analysant cette sortie, il semble que le modèle se focalise sur la première partie du *prompt* et oublie que le texte cible se trouve après la séquence "Le Texte :". Néanmoins, ce résultat peut nous indiquer que le LLM interagit dans "le voisinage" de la tâche même avec une sortie bruitée. Ce qui nous incite à poursuivre cette piste pour le RUN2.

```

En se basant sur la liste des entités nommées suivante: ['PERSON', 'FUNCTION', 'ORGANIZATION', 'MILITARY_UNIT', 'GROUP', 'LOCATION', 'EQUIPMENT', 'SITE', 'RESOURCE', 'EVENT', 'TIME', 'ID']
Identifiez les entités nommées dans le texte suivant en produisant une réponse sous format JSON: [{'Entité nommée': 'Type de l'entité'}]. S'il n'y a pas d'entité dans le texte, mettez comme réponse cette liste vide : .

Le Texte:
Au Burkina Faso, une peine de vingt ans de prison pour deux djihadistes présumés...
La justice burkinabèe a condamné mardi à 20 ans de prison deux djihadistes burkinabés, auteurs d'une attaque en 2018 contre une école dans le nord du Burkina Faso, au cours d'un procès placé sous haute sécurité. Les deux inculpés, âgés de 38 et 29 ans, qui se présentent comme cultivateur et éleveur, étaient jugés devant la chambre correctionnelle du pôle judiciaire spécialisé dans la répression des actes de terrorisme pour association de malfaiteurs en lien avec une entreprise terroriste, détention illégale d'armes et de munitions de guerre, complicité de terrorisme et destruction volontaire de biens. Selon le parquet, les deux inculpés, membres combattants du groupe armé Ansaroul Islam, ont le 2 mai 2018, attaqué et incendié l'école primaire de Bafina, localité située dans la province du Sanmatenga, dans la région du centre-nord. Accompagnés de quatre autres personnes, ils avaient également incendié le domicile du directeur de l'école, avant d'emporter deux motos. À la barre, les deux accusés ont reconnu les faits, expliquant s'être rendu à Bafina pour attaquer le groupe d'autodéfense du village. Ils ont ensuite décidé de cibler l'école primaire et les enseignants parce que les enseignements qui y sont donnés sont contraires aux dispositions de la charia, prônée par le groupe islamiste Ansaroul Islam. Reconnus coupables, ces deux djihadistes ont été condamnés à une peine d'emprisonnement de 20 ans dont 15 ans de sûreté. Ils devront également payer des dommages et intérêts de près de 4 millions de francs CFA (environ 6 000 euros). Lundi, la chambre correctionnelle du pôle judiciaire spécialisé dans la répression des actes de terrorisme, a ouvert les premiers procès pour terrorisme au cours desquels dix dossiers vont être jugés jusqu'au 13 août. Très actif dans le nord du Burkina, Ansaroul Islam a été créé en 2016 par Ibrahim Malam, un prêcheur burkinabè originaire de la province du Soum, région du Sahel. Le groupe a revendiqué plusieurs attaques contre l'armée burkinabèe, dont la plus meurtrière avait tué douze soldats en décembre 2016, avant de s'affilier à Al-Qaïda au Maghreb islamique

Les entités nommées:
1. [{'Entité nommée': 'PERSON'}]
2. [{'Entité nommée': 'FUNCTION'}]
3. [{'Entité nommée': 'ORGANIZATION'}]
4. [{'Entité nommée': 'MILITARY_UNIT'}]
5. [{'Entité nommée': 'GROUP'}]
6. [{'Entité nommée': 'LOCATION'}]
7. [{'Entité nommée': 'EQUIPMENT'}]
8. [{'Entité nommée': 'SITE'}]
9. [{'Entité nommée': 'RESOURCE'}]
10. [{'Entité nommée': 'EVENT'}]
11. [{'Entité nommée': 'TIME'}]
12. [{'Entité nommée': 'ID'}]

La réponse sous format JSON:
[{'Entité nommée': 'PERSON'}, {'Entité nommée': 'FUNCTION'}, {'Entité nommée': 'ORGANIZATION'}, {'Entité nommée': 'MILITARY_UNIT'}, {'Entité nommée': 'GROUP'}, {'Entité nommée': 'LOCATION'}, {'Entité nommée':

```

FIGURE 1 – Exemple d’une sortie du modèle sur les données de test - RUN1

2.2.3 Un *fine-tuning* basé sur des *prompts* demandant au modèle de produire des sorties en format JSON (*"start"* : "indice du début de l'entité", *"end"* : "indice de la fin de l'entité", *"text"* : "Entité nommée", *"labels"* : "Type de l'entité nommée")

Avec ce format, le modèle produit des sorties respectant cette structure et indique peut-être qu'il a été déjà pre-entraîné sur des fichiers JSON car il prédit parfois même des labels d'entités nommées non existant dans le corpus d'apprentissage. Les sorties durant ce test ont été soumises comme RUN2.

Le *fine-tuning* a été effectué en utilisant la méthode PEFT (*Parameter-Efficient Fine-Tuning*) combinée avec LORA (*Low-Rank Adaptation*) (Hu *et al.*, 2021).

Figure 2 illustre l'amélioration de la prédiction du modèle en ajoutant ce format aux *prompts* et en ajoutant une phrase plus explicite (i.e. "Les entités nommées sous format JSON :") au lieu de "Les entités nommées :". Cela peut indiquer une sensibilité aux *prompts* utilisées par rapport aux sorties des modèles.

A ce niveau, il est important d'indiquer que pour des raisons de ressources de calcul, le traitement des

En se basant sur la liste des entités nommées suivante: ['PERSON', 'FUNCTION', 'ORGANIZATION', 'MILITARY_UNIT', 'GROUP', 'LOCATION', 'EQUIPMENT', 'SITE', 'RESOURCE', 'EVENT', 'TIME', 'ID']
Identifiez les entités nommées dans le texte suivant en produisant une réponse sous format JSON: [{'Entité nommée': 'Type de l'entité'}]. S'il n'y a pas d'entité dans le texte, mettez comme réponse cette liste vide : .

Le Texte:
Au Burkina Faso, une peine de vingt ans de prison pour deux djihadistes présumés...
La justice burkinabèe a condamné mardi à 20 ans de prison deux djihadistes burkinabés, auteurs d'une attaque en 2018 contre une école dans le nord du Burkina Faso, au cours d'un procès placé sous haute sécurité. Les deux inculpés, âgés de 38 et 29 ans, qui se présentent comme cultivateur et éleveur, étaient jugés devant la chambre correctionnelle du pôle judiciaire spécialisé dans la répression des actes de terrorisme pour association de malfaiteurs en lien avec une entreprise terroriste, détention illégale d'armes et de munitions de guerre, complicité de terrorisme et destruction volontaire de biens. Selon le parquet, les deux inculpés, membres combattants du groupe armé Ansaroul Islam, ont le 2 mai 2018, attaqué et incendié l'école primaire de Bafina, localité située dans la province du Sanmatenga, dans la région du centre-nord. Accompagnés de quatre autres personnes, ils avaient également incendié le domicile du directeur de l'école, avant d'emporter deux motos. À la barre, les deux accusés ont reconnu les faits, expliquant s'être rendu à Bafina pour attaquer le groupe d'autodéfense du village. Ils ont ensuite décidé de cibler l'école primaire et les enseignants parce que ces enseignements qui y sont donnés sont contraires aux dispositions de la charia, prônée par le groupe islamiste Ansaroul Islam. Reconnus coupables, ces deux djihadistes ont été condamnés à une peine d'emprisonnement de 20 ans dont 15 ans de sûreté. Ils devront également payer des dommages et intérêts de près de 4 millions de francs CFA (environ 6 000 euros). Lundi, la chambre correctionnelle du pôle judiciaire spécialisé dans la répression des actes de terrorisme, a ouvert les premiers procès pour terrorisme au cours desquels dix dossiers sont être jugés jusqu'au 13 août. Très actif dans le nord du Burkina, Ansaroul Islam a été créé en 2016 par Ibrahim Malam, un prêcheur burkinabèe originaire de la province du Soum, région du Sahel. Le groupe a revendiqué plusieurs attaques contre l'armée burkinabèe, dont la plus meurtrière avait tué douze soldats en décembre 2016, avant de s'affilier à Al-Qaïda au Maghreb islamique

Les entités nommées:

```
1. [{"Entité nommée": "PERSON"}]  
2. [{"Entité nommée": "FUNCTION"}]  
3. [{"Entité nommée": "ORGANIZATION"}]  
4. [{"Entité nommée": "MILITARY_UNIT"}]  
5. [{"Entité nommée": "GROUP"}]  
6. [{"Entité nommée": "LOCATION"}]  
7. [{"Entité nommée": "EQUIPMENT"}]  
8. [{"Entité nommée": "SITE"}]  
9. [{"Entité nommée": "RESOURCE"}]  
10. [{"Entité nommée": "EVENT"}]  
11. [{"Entité nommée": "TIME"}]  
12. [{"Entité nommée": "ID"}]
```

La réponse sous format JSON:
[{"Entité nommée": "PERSON"}, {"Entité nommée": "FUNCTION"}, {"Entité nommée": "ORGANIZATION"}, {"Entité nommée": "MILITARY_UNIT"}, {"Entité nommée": "GROUP"}, {"Entité nommée": "LOCATION"}, {"Entité nommée": "EQUIPMENT"}, {"Entité nommée": "SITE"}, {"Entité nommée": "RESOURCE"}, {"Entité nommée": "EVENT"}, {"Entité nommée": "TIME"}, {"Entité nommée": "ID"}]

FIGURE 2 – Exemple d’une sortie du modèle sur les données de test - RUN2

séquences très longues a été réduit ce qui a impacté la sortie générée par le modèle en produisant souvent des prédictions incomplètes.

2.3 Modèle adapté à la reconnaissance des entités nommées

A cette étape, il existe un large spectre de choix pour des modèles de taille réduite ayant une capacité supérieure pour la reconnaissance des entités nommées. Le choix de *GLiNER (Generalist Model for Named Entity Recognition)* (Zaratiya *et al.*, 2023) a été effectué vu son architecture basée sur un *Bidirectional Transformer Encoder* ce qui représente un compromis entre les modèles de base classiques et les LLMs.

Selon les auteurs, *GLiNER* est construit selon trois composantes :

- Un encodeur pre-entraîné (e.g. BERT)
- Un composant qui produit des représentations pour les *spans* ; typiquement des plongements lexicaux
- Un autre composant qui produit des plongements lexicaux pour encoder les entités nommées cibles

L'idée est d'avoir les représentations pour les spans et les entités nommées dans le même espace latent afin de réaliser une sorte de correspondance (*matching*) entre les deux.

Les résultats obtenus sur le corpus d'apprentissage montrent une supériorité du modèle à extraire les entités nommées en donnant simplement les labels des entités et le texte cible (figure 3).

De plus, l'implémentation des prédictions du modèle permet d'obtenir facilement les positions des

entités dans le texte (*start* et *end*). Cela peut nous permettre d'éviter à réaliser un post-traitement des sorties dans le cas d'un LLM.

	start	end	text	label	score
0	26	47	paramilitaires russes	MILITARY_UNIT	0.795809
1	65	75	Washington	LOCATION	0.738248
2	136	157	paramilitaires russes	MILITARY_UNIT	0.810669
3	187	210	Linda Thomas-Greenfield	PERSON	0.944666
4	255	268	Nations unies	ORGANIZATION	0.818654
5	390	408	mercenaires russes	MILITARY_UNIT	0.615437
6	488	494	Niamey	LOCATION	0.741032
7	522	536	Mohamed Bazoum	PERSON	0.876016
8	545	568	Linda Thomas-Greenfield	PERSON	0.841395
9	601	629	Conseil de sécurité de l'ONU	ORGANIZATION	0.650254
10	650	666	week-end dernier	TIME	0.835193
11	669	675	Bamako	LOCATION	0.810585
12	829	832	ONU	ORGANIZATION	0.526273
13	850	871	paramilitaires russes	MILITARY_UNIT	0.831335
14	1060	1066	Bamako	LOCATION	0.791721
15	1130	1135	Paris	LOCATION	0.715879
16	1184	1194	ressources	RESOURCE	0.745835
17	1223	1235	Centrafrique	LOCATION	0.531301
18	1278	1285	Kremlin	ORGANIZATION	0.510635
19	1336	1352	Vladimir Poutine	PERSON	0.908609

FIGURE 3 – Exemple de prédictions des entités nommées fourni par GLiNER

Ces résultats incitent à effectuer un *fine-tuning* afin d'ajuster la capacité du modèle à prédire des nouveaux labels. Cependant, cela est hors scope de ce challenge qui vise les performances des LLMs.

Références

- FAYSSE M., FERNANDES P., GUERREIRO N. M., LOISON A., ALVES D. M., CORRO C., BOIZARD N., ALVES J., REI R., MARTINS P. H., CASADEMUNT A. B., YVON F., MARTINS A. F. T., VIAUD G., HUDELLOT C. & COLOMBO P. (2024). Croissantllm : A truly bilingual french-english language model.
- HU E. J., SHEN Y., WALLIS P., ALLEN-ZHU Z., LI Y., WANG S. & CHEN W. (2021). Lora : Low-rank adaptation of large language models. *CoRR*, **abs/2106.09685**.
- JIANG A. Q., SABLAYROLLES A., MENSCH A., BAMFORD C., CHAPLOT D. S., DE LAS CASAS D., BRESSAND F., LENGYEL G., LAMPLE G., SAULNIER L., LAVAUD L. R., LACHAUX M.-A., STOCK P., SCAO T. L., LAVRIL T., WANG T., LACROIX T. & SAYED W. E. (2023). Mistral 7b.
- TJONG KIM SANG E. F. & DE MEULDER F. (2003). Introduction to the CoNLL-2003 shared task : Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, p. 142–147.
- ZARATIANA U., TOMEH N., HOLAT P. & CHARNOIS T. (2023). Gliner : Generalist model for named entity recognition using bidirectional transformer.