CEA-List@EvalLLM2024 : prompter un très grand modèle de langue ou affiner un plus petit ?

Robin Armingaud, Arthur Peuvot, Romaric Besançon, Olivier Ferret, Sondes Souihi, Julien Tourille

CEA-List@EvalLLM2024 : prompter un très grand modèle de langue ou affiner un plus petit ?

Robin Armingaud Arthur Peuvot Romaric Besançon Olivier Ferret
Sondes Souihi Julien Tourille
Université Paris-Saclay, CEA, List, F-91120, Palaiseau, France
{robin.armingaud,arthur.peuvot,romaric.besancon,olivier.ferret,
sondes.souihi,julien.tourille}@cea.fr

RÉSUMÉ

Le challenge EvalLLM2024 cherche à évaluer les résultats d'approches few-shot pour l'extraction d'information en français. Dans notre contribution à ce challenge, nous testons deux approches : l'une exploite les données annotées disponibles dans le contexte d'interrogation d'un LLM (*in context learning*) tandis que l'autre affine un modèle générique de reconnaissance d'entités (GLiNER) en exploitant les données annotées. Nos expériences montrent que cette seconde approche obtient les meilleurs résultats, surtout lorsqu'elle est enrichie par une augmentation de données exploitant, d'une part, le guide d'annotation et, d'autre part, des LLM pour la génération d'exemples synthétiques.

ABSTRACT

CEA-List@EvalLLM2024: prompting a large language model or fine-tuning a smaller one?

The EvalLLM2024 challenge aims to evaluate the results of few-shot approaches to information extraction in French. Our contribution to this challenge tests two approaches: one exploits the available annotated data in the prompt of an LLM (*in context learning*) while the other fine-tunes a generic entity recognition model (GLiNER) by exploiting the annotated data. Our experiments show that this second approach obtains the best results, especially when enriched by a data augmentation step exploiting the annotation guide and LLMs for the generation of synthetic examples.

MOTS-CLÉS : Reconnaissance d'entités nommées, approches few-shot, modèles de langue.

KEYWORDS: Named entity recognition, few-shot approach, language model.

1 Introduction

Le développement des approches dites d'Intelligence Artificielle (IA) génératives fondées sur des grands modèles de langue, à la suite des travaux de Brown *et al.* (2020), s'est accompagné du développement de jeux d'évaluation de plus en plus grands pour évaluer leurs capacités, poursuivant en cela la tendance déjà esquissée pour l'évaluation des modèles de type BERT (Devlin *et al.*, 2019) avec des jeux de données tels que GLUE (Wang *et al.*, 2018). HELM (Bommasani *et al.*, 2023), MMLU (Hendrycks *et al.*, 2021) et BIG-bench (Srivastava *et al.*, 2023) sont les représentants les plus saillants de ces jeux d'évaluation à grande échelle. Outre leur tendance à être très centrés sur la langue anglaise, ces jeux de données possèdent une caractéristique plus surprenante : ils ne comprennent qu'assez peu de tâches d'extraction d'information et plus généralement, d'annotation de séquences,

pourtant historiquement au cœur des travaux en Traitement Automatique des Langues (TAL).

L'atelier EvalLLM2024 proposé dans le cadre de la conférence JEP TALN RECITAL 2024 organise une évaluation contribuant à remédier à cet état de fait en se focalisant sur les systèmes de reconnaissance d'entités nommées et d'identification d'événements en français dans un cadre dit *few-shot* où peu de données annotées sont disponibles pour caractériser chaque type d'entités ou d'événements.

De façon plus générale, la thématique du few-shot en extraction d'information et plus spécifiquement appliquée à la reconnaissance d'entités nommées ou d'événements a fait l'objet de nombreux travaux ces dernières années. Ces travaux ont bien souvent comme trait commun de se situer dans un cadre de méta-apprentissage. Un certain nombre d'entre eux, à l'instar de (Fritzler *et al.*, 2019) pour les entités ou (Cong *et al.*, 2021; Tuo *et al.*, 2023) pour les événements, s'inscrivent dans la lignée de (Vinyals *et al.*, 2016) : ils associent un apprentissage épisodique – la création artificielle d'un grand nombre d'épisodes d'apprentissage few-shot à partir d'un ensemble de données annotées suffisamment large – et un cadre d'apprentissage de métriques. Mais, à la suite du développement de la notion de base d'instructions, avec FLAN (Chung *et al.*, 2024) ou les efforts dans ce sens liés au modèle T0 (Sanh *et al.*, 2022), une autre approche de méta-apprentissage s'est imposée plus récemment, en particulier pour les modèles de langue à base de transformeurs (Vaswani *et al.*, 2017) de type décodeur seul, aboutissant aux modèles InstructGPT (Ouyang *et al.*, 2022) et ChatGPT (Achiam *et al.*, 2023). L'idée sous-jacente à cette approche est qu'un grand modèle de langue entraîné à partir d'un ensemble conséquent d'instructions spécifiant un large spectre de tâches sous forme d'énoncés en langage naturel peut acquérir la capacité à réaliser des tâches nouvelles spécifiées sous forme d'instructions.

Cette approche pose le problème de la définition de grands ensembles d'instructions. Alors que Chung *et al.* (2024) et Sanh *et al.* (2022) se sont appuyés sur la conversion sous forme d'instructions de jeux de données d'évaluation existants en TAL, ChatGPT a surtout exploité des instructions directement fournies par des humains. Afin de minimiser le coût de cette dernière option, Wang *et al.* (2023b) se sont tournés vers la distillation de modèle en utilisant ChatGPT pour produire des instructions destinées à entraîner un modèle de langue plus petit.

Dans le contexte de l'extraction d'information, et plus particulièrement de tâches comme la reconnaissance d'entités nommées ou d'événements, la conjugaison du méta-apprentissage par instruction et de la distillation de modèle a fait l'objet récemment de plusieurs travaux intéressants, aboutissant à la notion d'*Open Named Entity Recognition*. Plus spécifiquement, le méta-apprentissage est utilisé dans ce cadre non pas pour offrir la possibilité de réaliser de nouvelles tâches mais pour traiter de nouveaux types d'entités. L'idée est qu'entraîner un modèle de langue avec un grand nombre de types d'entités différents lui conférera la capacité à reconnaître de nouveaux types d'entités, soit sans exemple, à partir d'un intitulé ou un descriptif de type d'entité, soit avec quelques exemples, ce qui correspond au paradigme du *In-Context-Learning* (ICL).

Une première mise en œuvre de cette philosophie se retrouve dans InstructUIE (Wang *et al.*, 2023a), avec deux différences : le méta-apprentissage couvre plusieurs tâches d'extraction d'information et les instructions sont obtenues de façon comparable à FLAN, à partir de jeux de données annotés manuellement. La même approche se retrouve globalement dans GoLLIE (Sainz *et al.*, 2024), la différence se situant principalement au niveau du modèle de langue utilisé et de la forme des instructions : tandis qu'InstructUIE s'appuie sur un modèle instruit généraliste – FlanT5 (Chung *et al.*, 2024) – et des instructions de génération de conversations, GoLLIE exploite un modèle spécialisé pour la génération de code – Code-LLaMA (Roziere *et al.*, 2023) – et des instructions visant cette même tâche. UniversalNER (Zhou *et al.*, 2024) s'aligne plus directement sur la notion d'*Open Named Entity Recognition* en se concentrant sur la seule reconnaissance d'entités nommées et en constituant

un très grand ensemble d'instructions grâce à ChatGPT. Mais l'application de cette philosophie ne s'est pas arrêtée aux modèles de type décodeur seul. GLiNER (Zaratiana *et al.*, 2024) l'applique ainsi aux modèles de type BERT en réutilisant les données constituées par UniversalNER. Dans ce cas, le few-shot ne correspond plus à de l'ICL mais reprend la forme d'un affinage plus traditionnel. NuNER (Bogdanov *et al.*, 2024) s'inscrit dans la même optique, en proposant en particulier un ensemble de données d'entraînement différent de celui d'UniversalNER. Pour finir, MetaIE (Peng *et al.*, 2024) étend le champ d'application de cette façon de faire à un ensemble plus large de tâches d'extraction d'information.

Pour l'évaluation EvalLLM2024, qui se focalise sur un contexte few-shot, nous avons choisi de tester l'intérêt comparé d'une approche de type ICL, adossée à un modèle décodeur seul, et d'une approche de type affinage avec un modèle encodeur, les modèles ayant été préalablement pré-entraînés dans les deux cas avec de nombreux types d'entités différents.

2 Approches

2.1 Affinage d'un modèle encodeur : le modèle GLiNER

Le modèle GLiNER (Generalist Model for Named Entity Recognition using Bidirectional Transformer) est un modèle de reconnaissance d'entités nommées (REN) fondé sur un modèle de langue bidirectionnel (BERT, RoBERTa (Liu *et al.*, 2019), DeBERTa (He *et al.*, 2020)...) et entraîné sur un ensemble de données synthétiques générées par ChatGPT comportant une grande variété d'entités nommées (Zhou *et al.*, 2024). GLiNER traite le problème de reconnaissance d'entités nommées en étant entraîné à faire correspondre les plongements lexicaux des étiquettes des types d'entités avec les entités correspondantes, permettant d'obtenir de bonnes performances dans un contexte *zero-shot*. Grâce à son architecture, GLiNER est capable de classifier chaque sous-séquence de mots au sein d'une phrase et est donc facilement utilisable dans un contexte où les entités nommées peuvent être imbriquées, comme pour la tâche d'EvalLLM2024.

2.2 ICL et modèle décodeur : le modèle GoLLIE

GoLLIE est un grand modèle de langue (LLM) conçu pour suivre des directives d'annotation, sous forme de code Python, pour réaliser des tâches d'extraction d'information (EI). Il existe en trois versions : 7B, 13B et 34B, fondées respectivement sur les modèles Code-LLAMA 7B, 13B et 34B. Contrairement aux approches classiques d'EI, où un modèle est entraîné sur des données spécifiques, GoLLIE a été affiné pour suivre des directives d'annotation, ce qui améliore ses performances dans des contextes zero-shot et few-shot. Lors de l'entraînement, des techniques de régularisation telles que le masquage des noms de classes et l'abandon (dropout) de classes ont été utilisées, permettant ainsi au modèle de généraliser à des tâches ou domaines non vus pendant l'entraînement. Sainz *et al.* (2024) ont montré que dans un contexte zero-shot, ses performances pour des étiquettes non vues sont proches de celles pour des étiquettes déjà vues.

2.3 Post-traitements

Afin de suivre au plus près le guide d'annotation fourni et ainsi améliorer la précision des prédictions, en particulier ce qui concerne la délimitation des entités, un post-traitement spécifique est appliqué aux entités nommées prédites par les différents modèles en s'appuyant sur Spacy [1] pour déterminer les catégories grammaticales des mots ainsi que sur des listes prédéfinies de locutions prépositives de lieu, connecteurs logiques de temps et marqueurs de fréquence. Ce post-traitement effectue les ajustements suivants :

- Les composants des entités de type ORGANIZATION constitués de mots composés sont ajoutés individuellement comme entités de type ORGANIZATION.
- Si le mot précédant une entité de type TIME, LOCATION ou SITE est une préposition, il est intégré à l'entité.
- Si le mot précédant une entité de type UNKNOWN, FUNCTION, ORGANIZATION, MILITARY_UNIT, GROUP, SITE ou EQUIPMENT est un article et que le premier mot de l'entité est un nom commun, alors l'article est intégré à l'entité.
- Si le premier mot d'une entité de type PERSON est un article, il est retiré de l'entité.
- Si plusieurs prédictions de type TIME sont imbriquées, seule l'entité la plus longue, supposément la plus précise, est conservée. De même pour les types ID et PERSON.
- Si les mots précédant une entité de type LOCATION sont des locutions prépositives, ils sont intégrés à l'entité.
- Les connecteurs logiques et les marqueurs de fréquence sont détectés et ajoutés comme entités de type TIME.
- Les doublons prédits sont éliminés pour éviter les erreurs de sur-représentation des entités.

L'impact de ces post-traitements sur les prédictions est évalué en détail à la section 4.

3 Expérimentations

3.1 Cadre d'évaluation

En raison de la faible quantité de données d'entraînement fournies dans le cadre du challenge (5 documents) et afin d'obtenir une évaluation fiable sur l'ensemble des documents lors de la mise au point de nos modèles, nous nous sommes placés dans un cadre de validation croisée d'un contre tous (*leave-one-out*) pour toutes les expérimentations avec GLiNER et GoLLIE exploitant des données annotées. Cette méthode est appliquée ici à l'échelle du document : chaque document est utilisé tour à tour comme ensemble de test tandis que les quatre autres documents servent d'ensemble d'entraînement. Pour les expérimentations zero-shot avec GoLLIE, l'évaluation se fait sur les exemples de tous les documents.

3.2 GLiNER

Suite à des expérimentations préliminaires, nous avons observé que le modèle GLiNER affiche surtout de bonnes performances après une phase de réentraînement sur des données de la tâche cible. Ces

1. Modèle fr_core_news_sm : https://spacy.io/models/fr

performances sont rapportées dans le tableau 1.

Métrique	Précision	Rappel	F1-Score
Total micro	68,49	66,01	67,23
Total macro	55,00	49,54	51,52

TABLE 1 – Performances globales du modèle GLiNER en validation croisée d'un contre tous, avec les post-traitements décrits à la section 2.3.

3.2.1 Augmentation du corpus d'entraînement

Exemples du guide d'annotation. Pour élargir l'ensemble d'entraînement initial, en particulier pour les types d'entités les moins représentés, nous avons extrait des exemples supplémentaires du guide d'annotation selon la procédure décrite à l'annexe D.1. Le nombre d'exemples d'entraînement est ainsi passé de 1 à 6 pour le type *UNKNOWN* et de 4 à 7 pour le type *ID* mais implique une répétition de certaines entités, discutée à la section 3.2.3. Le tableau 2 illustre l'impact très positif de cet élargissement de l'ensemble d'entraînement.

Métrique	Précision	Rappel	F1-Score
Total micro	73,22	72,24	72,73
Total macro	64,82	61,78	62,32

TABLE 2 – Résultats globaux de GLiNER avec l'ajout du guide d'annotation.

Augmentation de données à l'aide d'un grand modèle de langue. À partir du corpus constitué des éléments d'entraînement du challenge et des exemples du guide d'annotation, nous proposons une méthode permettant de créer de nouveaux exemples d'entraînement synthétiques en faisant varier les entités de ces exemples initiaux. Plus précisément, pour chaque exemple synthétique à générer et pour chaque entité présente dans l'exemple d'origine, on effectue un tirage selon une loi de Bernouilli de probabilité p, variant de 0,1 à 0,9, avec un pas de 0,2. En cas de réussite du tirage, l'entité est remplacée par une entité générée par le LLM Mixtral-8x7B avec une température de 0,7, en contraignant l'inférence afin d'obtenir un fichier JSON valide grâce à vLLM (Kwon *et al.*, 2023). Chaque exemple est augmenté d'un facteur n égal à 5, 10, 15 ou 20, ce qui produit finalement, avec les différentes probabilités p, 20 ensembles de données augmentés différents. Le tableau 3 donne les résultats de l'évaluation de ces ensembles.

	$n=5$		$n=10$		$n=15$		$n=20$		Moy.	
	Micro F1	Macro F1	Micro F1	Macro F1	Micro F1	Macro F1	Micro F1	Macro F1	Micro F1	Macro F1
$p=0,1$	72,18	63,07	74,39	65,76	74,02	65,47	74,23	72,92	73,71	65,30
$p=0,3$	74,20	65,33	72,67	65,72	74,96	74,01	74,96	66,89	74,20	68,17
$p=0,5$	73,20	66,07	72,14	72,12	73,20	64,39	73,78	63,39	73,08	66,49
$p=0,7$	73,38	65,95	70,91	60,65	72,76	61,99	73,63	73,90	72,67	65,62
$p=0,9$	72,70	65,41	72,16	63,16	71,06	61,03	72,92	62,53	72,21	63,03
Moy.	73,13	65,17	72,45	65,48	73,20	65,38	73,90	67,93	73,17	65,99

TABLE 3 – Résultats de GLiNER sur chaque ensemble de données synthétiques.

3.2.2 Combinaison de modèles

La combinaison de modèles est une méthode utilisée classiquement pour améliorer la performance des systèmes. Dans le cas de modèles neuronaux tels que GLiNER, il est possible de produire plusieurs modèles différents entraînés à partir des mêmes données en faisant varier la graine du générateur aléatoire utilisé. Pour combiner les prédictions issues de plusieurs de ces modèles GLiNER, nous avons envisagé les quatre méthodes classiques suivantes :

- Vote à la majorité absolue : si plus de la moitié des modèles prédisent une certaine classe, cette classe est choisie comme résultat final.
- Vote à la majorité relative : cette méthode est une variante de la majorité absolue où la classe avec le plus grand nombre de votes est choisie comme résultat final.
- Intersection : cette méthode ne conserve que les entités identifiées par l'ensemble des modèles.
- Union : cette méthode agrège toutes les prédictions faites par au moins un modèle de l'ensemble.

Les méthodes de vote à la majorité et l'intersection réduisent le nombre de faux positifs et permettent une meilleure précision globale en contrepartie d'un moins bon rappel tandis que l'union de toutes les prédictions augmente le rappel contre une baisse de la précision.

Nos expérimentations ont montré que la meilleure solution est une combinaison de ces méthodes. En pratique, une méthode différente est appliquée pour chaque type d'entités en fonction des performances des modèles sur l'ensemble d'entraînement : si les modèles ont en moyenne un rappel plus important sur un type d'entités, un vote à la majorité relative est appliqué. En revanche, si les modèles ont en moyenne une précision plus importante, l'union des prédictions est utilisée.

3.2.3 Discussions

Modèles sélectionnés. Les modèles utilisés pour le challenge EvalLLM2024 sont :

- Un modèle GLiNER entraîné à partir de toutes les données d'entraînement et de toutes les données d'augmentation (paramètres $n = 15$ et $p = 0,3$ pour l'augmentation par LLM).
- Un ensemble de 5 modèles GLiNER entraînés sur le même ensemble de données avec des graines (*seed*) différentes et assemblés selon la méthode décrite à la section 3.2.2.

Contamination des données avec le guide d'annotation. L'introduction des exemples inclus dans le guide d'annotation implique un certain degré de contamination car certaines entités présentes dans les documents sont reprises dans le guide. En étudiant de façon plus détaillée la proportion d'entités communes entre les documents annotés et le guide d'annotation, nous observons que 22,6 % des entités du guide d'annotation sont présentes dans deux documents ou plus de l'ensemble d'entraînement et que la contamination est particulièrement importante pour les documents 75 et 74 dont 35 % et 38 % des entités sont contenues dans le guide d'annotation.

Du fait de cette contamination, les résultats obtenus avec l'ajout des exemples du guide d'annotation ne sont pas comparables avec ceux du modèle entraîné uniquement sur les données d'entraînement fournies par le challenge. Par ailleurs, certaines entités sont présentes avec une fréquence plus importante dans l'entraînement des modèles finaux sur l'ensemble des données.

Méthodes d'augmentation. La méthode d'augmentation employée s'apparente à l'augmentation au niveau des entités (*entity-level*) présentée dans (Ye *et al.*, 2024) et se distingue des méthodes classiques d'augmentation présentées par Dai & Adel (2020). Ces dernières nous semblent peu pertinentes pour le challenge en raison de la difficulté et des règles très spécifiques de la tâche d'annotation du challenge. L'augmentation au niveau des entités à l'aide d'un LLM présente de plus de meilleures performances d'après (Ye *et al.*, 2024) tout en préservant la lisibilité et la structure grammaticale des phrases générées. Néanmoins, d'autres méthodes d'augmentation au niveau de la phrase complète sont également possibles et pourraient être des pistes d'amélioration intéressantes.

3.3 GoLLIE

Différentes expérimentations ont été menées pour évaluer les performances de GoLLIE sur le benchmark EvalLLM2024. Les directives d'annotation utilisées pour chaque classe se fondent sur le guide d'annotation fourni (cf. annexe F pour leur définition). Toutes les expériences ont été réalisées à l'aide de deux GPUs NVIDIA A100 et de la bibliothèque Python vLLM. Les hyperparamètres utilisés sont détaillés à l'annexe B. Le tableau 4 détaille les résultats des premières expérimentations réalisées dans un contexte zero-shot et montre que parmi les trois versions du modèle, GoLLIE-13B obtient les meilleurs résultats. Par la suite, c'est donc ce modèle 13B que nous utilisons.

	Micro F1-Score	Macro F1-Score
GoLLIE-7B	22,06	15,88
GoLLIE-13B	**29,11**	**18,97**
GoLLIE-34B	24,75	15,00

TABLE 4 – Performance des modèles GoLLIE en zero-shot sur tout le corpus d'entraînement, avec les post-traitements décrits à la section 2.3.

3.3.1 Zero-shot : stratégies de prompts

Lors des évaluations précédentes, chaque prompt contenait les directives d'annotation de toutes les classes ainsi que l'entièreté du texte à analyser. Pour améliorer les résultats, différentes stratégies de prompts ont été explorées. Toutes les configurations des éléments suivants ont été évaluées :

Concernant le texte :

- Le texte entier dans un seul prompt.
- Un prompt par phrase. Spacy a été utilisé pour découper les textes en phrases.

Concernant les directives d'annotation :

- Un prompt contenant les directives d'annotation de toutes les classes.
- Un prompt par groupe de directives d'annotation. Les classes ayant des définitions proches ont été regroupées pour permettre une meilleure distinction de leur sens par le LLM. Par exemple, les classes GROUP, ORGANIZATION, MILITARY_UNIT ont été traitées ensemble. Les groupes formés sont détaillés à l'annexe F.
- Un prompt par directive d'annotation pour interroger le LLM sur une seule classe à la fois.

Comme le montre le tableau 5, les meilleurs résultats sont obtenus en utilisant un prompt par phrase et par directive d'annotation. En comparaison avec les premières expérimentations, cette stratégie de prompts permet de gagner 4 points de micro F1-score et 10 points de macro F1-score. Cette configuration a été utilisée pour toutes les expérimentations en few-shot décrites ci-après. Il est important de noter que du fait du nombre de prompts qu'elle produit, c'est la configuration requérant le plus long temps d'exécution et donc, la plus énergivore.

	Micro F1-Score	Macro F1-Score
1 prompt par texte et avec toutes les classes	29,11	18,97
1 prompt par texte et par groupe de classes	31,98	23,19
1 prompt par texte et par classe	24,90	24,22
1 prompt par phrase et avec toutes les classes	34,80	28,50
1 prompt par phrase et par groupe de classes	35,84	25,73
1 prompt par phrase et par classe	**35,89**	**28,58**

TABLE 5 – Performance des différentes stratégies de prompts pour GoLLIE-13B en zero-shot sur tout le corpus d'entraînement, avec les post-traitements décrits à la section 2.3.

3.3.2 Few-shot : stratégies de choix d'exemples

GoLLIE-13B a également été évalué dans des contextes few-shot. Dans cette configuration, les prompts contiennent toujours des exemples de la classe dont les directives d'annotation sont décrites dans le prompt. Les tests ont été réalisés avec 5 exemples. Afin de déterminer si les exemples choisis ont un impact sur les prédictions, deux stratégies de choix d'exemples ont été explorées :

— Choix aléatoire des exemples : différentes graines ont été utilisées.

— Choix fondé sur un score de similarité : Guo *et al.* (2023) ont montré que sélectionner des exemples similaires au texte à analyser dans un contexte de In-context-learning peut améliorer les performances. Un modèle est utilisé pour générer les représentations vectorielles des phrases puis les scores de similarité sont calculés grâce à la mesure cosinus. Pour chaque phrase, les 5 exemples ayant la plus grande similarité sont sélectionnés. Des expérimentations avec les modèles Solon-embeddings-large-0.1 [2] et sentence-camembert-large [3] ont été menées.

Les classes ID et UNKNOWN n'ayant pas un nombre suffisant d'exemples dans le corpus d'entraînement, les prompts contenaient respectivement 4 et 1 exemples.

Le tableau 6 détaille les résultats de ces expérimentations. Lorsque les exemples sont sélectionnés de manière aléatoire, les résultats varient selon la graine choisie mais dans la majorité des cas, ils surpassent les performances en zero-shot. En moyenne, cette approche apporte une amélioration d'environ 3 et 6 points respectivement en micro et macro F1-Score comparé à la meilleure approche zero-shot. Les performances en choisissant des exemples similaires avec les modèles sentence-camembert-large ou Solon-embeddings-large-0.1 sont inférieures à celles avec des exemples aléatoires. De plus, les calculs de similarité induisent un temps d'exécution significativement plus long, entraînant des coûts énergétiques plus élevés. Les coûts énergétiques et environnementaux de l'ensemble des expérimentations avec GoLLIE sont détaillés dans le tableau 12 de l'annexe C.

2. https://huggingface.co/OrdalieTech/Solon-embeddings-large-0.1
3. https://huggingface.co/Lajavaness/sentence-camembert-large

		Micro F1-Score	Macro F1-Score
Exemples aléatoires Graine :	0	40,63	37,20
	12	33,49	31,62
	76	38,25	34,90
	233	40,62	36,03
	407	**41,08**	**37,53**
	859	39,18	34,38
	Moyennes	38,88	35,28
Exemples similaires Modèle :	sentence-camembert-large	35,64	33,76
	Solon-embeddings-large-0.1	31,90	28,92

TABLE 6 – Performance des différentes stratégies de choix d'exemples en few-shot, avec les post-traitements décrits à la section 2.3. L'évaluation se fait en validation croisée d'un contre tous sur le corpus d'entraînement.

3.3.3 Discussion

Modèles sélectionnés. Le modèle sélectionné pour le challenge EvalLLM2024 est GoLLIE-13B dans un mode few-shot. Chaque prompt est composé d'une phrase, des directives d'annotation d'une seule classe ainsi que de 5 exemples associés à cette classe.

Directives d'annotation. La définition des directives d'annotation joue un rôle crucial dans la variation des performances. Il est essentiel d'adapter ces directives à chaque benchmark. De plus, le choix des exemples ajoutés peut exercer une influence significative sur les prédictions du modèle : ces exemples sélectionnés doivent s'aligner étroitement avec le guide d'annotation.

4 Évaluation de l'impact des post-traitements

Le tableau 7 montre l'impact des post-traitements, détaillés à la section 2.3, sur les prédictions des différents modèles. Sur les prédictions de GLiNER, les post-traitements ont un impact minime. En revanche, pour GoLLIE en zero et few-shot, les post-traitements permettent une réelle amélioration des performances, avec des gains de plus de 18 et 13 points en micro et macro F1-Score en zero-shot et 14 points en micro et macro F1-Score en few-shot.

Durant son entraînement, GLiNER a pu intégrer les particularités et les nuances des annotations attendues. Ainsi, les prédictions brutes de GLiNER sont déjà alignées avec le guide d'annotation, rendant les post-traitements superflus, voire contre-productifs. Au contraire, les post-traitements sont essentiels à GoLLIE pour améliorer ses performances, en ajustant les prédictions pour qu'elles correspondent mieux aux exigences du guide. Par exemple, GoLLIE ne va pas toujours ajouter l'article précédent une entité de type UNKNOWN, FUNCTION, ORGANIZATION, MILITARY_UNIT, GROUP, SITE ou EQUIPMENT et ce, même si des exemples lui sont fournis dans le prompt. Une des limitations de GoLLIE est donc sa capacité limitée à s'adapter à un guide d'annotation très spécifique.

Méthode	Avant post-traitements		Après post-traitements	
	Micro F1	Macro F1	Micro F1	Macro F1
GLiNER	74,22	71,89	73,83	72,12
Ensemble de 5 GLiNER	**75,69**	**74,14**	**75,26**	**74,30**
GoLLIE-13B zero-shot	17,65	15,64	35,89	28,58
GoLLIE-13B few-shot - exemples aléatoires	25,93	22,92	40,63	37,20

TABLE 7 – Performances avant et après les post-traitements des modèles utilisés pour la soumission. L'évaluation se fait en validation croisée d'un contre tous sur le corpus d'entraînement.

5 Résultats de l'évaluation sur les données de test

Le tableau 8 donne les résultats de nos trois soumissions sur les données de test, notre meilleure soumission ayant une macro-F1 de 59,72 % tandis que la médiane des scores de macro-F1 se situe aux alentours de 28 %.

Le premier constat est que ces résultats ont globalement le même profil que nos évaluations en validation croisée sur le corpus d'entraînement : GoLLIE est nettement surclassé par les modèles GLiNER et la combinaison de 5 modèles GLiNER obtient les meilleurs résultats. Les différences s'observent surtout entre les scores macro et micro. Ainsi, la différence entre micro et macro-F1 est limitée pour GoLLIE dans le tableau 7 alors qu'elle est plus importante au niveau du tableau 8. Pour un modèle GLiNER seul, le point de comparaison se trouve plutôt au niveau du tableau 1 compte tenu de la tendance à la surévaluation des résultats introduite par les exemples du guide d'annotation. Tandis que le score de macro-F1 sur le test dépasse celui du tableau 1, la tendance est inverse pour la micro-F1. L'annexe A fournit quelques résultats plus détaillés mais en l'absence d'information quant au nombre de mentions de chaque type d'entités dans le corpus de test, il est difficile d'interpréter ce constat et de déterminer en particulier s'il résulte d'une différence de distribution des données entre le corpus d'entraînement et le corpus de test ou s'il trouve son origine dans l'évaluation en validation croisée d'un contre tous.

Modèles	Macro-P	Macro-R	Macro-F1	Micro-P	Micro-R	Micro-F1
GLiNER ensemble	**62,36**	**57,65**	**59,72**	**66,91**	**63,72**	**65,28**
GLiNER	58,14	54,11	55,81	65,88	62,86	64,34
GoLLIE	43,41	30,45	34,4	44,27	22,05	29,43

TABLE 8 – Résultats de nos trois soumissions sur les données de test.

6 Conclusion et perspectives

Dans cet article, nous avons présenté les méthodes sous-tendant nos trois soumissions à l'évaluation EvalLLM2024 portant sur la reconnaissance d'entités nommées dans un contexte few-shot. Nous avons en particulier mis en regard, pour l'exploitation des exemples fournis, l'utilisation de l'ICL par un LLM génératif dédié initialement à la génération de code couplée à de l'augmentation de données

et l'affinage d'un modèle de type encodeur. Dans les deux cas, les modèles avaient été préentraînés sur de grands corpus annotés en entités nommées. Nos évaluations et nos résultats sur le jeu de test ont montré la nette supériorité de la seconde option.

Les expérimentations menées montrant l'intérêt d'affiner les modèles avec le peu de données disponibles, une extension très directe de ce travail serait d'affiner un LLM génératif avec ces données. Ce processus est coûteux compte tenu de la taille des modèles utilisés mais l'utilisation de méthodes de quantification et d'adapteurs permet de réduire fortement ce coût. Par ailleurs, nos expérimentations ont également montré l'intérêt de l'utilisation de LLM génératifs pour la génération de données synthétiques, piste que nous souhaiterions explorer plus avant.

Remerciements

Ces travaux ont été réalisés grâce au supercalculateur Factory-IA, financé par le Conseil Régional d'Île-de-France. Ils ont également bénéficié du soutien des projets ARIEN et VANGUARD, financés par l'union européenne dans le cadre du programme de recherche et d'innovation Horizon 2020 (n° de convention de subvention 101121329 pour ARIEN et 101121282 pour VANGUARD), ainsi que du projet DataFIX, financé par le gouvernement dans le cadre du Programme France 2030, opéré par Bpifrance.

Références

ACHIAM J., ADLER S., AGARWAL S., AHMAD L., AKKAYA I., ALEMAN F. L., ALMEIDA D. *et al.* (2023). GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

BOGDANOV S., CONSTANTIN A., BERNARD T., CRABBÉ B. & BERNARD E. (2024). NuNER: Entity Recognition Encoder Pre-training via LLM-Annotated Data. *arXiv preprint arXiv:2402.15343*.

BOMMASANI R., LIANG P. & LEE T. (2023). Holistic evaluation of language models. *Annals of the New York Academy of Sciences*, **1525**(1), 140–146.

BROWN T., MANN B., RYDER N., SUBBIAH M., KAPLAN J. D., DHARIWAL P., NEELAKANTAN A. *et al.* (2020). Language Models are Few-Shot Learners. In H. LAROCHELLE, M. RANZATO, R. HADSELL, M. BALCAN & H. LIN, Éds., *Thirty-fourth Annual Conference on Neural Information Processing Systems (NeurIPS 2020)*, volume 33, p. 1877–1901, Virtual.

CHUNG H. W., HOU L., LONGPRE S., ZOPH B., TAY Y., FEDUS W., LI Y. *et al.* (2024). Scaling Instruction-Finetuned Language Models. *Journal of Machine Learning Research*, **25**(70), 1–53.

CONG X., CUI S., YU B., LIU T., YUBIN W. & WANG B. (2021). Few-Shot Event Detection with Prototypical Amortized Conditional Random Field. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, p. 28–40, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.findings-acl.3.

DAI X. & ADEL H. (2020). An analysis of simple data augmentation for named entity recognition. *arXiv preprint arXiv:2010.11683*.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for language understanding. In *Proceedings of the 2019 Conference*

of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), p. 4171–4186, Minneapolis, Minnesota : Association for Computational Linguistics. DOI : 10.18653/v1/N19-1423.

FRITZLER A., LOGACHEVA V. & KRETOV M. (2019). Few-shot classification in named entity recognition task. In *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing*, p. 993—-1000, Limassol, Cyprus : Association for Computing Machinery. DOI : 10.1145/3297280.3297378.

GUO Y., LI Z., JIN X., LIU Y., ZENG Y., LIU W., LI X. *et al.* (2023). Retrieval-augmented code generation for universal information extraction. *arXiv preprint arXiv:2311.02962*.

HE P., LIU X., GAO J. & CHEN W. (2020). Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

HENDRYCKS D., BURNS C., BASART S., ZOU A., MAZEIKA M., SONG D. & STEINHARDT J. (2021). Measuring Massive Multitask Language Understanding. In *The Ninth International Conference on Learning Representations (ICLR 2021)*, Virtual.

JIANG A. Q., SABLAYROLLES A., ROUX A., MENSCH A., SAVARY B., BAMFORD C., CHAPLOT D. S. *et al.* (2024). Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

KWON W., LI Z., ZHUANG S., SHENG Y., ZHENG L., YU C. H., GONZALEZ J. E. *et al.* (2023). Efficient Memory Management for Large Language Model Serving with PagedAttention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

LIU Y., OTT M., GOYAL N., DU J., JOSHI M., CHEN D., LEVY O. *et al.* (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

OUYANG L., WU J., JIANG X., ALMEIDA D., WAINWRIGHT C., MISHKIN P., ZHANG C. *et al.* (2022). Training language models to follow instructions with human feedback. In S. KOYEJO, S. MOHAMED, A. AGARWAL, D. BELGRAVE, K. CHO & A. OH, Éds., *Advances in Neural Information Processing Systems (NeurIPS 2022)*, volume 35, p. 27730–27744 : Curran Associates, Inc.

PENG L., WANG Z., YAO F., WANG Z. & SHANG J. (2024). MetaIE: Distilling a Meta Model from LLM for All Kinds of Information Extraction Tasks. *arXiv preprint arXiv:2404.00457*.

ROZIERE B., GEHRING J., GLOECKLE F., SOOTLA S., GAT I., TAN X. E., ADI Y. *et al.* (2023). Code llama: Open foundation models for code. *arXiv preprint arXiv:2308.12950*.

SAINZ O., GARCÍA-FERRERO I., AGERRI R., DE LACALLE O. L., RIGAU G. & AGIRRE E. (2024). GoLLIE: Annotation Guidelines improve Zero-Shot Information-Extraction. In *The Twelfth International Conference on Learning Representations (ICLR 2024)*.

SANH V., WEBSON A., RAFFEL C., BACH S., SUTAWIKA L., ALYAFEAI Z., CHAFFIN A. *et al.* (2022). Multitask Prompted Training Enables Zero-Shot Task Generalization. In *The Tenth International Conference on Learning Representations (ICLR 2022)*, Virtual.

SRIVASTAVA A., RASTOGI A., RAO A., SHOEB A. A. M., ABID A., FISCH A., BROWN A. R. *et al.* (2023). Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*.

TUO A., BESANÇON R., FERRET O. & TOURILLE J. (2023). Trigger or not Trigger: Dynamic Thresholding for Few Shot Event Detection. In J. KAMPS, L. GOEURIOT, F. CRESTANI, M. MAISTRO, H. JOHO, B. DAVIS, C. GURRIN, U. KRUSCHWITZ & A. CAPUTO, Éds., *45th European Conference on Information Retrieval (ECIR 2023): Advances in Information Retrieval*, volume 13981 de *Lecture Notes in Computer Science*, p. 637–645, Dublin, Ireland : Springer Nature Switzerland.

VASWANI A., SHAZEER N., PARMAR N., USZKOREIT J., JONES L., GOMEZ A. N., KAISER L. U. *et al.* (2017). Attention is All you Need. In I. GUYON, U. V. LUXBURG, S. BENGIO, H. WALLACH, R. FERGUS, S. VISHWANATHAN & R. GARNETT, Éds., *Advances in Neural Information Processing Systems (NIPS 2017)*, volume 30 : Curran Associates, Inc.

VINYALS O., BLUNDELL C., LILLICRAP T., KAVUKCUOGLU K. & WIERSTRA D. (2016). Matching Networks for One Shot Learning. In D. LEE, M. SUGIYAMA, U. LUXBURG, I. GUYON & R. GARNETT, Éds., *30th International Conference on Neural Information Processing Systems (NIPS 2016)*, volume 29, p. 3637—-3645, Barcelona, Spain.

WANG A., SINGH A., MICHAEL J., HILL F., LEVY O. & BOWMAN S. (2018). GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In T. LINZEN, G. CHRUPAŁA & A. ALISHAHI, Éds., *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, p. 353–355, Brussels, Belgium : Association for Computational Linguistics. DOI : 10.18653/v1/W18-5446.

WANG X., ZHOU W., ZU C., XIA H., CHEN T., ZHANG Y., ZHENG R. *et al.* (2023a). InstructUIE: multi-task instruction tuning for unified information extraction. *arXiv preprint arXiv:2304.08085*.

WANG Y., KORDI Y., MISHRA S., LIU A., SMITH N. A., KHASHABI D. & HAJISHIRZI H. (2023b). Self-Instruct: Aligning Language Models with Self-Generated Instructions. In A. ROGERS, J. BOYD-GRABER & N. OKAZAKI, Éds., *61st Annual Meeting of the Association for Computational Linguistics (ACL 2023)*, p. 13484–13508, Toronto, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/2023.acl-long.754.

YE J., XU N., WANG Y., ZHOU J., ZHANG Q., GUI T. & HUANG X. (2024). LLM-DA: Data Augmentation via Large Language Models for Few-Shot Named Entity Recognition. *arXiv preprint arXiv:2402.14568*.

ZARATIANA U., TOMEH N., HOLAT P. & CHARNOIS T. (2024). GLiNER: Generalist Model for Named Entity Recognition using Bidirectional Transformer. In *2024 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2024)*.

ZHOU W., ZHANG S., GU Y., CHEN M. & POON H. (2024). UniversalNER: Targeted Distillation from Large Language Models for Open Named Entity Recognition. In *The Twelfth International Conference on Learning Representations (ICLR 2024)*.

A Résultats complémentaires sur le jeu de données de test

Le tableau 9 se focalise sur le problème des entités discontinues. Le relâchement des contraintes s'avère bénéfique pour GLiNER ensemble, assez neutre pour GLiNER seul et assez nettement défavorable à GoLLIE. Il reste un travail d'analyse à faire à ce niveau.

Modèles	Macro-P	Macro-R	Macro-F1
GLiNER ensemble	62,78	60,6	61,46
GLiNER	56,26	54,65	55,31
GoLLIE	38,69	27,67	30,30

TABLE 9 – Résultats en considérant seulement les frontières englobantes pour les entités discontinues.

Le tableau 10 donne quant à lui les résultats obtenus sur les données de test en faisant abstraction du type d'événements EVENT, très représenté dans les données d'entraînement. La comparaison avec le tableau 8 montre que nos approches obtiennent des résultats très comparables pour les types d'entités EVENT par rapport à l'ensemble des autres types d'entités.

Modèles	Macro-P	Macro-R	Macro-F1
GLiNER ensemble	63,85	56,62	59,49
GLiNER	59,30	52,79	55,36
GoLLIE	42,97	32,13	35,83

TABLE 10 – Résultats sans le type d'entités EVENT.

Ce constat global cache néanmoins un certain nombre de disparités entre types d'entités comme le montre le tableau 11, qui donne le détail des F1-scores par type d'entités, illustré également par la figure 1.

Modèles	PERSON	FUNC.	GROUP	ORG.	MILIT.	LOC.	SITE	EQUIP.	RESRCE	TIME	EVENT	ID	UNK
GLiNER ens.	90,11	59,06	58,02	65,87	50,00	55,91	42,59	68,39	32,43	65,57	73,91	75,00	40,00
GLiNER	89,36	59,72	57,14	67,31	49,15	50,00	41,30	66,85	30,77	67,97	72,14	44,44	30,00
GoLLIE	86,32	33,85	14,97	20,00	53,97	34,93	23,81	29,10	5,56	57,06	16,87	33,33	35,90

TABLE 11 – Résultats par type d'entités (F1-score).

Ce détail permet de faire plusieurs constats. Le premier est qu'il y a globalement peu de différences, en termes de profil de résultats, entre le modèle GLiNER seul et la combinaison de modèles GLiNER. Ce constat admet toutefois quelques exceptions : pour les entités de type ID, et UNKNOWN dans une moindre mesure, GLiNER ensemble est nettement meilleur que GLiNER seul. La très faible représentation de ces entités dans l'ensemble d'apprentissage peut expliquer la prime apportée par la combinaison de modèles. Pour les autres types, la différence est le plus souvent faible, généralement à l'avantage de GLiNER ensemble, mais pas systématiquement (cf. types TIME, ORGANIZATION et FUNCTION).

Le second constat notable est que bien que les performances de GoLLIE soient globalement très en dessous de celles des modèles GLiNER, les profils de résultats des deux types de modèles sont

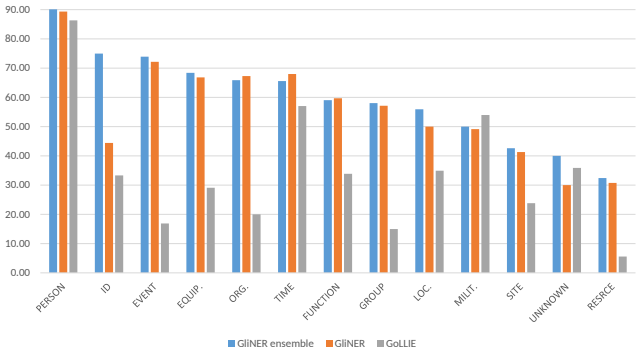

FIGURE 1 – Histogramme des résultats par type d'entités (F1-score).

différents. Ainsi, GoLLIE arrive même à dépasser GLiNER seul pour le type UNKNOWN et tous les modèles GLiNER pour le type MILITARY_UNIT. Pour UNKNOWN, la rareté de ce type d'entités est comme ci-dessus une explication possible. Le cas de MILITARY_UNIT est moins directement interprétable, sauf si les entités de ce type sont mieux représentées dans le corpus de préentraînement de GoLLIE que dans celui de GLiNER [4]. Par ailleurs, GoLLIE s'approche des modèles GLiNER pour les types PERSON et TIME. Dans le cas du type PERSON, sa forte présence dans les corpus de préentraînement des deux types de modèles peut expliquer à la fois le très haut niveau de performance atteint et la proximité des modèles, logique qui prévaut aussi pour TIME dans une moindre mesure. Mais cette caractéristique ne produit pas les mêmes effets pour les deux autres types fréquents dans les corpus de préentraînement, ORGANIZATION et LOCATION, dont les performances sont nettement en deça de celles de PERSON, avec même un niveau assez faible pour GoLLIE concernant ORGANIZATION. GoLLIE est aussi particulièrement faible pour EVENT, au contraire des modèles GLiNER, et une forte différence se retrouve aussi pour GROUP et RESOURCE, avec un niveau un peu moindre pour les modèles GLiNER. Finalement, la présence de plusieurs facteurs – en particulier le type des modèles, la fréquence des entités dans le corpus de préentraînement des modèles et la fréquence des entités dans le corpus d'entraînement EvalLLM2024 – nécessiterait de mener une analyse plus fouillée en essayant de collecter plus de données sur ces différents facteurs.

4. Ce que nous n'avons pas vérifié.

B Hyperparamètres

GLiNER

Les hyperparamètres choisis pour l'entraînement des modèles sont les suivants :

Hyperparamètre	**Valeur**
Batch size	16
Taux d'apprentissage de l'encodeur	1e-5
Taux d'apprentissage des autres éléments du modèle	5e-5
Nombre d'étapes d'entraînement	250 sans augmentation n x 100 avec augmentation de ratio n
Pas de mélange dans l'ordre des étiquettes ni de suppression aléatoire des étiquettes durant l'entraînement	

Seuls le nombre d'étapes et le choix de mélanger l'ordre des étiquettes ou de supprimer aléatoirement des étiquettes pendant l'entraînement ont été modifiés ; les autres paramètres sont ceux recommandés par les auteurs de GLiNER.

GoLLIE

Les hyperparamètres utilisés pour l'inférence de GoLLIE avec la librairie Python vLLM sont les suivants :

Hyperparamètre	**Valeur**
dtype	float16
gpu_memory_utilization	0,8
tensor_parallel_size	2
temperature	0
max_tokens	500
top_p	1
frequency_penalty	0
presence_penalty	0
best_of	1

C Impact environnemental

Nos expérimentations ont été exécutées sur le serveur FactoryIA hébergé par le Très Grand Centre de Calcul du CEA. Les coûts énergétiques ont été calculés à l'aide du calculateur Green Algorithm [5].

GLiNER

L'ensemble des calculs pour les expérimentations impliquant GLiNER (augmentation, entraînement, inférence) ont été effectués avec la configuration moyenne suivante pendant un temps total estimé à 123 heures et 18 minutes :

- 1 GPU A100 80G (TDP 400W)
- 64 GiB de mémoire RAM
- 8 cœurs CPU AMD EPYC 7543 (TDP 7W/cœur)

L'impact carbone total estimé est de 5,07 kg CO2e pour une consommation énergétique totale de 98,8 kWh. Parmi ces expériences, l'entraînement des modèles finaux et l'inférence sur les données de test ont demandé 2h24, soit 1,92 kWh et 98,62 g CO2e.

GoLLIE

Les coûts énergétiques et environnementaux des expérimentations avec GoLLIE sont détaillés dans le tableau 12. La configuration pour utiliser GoLLIE en inférence a toujours été la suivante :

- 2 GPU A100 40G (TDP 400W)
- 64 GiB de mémoire RAM
- 12 cœurs CPU AMD EPYC 7543 (TDP 7W/cœur)

Au total, ces expérimentations ont demandé 7,51 kWh et 384 g CO2e. L'inférence sur les données de test a demandé 50 minutes, soit 1,29kWh et 65,95 g C02e.

5. https://calculator.green-algorithms.org/

Expérience	Configuration	Énergie (Wh)	Empreinte carbone (g CO2e)
Zero-shot Train set	GoLLIE-7B 1 prompt par texte et avec toutes les classes	41,69	2,14
	GoLLIE-13B 1 prompt par texte et avec toutes les classes	51,29	2,63
	GoLLIE-13B 1 prompt par texte et par groupe de classes	45,48	2,33
	GoLLIE-13B 1 prompt par texte et par classe	55,59	2,85
	GoLLIE-13B 1 prompt par phrase et avec toutes les classes	54,33	2,79
	GoLLIE-13B 1 prompt par phrase et par groupe de classes	63,68	3,27
	GoLLIE-13B 1 prompt par phrase et par classe	72,01	3,69
	GoLLIE-34B 1 prompt par texte et avec toutes les classes	162,22	8,32
Few-shot Exemples aléatoires Train set	Seed=0	547,06	28,05
	Seed=12	676,68	34,7
	Seed=76	572,33	29,35
	Seed=233	552,87	28,35
	Seed=407	522,55	26,8
	Seed=859	533,66	27,37
	Moyennes	509,62	26,13
Few-shot Exemples similaires Train set	Solon-embeddings-large-0.1	882,11	45,23
	sentence-camembert-large	876,30	44,94
Few-shot Exemples aléatoires Test set	Seed=0	1290,00	65,95

TABLE 12 – Coûts énergétiques et environnementaux des expérimentations GoLLIE.

D Augmentation des données

D.1 Extraction d'exemples du guide d'annotation

Le processus d'extraction d'exemples du guide d'annotation se compose des étapes suivantes :

— Préparation des données : le guide d'annotation a été converti en fichier textuel à l'aide d'un outil de reconnaissance optique de caractères (OCR) [6].

— Segmentation du texte à l'aide d'une expression régulière pour isoler les exemples du guide d'annotation [7].

— Extraction des exemples à l'aide du modèle Mixtral-8x7B (Jiang *et al.*, 2024) avec une température de 0 (cf. section suivante).

— Correction manuelle de quelques erreurs.

6. https://github.com/ocrmypdf/OCRmyPDF
7. (Exemple\s*\d*)(.*?)(?=Exemple\s*\d*|$)

D.2 Prompt Mixtral d'extraction des exemples du guide d'annotation

```
**Task:** Extract labeled examples from the provided French text and format them into a JSON
object.

**Input Text Format:**
1. The text will include one or more sentences.
2. Labels to be extracted will be annotated within the text.
3. The main sentence and annotations are contained within quotation marks.
4. Annotations will identify different types of entities in the following list :
['EQUIPMENT',
 'EVENT',
 'FUNCTION',
 'GROUP',
 'ID',
 'LOCATION',
 'MILITARY_UNIT',
 'ORGANIZATION',
 'PERSON',
 'RESOURCE',
 'SITE',
 'TIME',
 'UNKNOWN'].


**Example Input:**
```
Exemple 2:
« La secrétaire au bureau du Conseil de Sécurité de l'ONU. » On annote trois organisations qui
 sont imbriquées: « bureau du Conseil de Sécurité de l'ONU », « Conseil de Sécurité de l'ONU »
, « l'ONU ». On annote aussi « la secrétaire au bureau du Conseil de Sécurité de l'ONU » comme
 FUNCTION (voir 3.3).
```

**Expected JSON Output:**
```json
{
    "text": "La secrétaire au bureau du Conseil de Sécurité de l'ONU.",
    "label": [
        {"text": "bureau du Conseil de Sécurité de l'ONU", "label": "ORGANIZATION"},
        {"text": "Conseil de Sécurité de l'ONU", "label": "ORGANIZATION"},
        {"text": "l'ONU", "label": "ORGANIZATION"},
        {"text": "la secrétaire au bureau du Conseil de Sécurité de l'ONU", "label": "FUNCTION
"}
    ]
}
```

Steps for Extraction:
1. Identify the main sentence from the input text (contained within quotation marks).
2. Extract all the annotated entities and their corresponding labels.
3. Format the extracted data into the specified JSON structure.

Additional instructions
1. Do not translate, the text must match label texts.
2. Do not annotate twice the same span.
3. Only label explicitely labeled entities.

Input:
```

D.3 Prompt Mixtral d'augmentation des données

```
You are a French synonym system that can provide a similar string given an input string.
Provide a similar string to the following string, by changing the number of words and/or
modifying the meaning of one word. You must keep the grammatical structure and you will only
respond in French with a JSON object with one key "New_String". Do not provide explanation or
any other element.
### Examples :
Le Ministère de la Santé => {"New_String":"Le Département des affaires étrangères"}
3 => {"New_String":"5"}
Le fils du président => {"New_String":"Les cousins du ministre"}
Messieurs => {"New_String":"Madame"}
Espagnol => {"New_String":"Belge"}
premier => {"New_String":"dixième"}

### Context :

### Input String :

### Answer :
```

E Stratégie d'ensemble


```
{
    "EQUIPMENT": "union",
    "PERSON": "relative-majority",
    "LOCATION": "relative-majority",
    "ORGANIZATION": "relative-majority",
    "EVENT": "relative-majority",
    "FUNCTION": "relative-majority",
    "GROUP": "relative-majority",
    "ID": "relative-majority",
    "MILITARY_UNIT": "relative-majority",
    "RESOURCE": "relative-majority",
    "SITE": "relative-majority",
    "TIME": "relative-majority",
    "UNKNOWN": "union"
}
```

F Directives d'annotation GoLLIE

Les directives d'annotation du benchmark EvalLLM2024 pour GoLLIE, définies à partir du guide d'annotation fourni, sont détaillées ci-dessous :

```
from typing import List
from src.utils.utils_GoLLIE.utils_typing import Entity, dataclass

"""
Entity definitions
"""


@dataclass
class Person(Entity):
    """Désigne des individus, y compris des personnes réelles, des avatars et des
    noms propres.
    Cette classe englobe les noms propres d'individus ou d'avatars, y compris les
    titres (M., Mme, etc.)
    et les acronymes (par exemple, "JFK" pour John Fitzgerald Kennedy)."""

    span: str  # Such as: "Jean Dupont", "Céline Martin", "Mr. Durand", "Dr.
    Nguyen", "Louis XIV", "M. Le Président", "Juliette Rousseau", "Sophie Martin",
     "Napoléon Bonaparte", "CDG"


@dataclass
class Function(Entity):
    """Désigne les rôles, titres ou fonctions occupés par des individus, y compris
     les titres professionnels,les grades militaires et les fonctions nommées.
    Cette classe englobe les fonctions, les titres et les grades militaires, ainsi
     que les individus désignés par leur profession."""

    span: str  # Such as: "Président de la France", "Chef Chirurgien", "PDG de la
    société XYZ", "Chef Pâtissier Exécutif", "Docteur", "Ingénieurs", "Professeur
    ", "Général de la 2e division", "Capitaine", "Manager"


@dataclass
class Unknown(Entity):
    """Désigne des individus ou entités non identifiés qui ne peuvent pas être
    classés comme Personne ou Fonction.
    Cette classe représente des individus non nommés, y compris les titres de
    civilité non suivis d'un nom propre,
    les formulations qui ne permettent pas d'identification précise et les
    mentions de relations."""

    span: str  # Such as: "un suspect", "un informateur", "une victime", "l'auteur
    ", "une source anonyme", "le défunt", "le coupable", "un passant", "l'
    assaillant", "le malfaiteur"


@dataclass
class Organization(Entity):
    """Désigne des structures organisées ayant une reconnaissance légale, y
    compris des États, des gouvernements,
    des institutions publiques, des organisations internationales, judiciaires,
    diplomatiques, scientifiques, éducatives,
```

```
    culturelles, bancaires, sportives et sanitaires, ainsi que des entités privées
     telles que des entreprises."""

    span: str  # Such as: "Organisation des Nations Unies", "Organisation mondiale
     de la santé", "Google", "Apple Inc.", "Union européenne", "Croix-Rouge", "
    Fonds monétaire international", "NASA", "FIFA", "UNESCO"

@dataclass
class MilitaryUnit(Entity):
    """Désigne toutes les organisations qui relèvent du militaire.
    Cela inclut les armées et les composantes d'armée régulières, ainsi que les
    désignations « vagues » d'armées.
    Les bâtiments de la Marine sont également concernés, tels que les frégates,
    portehélicoptères et porte-avions, etc."""

    span: str  # Such as: "122e bataillon d'infanterie russe", "PROVENCE", "BAT10
    ", "septième unité d'infanterie", "commandement américain", "forces de
    sécurité afghanes", "la coalition", "troupes de la coalition", "marine
    argentine", "armées opposées"


@dataclass
class Group(Entity):
    """Désigne les groupes d'individus, organisés ou non, qui ne relèvent pas de
    la définition d'Organization,, MilitaryUnit ou Function.
    Cette étiquette inclut les groupes criminels organisés, la piraterie, la mafia
    , les hackers, les trafiquants, les manifestants, les
    migrants, une population, une tribu, les terroristes, une secte, une religion,
     une ethnie.
    Un sous-ensemble d'individus provenant d'une organisation, qui ne représente
    pas une organisation en soi, est considéré comme un Group."""

    span: str  # Such as: "Yakuza", "Pirates somaliens", "Cosa Nostra", "Anonymous
    ", "Cartels de drogue", "Manifestants à Hong Kong", "Réfugiés syriens", "Tribu
     Maasaï", "Al-Qaïda", "Scientology"



@dataclass
class Location(Entity):
    """Désigne les noms de lieux, qu'ils soient géographiques ou administratifs (
    ville, pays, département, région, continent, etc.).
    On prend aussi en compte les adresses et les coordonnées GPS, peu importe leur
     système de notation : latitude/longitude, le système standard de lOTAN (MGRS)
    , WGS84, etc."""

    span: str  # Such as: "Paris", "France", "New York City", "Los Angeles", "Mont
     Everest", "Tour Eiffel", "1600 Pennsylvania Avenue", "Latitude: 40.7128,
    Longitude: -74.0060", "N 48° 51' 32.3400", "E 2° 21' 49.7400"


@dataclass
class Site(Entity):
    """Désigne les lieux occupés dans un but précis ou pour un intérêt ponctuel
    dans un évènement, une construction ou un lieu qui fait lobjet d'un
    aménagement.
    On considère comme site les lieux importants pour une activité industrielle,
    économique ou militaire («pont», «centrale électrique », « centre commercial »
    ).
```

```
    L'occupation peut être permanente (« base militaire »), éphémère (« ligne de
    front ») ou en transit.
    Les sites sont des lieux à léchelle d'un bâtiment ou dune infrastructure."""

    span: str  # Such as: "Pont du Gard", "Barrage des Trois-Gorges", "Centre
    commercial Les Quatre Temps", "Zone 51", "Tour Eiffel", "Stand de tir", "
    Raffinerie de pétrole", "Pas de tir spatial", "Chantier de construction", "
    Base militaire"

@dataclass
class Equipment(Entity):
    """Désigne lensemble du matériel pouvant appartenir à une organisation ou une
    personne.
    Ces équipements peuvent être des objets, des systèmes techniques (matériels)
    ou de télécommunication, des véhicules, etc."""

    span: str  # Such as: "Ordinateur", "Arme à feu", "Téléphone satellite", "
    Drone", "Machines de construction", "Équipement médical", "Véhicule militaire
    ", "Smartphone", "Avion", "Équipement de laboratoire"

@dataclass
class Resource(Entity):
    """Désigne moyens et produits dont disposent les organisations.
    Cette catégorie inclut les ressources naturelles, les produits industriels ou
    agricoles, les objets numériques (services, logiciels, protocoles de
    communication, réseaux sociaux, programmes),
    les ressources monétaires, et les documents tels que les papiers d'identité,
    les rapports et les documents multimédias."""

    span: str  # Such as: "Pétrole", "Électricité", "Médicaments", "Logiciel", "
    Internet", "Billets de banque", "Passeports", "Rapports officiels", "Documents
     financiers", "Enregistrements vidéo"

@dataclass
class Time(Entity):
    """Désigne toutes les dates absolues, relatives, les périodes, etc.
    On considère aussi toutes les marques de fréquence et les connecteurs logiques
    .
    On n'annote que la mention la plus précise, c'est-à-dire quon ne considère pas
     d'imbrication."""

    span: str  # Such as: "1er janvier 2022", "Hier", "La semaine dernière", "
    Chaque lundi", "L'année prochaine", "De 2010 à 2020", "Dans les années 1990",
    "Pendant l'été 2018", "Pour les trois prochains mois", "12 avril 2024"


@dataclass
class Id(Entity):
    """Désigne l'ensemble des informations intervenant dans un système
    informatique et/ou représentant une personne, permettant le contact par des
    moyens électroniques ou technologiques.
    Cela inclut les adresses IP, les URL de site web, les e-mails, les numéros de
    téléphone, les noms d'utilisateur, les identifiants de connexion, les mots de
    passe, les numéros d'immatriculation, les numéros d'enregistrement, etc."""

    span: str  # Such as: "192.168.1.1", "www.example.com", "user@example.com",
    "+1234567890", "john_doe123", "admin", "password123", "user123", "ABC123",
    "123456"
```

```
@dataclass
class Event(Entity):
    """Désigne les évènements historiques et ceux au sein du texte, indiqués par
    des amorces d'évènements.
    Un évènement recoupe tous les faits significatifs, y compris ceux avec des
    conséquences, marquant un changement d'état ou influençant leur contexte de
    production.
    Les verbes de parole sont systématiquement annotés en EVENT, que le sujet soit
     une personne ou toute autre entité."""

    span: str  # Such as: "La seconde Guerre Mondiale", "les guerres
    napoléoniennes", "le communiqué de presse", "l'annonce", "l'accord", "la
    réunion au sommet", "le jugement du tribunal", "le verdict", "la négociation",
     "la résolution du conflit", "déclare", "affirme", "confirme"


ENTITY_DEFINITIONS: List[Entity] = [
    Person,
    Function,
    Unknown,

    Organization,
    MilitaryUnit,
    Group,

    Location,
    Site,

    Equipment,
    Resource,

    Time,

    Id,

    Event
]

if __name__ == "__main__":
    cell_txt = In[-1]
```

Listing 1 – Directive d'annotation

Groupes de classes. Les classes identifiées comme ayant des définitions proches ont été regroupées en groupes dans le cadre de certaines expérimentations de la section 3.3.1. Certaines classes n'ayant pas de lien avec une ou plusieurs autres classes n'ont pas été associées à un groupe. Voici les groupes formés :

- PERSON, FUNCTION, UNKNOWN
- ORGANISATION, MILITARY_UNIT, GROUP
- LOCATION, SITE
- EQUIPMENT
- RESOURCE
- TIME
- ID

— EVENT