Actes de CORIA-TALN 2026
Nantes, France, 29 juin au 3 juillet 2026

Évaluation d'un modèle bi-encodeur généraliste pour l'extraction de relations documentaires en contexte de données limitées

Robin Armingaud Romaric Besançon
Université Paris-Saclay, CEA, List, F-91120, Palaiseau, France
{robin.armingaud, romaric.besancon}@cea.fr

RÉSUMÉ

L'extraction de relations documentaires (ER-DOC) est complexe en raison des interactions distantes entre entités. Bien que des modèles performants comme ATLOP existent, leur efficacité avec peu de données est peu étudiée. Nous proposons d'évaluer un modèle compact bi-encodeur pour l'ER-DOC, pré-entraîné sur des données synthétiques distillées, efficace en supervisé et en *few-shot*. Des tests sur trois datasets montrent que ce modèle surpasse les méthodes existantes en contexte de données limitées et atteint des performances en *zero-shot* comparables à celles de grands modèles, à moindre coût. Le code et le modèle seront disponibles publiquement.

ABSTRACT

A bi-encoder model for few-shot document-level relation extraction

Document-level relation extraction (DocRE) is complex due to distant interactions between entities. While effective models like ATLOP exist, their performance with limited data is understudied. We propose evaluating a compact bi-encoder model for DocRE, pre-trained on distilled synthetic data, effective in both supervised and *few-shot* settings. Tests on three datasets show that this model outperforms existing methods in low-resource scenarios and achieves *zero-shot* performance comparable to large models at a lower cost. The code and model will be publicly available.

MOTS-CLÉS : Extraction de relations documentaires, few-shot, pré-entraînement.

KEYWORDS: Document-level relation extraction, few-shot, pretraining.

1 Introduction

L'extraction de relations documentaires (ER-DOC) a pour but d'identifier des relations entre entités à l'échelle du document (Figure 1). Cette tâche constitue un défi majeur du traitement du langage naturel car elle exige l'identification de liens entre entités parfois très distantes dans le texte. Contrairement à l'extraction au niveau de la phrase, la tâche d'ER-DOC ne se limite pas à la simple classification d'une unique paire d'entités et offre un cadre d'évaluation complexe mais plus réaliste et plus proche des applications industrielles ou biomédicales où les relations dépassent les frontières de la phrase et où les paires d'entités d'intérêt ne sont pas connues à l'avance. Cette complexité est accentuée par la croissance quadratique des paires négatives par rapport au nombre d'entités dans un document. L'évaluation de référence s'appuie généralement sur des jeux de données tels que DocRED ou Re-DocRED (Yao *et al.*, 2019; Tan *et al.*, 2022b).

Cependant, si l'extraction de relations au niveau de la phrase dans un contexte avec peu de ressources

Actes des 33ème Conférence sur le Traitement Automatique des Langues Naturelles.
Volume 1 : articles scientifiques originaux, pages 221–237.

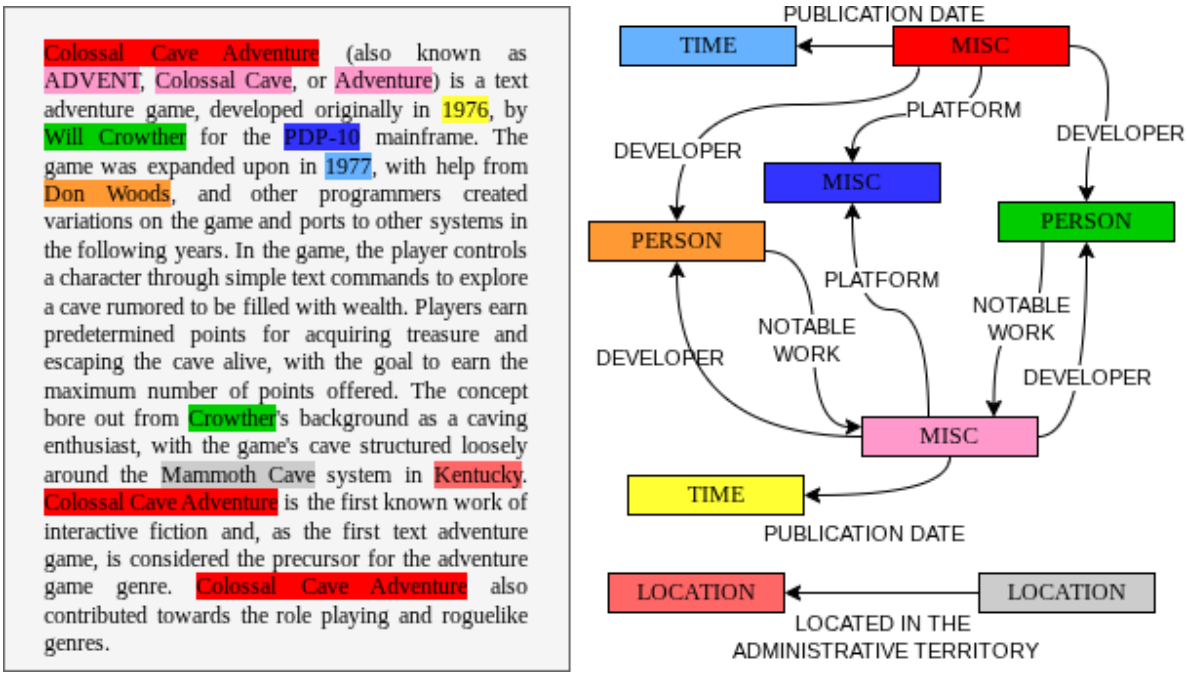

FIGURE 1 – Exemple d'extraction de relations au niveau d'un document : le texte et les annotations sont issus du jeu de données Re-DocRED, les mentions d'une même entité sont surlignées de la même couleur et, dans le graphe des relations, les mentions sont représentées par le type d'entité auquel elles sont associées.

a été largement documentée (Boylan *et al.*, 2025; Li *et al.*, 2024a; Lan *et al.*, 2023), les recherches sur l'évaluation de la tâche d'ER-DOC dans des conditions similaires restent limitées. Dans un contexte *zero-shot*, si les grands modèles de langage (LLM) sont très performants pour la reconnaissance d'entités nommées (NER) et l'extraction de relations (ER) (Sainz *et al.*, 2024; Zhou *et al.*, 2023; Wang *et al.*, 2023; Wei *et al.*, 2023), leurs performances sur l'ER-DOC *zero-shot* demeurent restreintes (Li *et al.*, 2023; Xue *et al.*, 2024). À l'opposé, des modèles encodeurs légers comme GLiNER (Zaratiana *et al.*, 2024) montrent de bonnes performances en exploitant la similarité entre les représentations des types d'entités et les segments textuels. Cette approche, grâce à l'utilisation d'encodeurs bidirectionnels, résout les problèmes de mise à l'échelle des modèles autorégressifs et surpasse souvent des LLM bien plus coûteux. De plus, ces modèles *zero-shot* ne sont pas restreints à un ensemble d'étiquettes appris pendant l'entraînement. Dans cette optique, cet article se concentre sur l'évaluation du modèle bi-encodeur GLiDRE (Armingaud, 2025), une adaptation de GLiNER pour l'ER-DOC. Nos contributions sont les suivantes :

— Nous menons une analyse approfondie des performances de GLiDRE dans un large spectre de situations sur les jeux de données DocRED, Re-DocRED, FREDo et Re-FREDo, en montrant que ce modèle léger spécialisé présente des performances compétitives dans les scénarios avec peu de données.
— Nous démontrons que des modèles encodeurs plus petits et spécialisés peuvent être une alternative efficace à des grands modèles de langages pour la tâche complexe d'ER-DOC tout en présentant une évaluation à jour de ces modèles sur le jeu de données Re-DocRED dans un contexte *zero-shot*.

2 État de l'art

Méthodes d'ER au niveau document De nombreuses approches supervisées en ER-DOC étendent le modèle ATLOP (Zhou *et al.*, 2021), qui introduit un module de contexte local et une fonction de perte à seuil adaptatif. DREEAM (Ma *et al.*, 2023) utilise en plus des annotations de preuves, tandis que KD-DocRE (Tan *et al.*, 2022a) emploie une attention axiale et une distillation de connaissances. D'autres travaux ajoutent des contraintes de règles (Liu *et al.*, 2023; Zhang *et al.*, 2025) ou exploitent les LLM comme LMRC (Li *et al.*, 2024b), combinant un classifieur dédié et les capacités génératives d'un LLM.

Jeux de données d'ER-DOC Les modèles sont souvent évalués sur DocRED (Zhou *et al.*, 2021) et Re-DocRED (Tan *et al.*, 2022c), qui contiennent des documents complexes annotés avec des entités et paires candidates. D'autres jeux de données incluent HacRED (Cheng *et al.*, 2021) et SciERC (Luan *et al.*, 2018), mais DocRED et Re-DocRED restent les seuls avec des annotations de preuves, essentielles pour des modèles comme DREEAM.

ER-DOC dans un contexte *few-shot* L'évaluation de la tâche en contexte *few-shot* a reçu une attention limitée. Une stratégie directe d'évaluation consiste à entraîner les modèles dans des régimes à données contraintes, où seul un petit sous-ensemble d'exemples étiquetés est disponible. Allant au-delà de ce paradigme, FREDo (Popovic & Färber, 2022) et Re-FREDo (Meng *et al.*, 2023) formulent l'extraction de relations comme un problème de méta-apprentissage avec tâches épisodiques, utilisant DocRED et Re-DocRED. SciERC sert de test hors domaine pour évaluer la généralisation. Des approches comme RAPL (Meng *et al.*, 2023) et TPN (Zhang & Kang, 2024) utilisent un apprentissage par prototypes sur ces benchmarks.

3 Méthodologie

3.1 Extraction de relations documentaires

On représente un document par une séquence de mots. Dans ce document, on considère un ensemble de m entités pré-identifiées $E = \{e_1, e_2, \ldots, e_m\}$, où chaque entité e_i est associée à une ou plusieurs mentions au sein du texte (les mentions sont données). L'objectif de l'ER-DOC est d'identifier les triplets relationnels valides. Chaque triplet est de la forme (e_h, r, e_t), où $e_h, e_t \in E$ représentent respectivement les entités tête et queue (où h peut être égal à t), et r correspond à la relation établie entre elles. Selon cette définition, l'ER-DOC correspond à une tâche de classification multi-étiquettes sur chaque paire d'entités, où les étiquettes correspondent aux types de relations.

3.2 Architecture de GLiDRE

L'architecture de GLiDRE (Figure 2) s'inspire de la variante bi-encodeur de GLiNER [1], et utilise deux encodeurs distincts, l'un pour l'encodage du document et l'autre pour celui des étiquettes

1. https://blog.knowledgator.com/meet-the-new-zero-shot-ner-architecture-30ffc2cb1ee

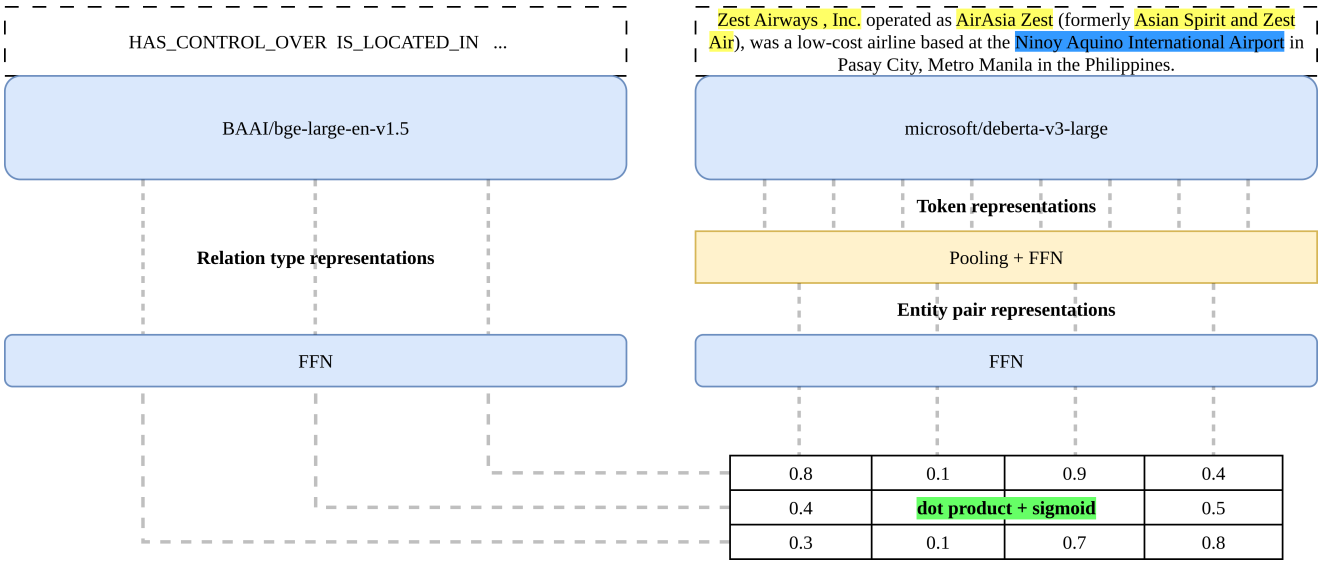

FIGURE 2 – Architecture de GLiDRE : modèle bi-encodeur pour représenter le texte et les types de relations.

de relations, ce qui offre plusieurs avantages : contrairement à la version uni-encodeur, l'approche bi-encodeur évite de concaténer les étiquettes et les jetons du document dans le contexte d'un unique encodeur. À la place, le document et les étiquettes sont encodées séparément. Cela élimine le besoin de *tokens* de séparation spéciaux et permet de s'affranchir des limites de longueur de contexte de l'encodeur, autorisant ainsi un ensemble d'étiquettes potentiellement illimité. De plus, les plongements lexicaux d'étiquettes peuvent être précalculés, ce qui accélère l'inférence. Toutefois, l'architecture bi-encodeur augmente l'empreinte mémoire et ne permet pas d'attention croisée entre les étiquettes. Par conséquent, elle peut éprouver des difficultés à désambiguïser des étiquettes sémantiquement proches.

Représentations des étiquettes et des mots Les plongements lexicaux d'étiquettes sont obtenus par la moyenne des représentations des mots constituant les noms de ces étiquettes. Les plongements d'étiquettes ainsi calculés passent par un réseau *feed-forward* à deux couches si la dimension de sortie de l'encodeur d'étiquettes diffère de la dimension de l'espace latent configurée pour GLiDRE. Pour les mots du document, la stratégie classique des modèles de NER standards est adoptée : pour les mots décomposés en sous-mots, la représentation du premier sous-mot est extraite.

Représentations des relations À partir des plongements lexicaux $H \in \mathbb{R}^{L \times D}$ issus de l'encodeur de document (où L est la longueur de la séquence et D la dimension cachée de l'encodeur), une représentation pour chaque relation candidate définie par une paire d'entités est construite : (1) pour chaque mention d'une entité, la représentation de la mention est calculée en effectuant l'agrégation moyenne des plongements des mots qui la constituent ; (2) la représentation d'une entité h_e est dérivée en calculant la moyenne des représentations de ses mentions ; (3) la représentation h_r de la relation entre e_h et e_t est obtenue par la concaténation (notée $\otimes$) de leurs représentations d'entités respectives et en les traitant via un réseau *feed-forward* à deux couches.

$$h_r = \text{FFN}(h_{e_h} \otimes h_{e_t}) \tag{1}$$

Calcul des scores de relation Pour déterminer si une paire d'entités candidate (e_h, e_t) est un type de relation t donné, le score de similarité suivant entre la représentation de la relation h_r et la représentation du type de relation h_t est mesuré :

$$s(e_h, e_t, t) = \sigma(h_r^\top h_t) \tag{2}$$

où $\sigma(x) = (1 + e^{-x})^{-1}$ est la fonction sigmoïde. Une relation t est assignée à la paire d'entités (e_h, e_t) si $s(e_h, e_t, t) > \tau$ où τ est un seuil de décision prédéfini.

3.3 Fonction de perte

Afin de remédier au problème de déséquilibre des classes dans les jeux de données d'ER-DOC, la fonction de perte *Focal Loss* proposée par Lin *et al.* (2017) est utilisée lors de l'entraînement. La *Focal Loss* est une fonction de perte d'entropie croisée mise à l'échelle dynamiquement, conçue pour réduire le poids de la contribution des exemples bien classés et concentrer l'apprentissage sur les exemples difficiles. Elle est définie comme suit :

$$\mathrm{FL}(p_t) = -\alpha_t (1 - p_t)^\gamma \log(p_t) \tag{3}$$

où p_t représente la probabilité estimée par le modèle pour la classe de la vérité terrain. Le paramètre de focalisation γ ajuste la vitesse à laquelle les exemples faciles sont dépondérés, et le facteur de pondération α_t permet d'équilibrer l'importance relative des exemples positifs et négatifs.

3.4 Contexte Local

Afin d'améliorer les représentations, plusieurs méthodes inspirées d'ATLOP (Zhou *et al.*, 2021) sont utilisées, notamment le *Localized Context Pooling*. Cette technique utilise les scores d'attention de l'encodeur du document pour se focaliser sur les parties les plus pertinentes pour une paire d'entités donnée. Soit $A \in \mathbb{R}^{L \times L}$ la matrice d'attention au niveau du document issue de la dernière couche de l'encodeur, moyennée sur l'ensemble des têtes d'attention. Pour l'entité tête e_h et l'entité queue e_t, leurs vecteurs d'attention correspondants A_h et A_t sont extraits en moyennant les scores d'attention sur les *tokens* de leurs mentions respectives. Un vecteur d'attention conjointe α est ensuite calculé via un produit d'Hadamard, puis normalisé et utilisé pour calculer une somme pondérée des plongements lexicaux, créant ainsi un vecteur de contexte localisé c_{loc}

$$\alpha = A_h \odot A_t \qquad\qquad c_{\text{loc}} = \sum_{i=1}^{L} \frac{\alpha_i}{\sum_{j=1}^{L} \alpha_j} H_i \tag{4}$$

Enfin, les représentations affinées pour les entités tête h'_{e_h} et queue h'_{e_t} sont calculées en concaténant leurs représentations initiales avec le vecteur de contexte localisé, puis en les traitant par deux réseaux *feed-forward* à deux couches :

$$h'_{e_h} = \tanh(\mathrm{FFN}_h(h_{e_h} \otimes c_{\text{loc}})) \qquad\qquad h'_{e_t} = \tanh(\mathrm{FFN}_t(h_{e_t} \otimes c_{\text{loc}})) \tag{5}$$

La représentation finale de h_r est alors calculée à partir de h'_{e_h} et h'_{e_t} en utilisant l'équation 1.

3.5 Données de pré-entraînement

Pour renforcer ses capacités de généralisation, GLiDRE distille de la connaissances à partir d'un grand modèle de langage. Conformément aux pratiques usuelles, nous qualifions cette étape de *pré-entraînement*. Son corpus de pré-entraînement a été constitué via une méthodologie d'annotation semi-automatisée inspirée de Zhou *et al.* (2023). Les documents sont d'abord échantillonnés aléatoirement à partir de FineWeb (Penedo *et al.*, 2024), un jeu de données de haute qualité construit à partir d'archives Common Crawl en anglais, filtrées et dédoublées.

Qwen3-30B-A3B-Thinking-2507 est ensuite utilisé pour générer des annotations pour les entités et pour les relations qui les lient sous un format JSON. Les sorties brutes sont filtrées pour éliminer les instances contenant du JSON mal formaté ou les documents dépassant un seuil de longueur prédéfini.

Le jeu de données résultant comprend 70 412 documents dotés d'étiquettes extrêmement variées, avec 45 047 types de relations uniques. Parmi les types de relations les plus fréquents, on trouve notamment `IS_LOCATED_IN`, `PLAYS_FOR`, `PART_OF` et `USES`. Ce jeu de données sert de base au pré-entraînement du modèle.

4 Expériences

4.1 Contextes d'évaluation

Entraînement supervisé standard Nous évaluons les performances du modèle sur le jeu de données anglais Re-DocRED (Tan *et al.*, 2022c), une version révisée par des experts humains de DocRED pour corriger le taux élevé de faux négatifs, les incohérences logiques et les erreurs de coréférence. Ce corpus contient environ 3000 documents d'entraînement et 500 documents pour la validation et le test, avec, en moyenne, une vingtaine d'entités par document et un trentaine de triplets à identifier. Nous utilisons trois configurations distinctes, en mettant l'accent sur les régimes avec peu de ressources où les avantages de GLiDRE sont les plus marqués : (1) nous affinons différents modèles sur des sous-ensembles de données d'entraînement échantillonnés aléatoirement, contenant 1, 5, 10, 50, 100, 500 ou 1000 documents, (2) nous comparons le modèle affiné sur l'ensemble de l'échantillon d'entraînement à des modèles de référence à l'état de l'art, (3) nous examinons les performances en *zero-shot* en comparant GLiDRE à des LLM plus massifs sans aucun ajustement spécifique à la tâche, soulignant ainsi la compétitivité du modèle dans des scénarios avec peu ou pas de données étiquetées.

Meta-apprentissage épisodique L'évaluation est faite sur les benchmarks FReDo (Popovic & Färber, 2022) et Re-FReDo (Meng *et al.*, 2023), qui réorganisent les documents de DocRED, Re-DocRED et SciERC dans un format épisodique. L'évaluation est structurée en deux tâches : une tâche intra-domaine comprenant environ 15 000 épisodes issus de DocRED, et une tâche hors-domaine plus complexe de 3 000 épisodes issus de SciERC. Chaque épisode se compose d'un petit ensemble de support de 1 ou 3 documents et d'un ensemble de requête. Pour garantir une évaluation *few-shot* rigoureuse, les types de relations pour l'entraînement, le développement et le test sont strictement disjoints. Nous avons adopté un protocole d'affinage en deux étapes afin de garantir une comparaison équitable avec les méthodes basées sur les prototypes. D'abord, nous réalisons une phase d'alignement initial en entraînant le modèle pendant 500 étapes sur l'ensemble

d'entraînement complet, principalement pour aligner les représentations des étiquettes (par exemple, pour la cohérence de la capitalisation). Ensuite, pour chaque épisode, nous affinons le modèle pendant 20 époques sur l'ensemble de support avec 1 ou 3 exemples fourni selon la configuration, avant d'évaluer ses performances sur l'ensemble de requête de l'épisode.

Métriques d'évaluation Nous utilisons les métriques standards de Re-DocRED (Tan *et al.*, 2022c) : *F1* correspond au score F1 en micro-moyenne sur les triplets de relations, et *Ign_F1* est le score F1 excluant les triplets de l'ensemble de test qui apparaissent également dans l'ensemble d'entraînement.

Pour la reproductibilité des résultats, les détails d'implémentation du modèle sont fournis en annexe A.1.

4.2 Résultats

4.2.1 Contexte supervisé avec peu de données

Dans les scénarios avec peu de ressources, nous évaluons l'efficacité du modèle en le comparant à deux modèles supervisés à l'état de l'art : **DREEAM** (Ma *et al.*, 2023) et **ATLOP** (Zhou *et al.*, 2021) (ces modèles sont détaillés en annexe A.2). De plus, nous comparons GLiDRE au LLM Qwen3-30B-A3B-Thinking-2507 en ajoutant les exemples d'entraînements dans le contexte du modèle (*in-context learning* ou ICL), en utilisant les mêmes sous-ensembles que ceux fournis aux modèles supervisés.

Cette comparaison permet de déterminer la taille critique du jeu de données à partir de laquelle ces méthodes supervisées conventionnelles égalent les performances du modèle, permettant ainsi de quantifier dans quels scénarios GLiDRE présente un avantage dans un contexte de données limitées.

Modèle	**N = 1**	**N = 5**	**N = 10**	**N = 50**	**N = 100**	**N = 500**	**N = 1000**
ATLOP	$4.32_{\pm 3.19}$	$18.76_{\pm 4.93}$	$29.48_{\pm 3.91}$	$50.36_{\pm 1.18}$	$57.17_{\pm 0.37}$	$68.91_{\pm 0.43}$	$71.70_{\pm 0.20}$
DREEAM	$4.27_{\pm 3.50}$	$16.02_{\pm 7.46}$	$27.07_{\pm 5.81}$	$52.05_{\pm 2.00}$	$58.14_{\pm 0.86}$	$\mathbf{69.32}_{\pm 0.31}$	$71.67_{\pm 0.32}$
GLiDRE (no pre.)	$10.53_{\pm 7.51}$	$16.95_{\pm 5.12}$	$27.24_{\pm 4.13}$	$45.71_{\pm 1.88}$	$52.21_{\pm 0.90}$	$65.26_{\pm 0.43}$	$69.10_{\pm 0.20}$
GLiDRE	$\mathbf{26.12}_{\pm 6.97}$	$\mathbf{35.17}_{\pm 5.09}$	$\mathbf{42.82}_{\pm 2.21}$	$\mathbf{54.99}_{\pm 1.77}$	$\mathbf{60.20}_{\pm 0.76}$	$69.01_{\pm 0.66}$	$\mathbf{72.01}_{\pm 0.32}$
LLM							
Qwen (ICL)	$25.43_{\pm 2.04}$	$24.71_{\pm 1.28}$	$27.02_{\pm 1.21}$	$27.32_{\pm 1.25}$	$24.06_{\pm 0.55}$	OOC	OOC

TABLE 1 – Performances en F1 (micro-moyenne) pour GLiDRE, ATLOP, DREEAM et Qwen3-30B-A3B-Thinking-2507 selon différentes tailles d'ensemble d'entraînement (N). Tous les modèles utilisent les mêmes 5 sous-ensembles pour chaque valeur de N. Les meilleurs résultats par colonne sont indiqués en **gras**.

Les résultats présentés dans le Tableau 1 démontrent clairement la supériorité de GLiDRE dans les scénarios avec peu de données. Dans les cas les plus limités ($N \leq 100$), GLiDRE devance largement ATLOP et le modèle de référence de l'état de l'art, DREEAM. Cet avantage significatif se maintient jusqu'à $N = 100$ échantillons, l'écart de performance ne commençant à se réduire que lorsque le volume de données d'entraînement atteint 500 et 1000 échantillons. L'étude d'ablation (*GLiDRE no pre.*) confirme que la phase de pré-entraînement synthétique constitue le principal moteur de cette efficacité dans un contexte de faibles ressources. Par ailleurs, bien que Qwen affiche des performances correctes avec un très faible nombre d'exemples, ses résultats n'augmentent pas de

manière proportionnelle à N. Dès $N = 500$, la quantité d'exemples dépasse la fenêtre de contexte du modèle (noté *OOC*, pour *Out of Context*).

4.2.2 Contexte de méta-apprentissage épisodique

Modèle	FREDo				ReFREDo			
	Intra-Domaine		Hors-Domaine		Intra-Domaine		Hors-Domaine	
	1-Doc F_1	3-Doc F_1	1-Doc F_1	3-Doc F_1	1-Doc F_1	3-Doc F_1	1-Doc F_1	3-Doc F_1
DL-Base	0.60	0.89	1.76	1.98	1.38	1.84	1.76	1.98
DL-MNAV	7.05	8.42	0.84	0.48	12.97	12.43	1.12	2.28
DL-MNAV$_{SIE}$	7.06	6.77	1.77	2.51	13.37	12.00	1.39	2.92
DL-MNAV$_{SIE+SBN}$	1.71	2.79	2.85	3.72	4.59	5.43	2.84	3.86
RAPL	8.75	10.67	3.33	5.35	15.20	16.35	3.51	5.48
TPN	9.12	8.64	3.98	4.48	15.54	15.73	4.72	5.02
GLiDRE	12.44	14.71	**8.87**	6.46	23.00	27.35	**9.05**	6.95
utilisant un LLM (Pan et al., 2025)								
ChatGPT$_{ICL}$	2.25	2.95	5.22	5.83	2.86	5.39	5.22	5.83
Llama-3-8B-Instruct$_{ICL}$	2.04	2.27	5.05	5.53	3.07	2.36	5.05	5.53
Llama-3-8B-Instruct$_{MetaICL}$	13.81	14.67	4.50	5.46	21.14	23.89	5.79	5.49
Llama-3-8B-Instruct$_{PT}$	**14.98**	**16.83**	7.42	**7.54**	**31.54**	**33.12**	8.10	**8.69**

TABLE 2 – Performances en *few-shot* episodique sur les jeux de données FREDo et ReFREDo, en macro-moyenne et reportés de (Meng *et al.*, 2023) (Zhang & Kang, 2024) et (Pan *et al.*, 2025). Les meilleurs résultats par colonne sont affichés en **gras**.

Pour cette évaluation, nous suivons le protocole épisodique défini par FREDo et Re-FREDo. Nous comparons GLiDRE à quatre méthodes prototypiques spécifiquement développées pour l'ER-DOC en *few-shot* épisodique : **DL-MNAV** (Popovic & Färber, 2022), **RAPL** (Meng *et al.*, 2023), **TPN** (Zhang & Kang, 2024) et **Prototype Tuning** (Pan *et al.*, 2025). Les détails de ces modèles sont fournis en annexe A.2.

Comme le montre le Tableau 2, GLiDRE atteint des performances comparables sur les deux benchmarks par rapport aux approches basées sur des LLM possédant 10× plus de paramètres. GLiDRE est particulièrement performant dans le contexte hors-domaine, où la tâche consiste à généraliser vers des relations issues du jeu de données SciERC, soulignant ainsi l'efficacité du pré-entraînement sur des données synthétiques pour le transfert de domaine.

4.2.3 Contexte supervisé sur tout l'ensemble d'entraînement

Pour l'apprentissage supervisé sur un large ensemble d'entraînement, nous comparons GLiDRE à plusieurs modèles à l'état de l'art : en complément de **DREEAM** et **ATLOP**, nous ajoutons les modèles **KD-DocRE** (Tan *et al.*, 2022a) et **TTM-RE** (Gao *et al.*, 2024), ainsi que **GLiREL** (Boylan *et al.*, 2025), une adaptation de GLiNER pour l'extraction de relations au niveau de la phrase (en utilisant les coréférences de la vérité terrain pour agréger les relations détectées au niveau phrastique).

Nous rapportons également les résultats des méthodes récentes s'appuyant sur les LLM, en comparant GLiDRE à la méthode introduite par Li *et al.* (2024b). Cette méthode utilise l'affinage de modèles Llama (Touvron *et al.*, 2023) via une adaptation de bas rang LoRA (Hu *et al.*, 2022), tout en améliorant

Modèle	Val F1	Val Ign F1	Test F1	Test Ign F1
Modèles ER-DOC				
ATLOP (Zhou *et al.*, 2021)	$76.15_{\pm 0.23}$	$75.88_{\pm 0.23}$	$77.81_{\pm 0.71}$	$76.13_{\pm 0.28}$
KD-DocRE (Tan *et al.*, 2022a)	$77.88_{\pm 0.42}$	$77.12_{\pm 0.49}$	$78.28_{\pm 0.72}$	$77.60_{\pm 0.25}$
TTM-RE (Gao *et al.*, 2024)	$78.13_{\pm 0.12}$	$78.05_{\pm 0.17}$	$79.95_{\pm 0.13}$	$78.20_{\pm 0.34}$
DREEAM (Ma *et al.*, 2023)	$79.42_{\pm 0.18}$	$78.36_{\pm 0.19}$	$80.20_{\pm 0.45}$	$78.56_{\pm 0.39}$
Méthode basée sur les LLM (Li et al.*, 2024b)*				
LMRC LLaMA2-13B-Chat	-	-	74.63	74.08
Modèle inspiré de GLiNER				
GLiREL (Boylan *et al.*, 2025)	-	-	54.13	53.24
GLiDRE	$77.64_{\pm 0.05}$	$76.54_{\pm 0.04}$	$77.80_{\pm 0.22}$	$76.73_{\pm 0.22}$

TABLE 3 – Performances sur le jeu de données Re-DocRED. Les scores F1 sont calculés sur les jeux de test et de validation. Les résultats reportés proviennent des articles correspondant excepté pour les modèles d'ER-DOC dont les résultats proviennent de (Gao *et al.*, 2024).

les performances à l'aide d'un classifieur chargé d'identifier des relations potentielles et de guider le LLM affiné. Nous ne rapportons que les meilleurs résultats.

Les résultats présentés dans le Tableau 3 montrent que GLiDRE atteint des scores honorables sur le jeu de données Re-DocRED alors que le modèle n'a pas été spécifiquement conçu pour être entraîné à grande échelle, ce qui souligne sa polyvalence et son adaptabilité à divers scénarios. Avec un score F1 de 77.8 sur l'ensemble de test, GLiDRE se place en compétition directe avec des modèles d'ER-DOC établis tels qu'ATLOP et KD-DocRE. Des méthodes comme DREEAM et TTM-RE obtiennent des scores plus importants, mais en intègrant des mécanismes additionnels, tels que l'utilisation de *preuves* pour DREEAM ou le module *Token Turing Machine* pour TTM-RE. Par ailleurs, GLiDRE dépasse l'approche LMRC, fondée sur les LLM, de plus de 3 points de F1, tout en étant nettement plus petit et plus économe en calculs, ce qui souligne l'avantage des architectures d'encodeurs spécialisés pour cette tâche par rapport à l'affinage de LLM généralistes. Enfin, la comparaison avec GLiREL souligne l'importance des adaptations architecturales de GLiDRE pour l'extraction au niveau du document.

4.2.4 Contexte *Zero-Shot*

Le cadre d'évaluation sans exemple (*zero-shot*) pour l'ER-DOC reste un domaine de recherche largement sous-exploré. Nous établissons une nouvelle référence en évaluant les performances de grands modèles de langage récents et performants. Notre méthodologie de *prompting* s'inspire de travaux récents sur l'extraction d'informations *zero-shot* (Yuan *et al.*, 2023) : nous annotons les mentions d'entités directement dans le texte d'entrée à l'aide de balises spéciales et interrogeons le modèle pour produire des triplets sous la forme `(tête, relation, queue)`.

Les résultats (Tableau 4) montrent que GLiDRE, avec ses 800 millions de paramètres, atteint des performances compétitives face à des LLM de taillle très supérieure. Avec un score F1 de $17,52$, GLiDRE surpasse Llama 3.3 70B, Qwen3-4B Instruct, Gemma3 27B Instruct et Qwen3 4B Thinking, tout en ne demandant qu'une fraction de leur coût d'inférence, montrant que, pour des tâches spéciali-

Modèles	Val F1	Val Ign F1	Test F1	Test Ign F1	Temps (s)
Grands Modèles de Language affinés en instruction					
Qwen3-4B	0.35	0.35	0.36	0.36	1200
Gemma3 27B	13.52	13.38	12.77	12.66	300
Llama 3.3 70B	15.67	15.57	15.81	15.71	1680
Qwen3-30B-A3B	17.29	16.92	17.87	17.52	4800
Mistral-Large 123B	18.49	18.33	18.61	18.50	1800
Grands Modèles de Language affinés en raisonnement					
Qwen3-4B	16.12	16.09	16.85	16.82	14400
Qwen3-30B-A3B	20.70	20.62	21.58	21.48	14400
Modèle étudié					
GLiDRE	17.47	17.08	17.52	17.20	40

TABLE 4 – Performance *zero-shot* sur les ensembles de test et de validation de Re-DocRED

sées comme l'ER-DOC, une architecture dédiée peut rivaliser avec les capacités de raisonnement de modèles massifs généralistes.

De plus, l'inférence de GLiDRE sur les 500 documents de test nécessite moins de 10 Go de mémoire vidéo et prend environ 40 secondes avec un seul GPU NVIDIA A100-80GB, alors que Mistral-Large (123B) demande une configuration distribuée d'au moins quatre GPU A100-80GB pour supporter son empreinte mémoire supérieure à 300 Go.Cette comparaison montre que GLiDRE est non seulement plus compact, mais aussi plus de 40 fois plus frugal en termes de puissance de calcul totale requise.

De plus, l'utilisation de LLM pour des tâches d'extraction nécessite une ingénierie complexe (prompts dédiés, inférence contrainte) pour produire des sorties structurées exploitables et limiter les hallucinations. À l'inverse, GLiDRE génère nativement des prédictions structurées. Cet avantage est illustré par nos tests avec Qwen3-4B-Instruct, qui produit fréquemment des sorties qui ne peuvent pas être décodées.

4.2.5 Analyses complémentaires

Pour compléter l'évaluation du modèle, nous avons effectué une analyse des impacts des différents choix architecturaux. L'analyse complète des résultats est détaillée en annexe A.3 mais nous résumons les éléments principaux de cette analyse ici :

— **LogSumExp vs. Mean Pooling** : plusieurs modèles d'ER-DOC ont montré une amélioration par l'utilisation de la fonction *LogSumExp* pour agréger les représentations des mentions d'entités : une comparaison des deux approches a montré qu'une simple moyenne donnait des meilleurs résultats pour le modèle GLiDRE ;

— **Pré-entraînement et contexte local** : une étude d'ablation détaillée a montré que, même en contexte entièrement supervisé, le pré-entraînement apporte un gain et que l'intégration du module de contexte local (LOP) permet également d'améliorer les résultats ;

— **Adaptation de la fonction de perte à seuil adaptatif** : l'ajout d'un seuil adaptatif, introduit dans le modèle ATLOP pour apprendre un seuil de décision dynamique par relation, n'apporte pas d'amélioration pour GLiDRE.

4.2.6 Limites

GLiDRE, malgré son efficacité, présente plusieurs limites. Le nombre de paires de relations candidates croît de manière quadratique avec le nombre d'entités, augmentant la consommation de mémoire. L'architecture bi-encodeur est limitée par la taille de contexte maximale de l'encodeur de document (ex. 512 tokens pour DeBERTa), mais cette restriction est partagée par les modèles DREEAM, ATLOP, KD-DocRE et TTM-RE et n'affecte pas les documents de Re-DocRED. Dans le cas général, les documents dépassant cette limite peuvent être tronqués ou segmentés, par exemple avec une fenêtre glissante, ce qui risque néanmoins de briser certaines dépendances inter-phrastiques lointaines. Une autre solution consiste simplement à utiliser un modèle encodeur ayant un contexte plus large, comme BGE-Large-V1.5 (8k tokens).

Le calcul du score de similarité indépendant pour chaque paire entité-relation dans GLiDRE ne permet pas de considérer les interactions entre les étiquettes. Une première analyse qualitative des résultats a montré que la plupart des erreurs vient d'une difficulté à distinguer des types de relations sémantiquement proches, ce qui pourrait résulter de cette architecture. Une architecture de type *cross-encoder* permettrait de prendre en compte ces dépendances mais ne serait dans la pratique pas traitable lorsque les types de relations sont trop nombreux.

L'évaluation réalisée suppose que les entités et les chaînes de coréférence sont fournies, mais GLiDRE n'effectue ni la reconnaissance d'entités nommées ni la résolution de coréférences. Son efficacité dans une chaîne de traitement complet dépend donc de la précision des modules en amont. Cette limite étant intrinsèque à la tâche considérée, elle est partagée par les autres modèles de référence auxquels nous nous comparons.

Le pré-entraînement de GLiDRE repose sur l'annotation synthétique de textes par un LLM, ce qui comporte un risque de propagation des biais. Ce risque est atténué par des directives d'annotation claires et pourrait être limité en combinant plusieurs prompts et modèles, ou en faisant réviser un sous-ensemble d'annotations par des réviseurs humains.

5 Conclusion

Nous proposons dans cet article une évaluation de GLiDRE, un nouveau modèle bi-encodeur léger pour l'extraction de relations documentaires qui reconceptualise l'ER-DOC comme un problème de mesure de similarité entre plongements lexicaux. Nous démontrons, à travers des expérimentations approfondies sur les jeux de données Re-DocRED, FREDo et Re-FREDo, que GLiDRE atteint des performances en *few-shot* épisodique qui rivalisent avec l'état de l'art et égale ou surpasse des LLM bien plus massifs en configuration *zero-shot*, tout en ne consommant qu'une fraction de leur coût calculatoire. Dans des contextes avec peu de ressources, GLiDRE rivalise des modèles entièrement supervisés et, de plus, il fournit des prédictions structurées sans nécessiter de *prompting* complexe ou d'inférence contrainte. Son efficacité en fait un choix logique pour les applications réelles d'extraction d'informations.

Des améliorations du modèle pourront explorer la génération synthétique de relations pour combler davantage les écarts de performance restants et étudieront des méthodes de seuil adaptatifs adaptées à la configuration bi-encodeur afin de renforcer la robustesse du modèle face des relations éloignées de son domaine de pré-entraînement.

Références

ARMINGAUD R. (2025). GLiDRE : Modèle généraliste pour l'extraction de relations à l'échelle de documents. In *Actes de l'atelier TextMine 2025*, Strasbourg, France. HAL : hal-04918406.

BOYLAN J., HOKAMP C. & GHALANDARI D. G. (2025). GLiREL - generalist model for zero-shot relation extraction. In *NAACL*, p. 8230–8245.

CHENG Q., LIU J., QU X., ZHAO J., LIANG J., WANG Z., HUAI B., YUAN N. J. & XIAO Y. (2021). HacRED : A large-scale relation extraction dataset toward hard cases in practical applications. In *ACL Findings*, p. 2819–2831.

GAO C., WANG X. & SUN J. (2024). TTM-RE : Memory-augmented document-level relation extraction. In *ACL*, p. 443–458.

HE P., LIU X., GAO J. & CHEN W. (2021). Deberta : Decoding-enhanced bert with disentangled attention. In *ICLR*.

HU E. J., SHEN Y., WALLIS P., ALLEN-ZHU Z., LI Y., WANG S., WANG L., CHEN W. *et al.* (2022). Lora : Low-rank adaptation of large language models. *ICLR*, **1**(2), 3.

JIA R., WONG C. & POON H. (2019). Document-level n-ary relation extraction with multiscale representation learning. In *NAACL*, p. 3693–3704.

LAN Y., LI D., ZHANG Y., ZHAO H. & ZHAO G. (2023). Modeling zero-shot relation classification as a multiple-choice problem. In *IJCNN*, p. 1–8.

LI G., WANG P., LIU J., GUO Y., JI K., SHANG Z. & XU Z. (2024a). Meta in-context learning makes large language models better zero and few-shot relation extractors. In *IJCAI*, p. 6350–6358.

LI J., JIA Z. & ZHENG Z. (2023). Semi-automatic data enhancement for document-level relation extraction with distant supervision from large language models. In *EMNLP*, p. 5495–5505.

LI X., CHEN K., LONG Y. & ZHANG M. (2024b). Llm with relation classifier for document-level relation extraction. *arXiv preprint arXiv :2408.13889*.

LIN T.-Y., GOYAL P., GIRSHICK R., HE K. & DOLLÁR P. (2017). Focal loss for dense object detection. In *ICCV*, p. 2980–2988.

LIU Y., ZHU Z., ZHANG X., FENG Z., CHEN D. & LI Y. (2023). Document-level relationship extraction by bidirectional constraints of beta rules. In *EMNLP*, p. 2256–2266.

LUAN Y., HE L., OSTENDORF M. & HAJISHIRZI H. (2018). Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. In *EMNLP*, p. 3219–3232.

MA Y., WANG A. & OKAZAKI N. (2023). DREEAM : Guiding attention with evidence for improving document-level relation extraction. In *EACL*, p. 1971–1983.

MENG S., HU X., LIU A., LI S., MA F., YANG Y. & WEN L. (2023). RAPL : A relation-aware prototype learning approach for few-shot document-level relation extraction. In *EMNLP*, p. 5208–5226.

PAN D., SUN Y., XU B., LI J., YANG Z., LUO L., LIN H. & WANG J. (2025). Prototype tuning : A meta-learning approach for few-shot document-level relation extraction with large language models. In *NAACL findings*, p. 1112–1128.

PENEDO G., KYDLÍČEK H., LOZHKOV A., MITCHELL M., RAFFEL C. A., VON WERRA L., WOLF T. *et al.* (2024). The fineweb datasets : Decanting the web for the finest text data at scale. *Advances in Neural Information Processing Systems*, **37**, 30811–30849.

POPOVIC N. & FÄRBER M. (2022). Few-shot document-level relation extraction. In M. CARPUAT, M.-C. DE MARNEFFE & I. V. MEZA RUIZ, Éds., *NAACL*, p. 5733–5746.

SABO O., ELAZAR Y., GOLDBERG Y. & DAGAN I. (2021). Revisiting few-shot relation classification : Evaluation data and classification schemes. *Transactions of the Association for Computational Linguistics*, **9**, 691–706.

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 *ICLR*.

TAN Q., HE R., BING L. & NG H. T. (2022a). Document-level relation extraction with adaptive focal loss and knowledge distillation. In *ACL Findings*, p. 1672–1681.

TAN Q., XU L., BING L., NG H. T. & ALJUNIED S. M. (2022b). Revisiting DocRED - addressing the false negative problem in relation extraction. In Y. GOLDBERG, Z. KOZAREVA & Y. ZHANG, Éds., *EMNLP*.

TAN Q., XU L., BING L., NG H. T. & ALJUNIED S. M. (2022c). Revisiting DocRED - addressing the false negative problem in relation extraction. In *EMNLP*, p. 8472–8487.

TOUVRON H., LAVRIL T., IZACARD G., MARTINET X., LACHAUX M.-A., LACROIX T., ROZIÈRE B., GOYAL N., HAMBRO E., AZHAR F. *et al.* (2023). Llama : Open and efficient foundation language models. *arXiv preprint arXiv :2302.13971*.

WANG X., ZHOU W., ZU C., XIA H., CHEN T., ZHANG Y., ZHENG R., YE J., ZHANG Q., GUI T. *et al.* (2023). Instructuie : Multi-task instruction tuning for unified information extraction. *arXiv preprint arXiv :2304.08085*.

WEI X., CUI X., CHENG N., WANG X., ZHANG X., HUANG S., XIE P., XU J., CHEN Y., ZHANG M. *et al.* (2023). Zero-shot information extraction via chatting with chatgpt. *arXiv preprint arXiv :2302.10205*.

XIAO S., LIU Z., ZHANG P. & MUENNIGHOFF N. (2023). C-pack : Packaged resources to advance general chinese embedding.

XUE L., ZHANG D., DONG Y. & TANG J. (2024). AutoRE : Document-level relation extraction with large language models. In *ACL*, p. 211–220.

YAO Y., YE D., LI P., HAN X., LIN Y., LIU Z., LIU Z., HUANG L., ZHOU J. & SUN M. (2019). DocRED : A large-scale document-level relation extraction dataset. In *ACL*, p. 764–777.

YUAN C., XIE Q. & ANANIADOU S. (2023). Zero-shot temporal relation extraction with chatgpt. *arXiv preprint arXiv :2304.05454*.

ZARATIANA U., TOMEH N., HOLAT P. & CHARNOIS T. (2024). GLiNER : Generalist model for named entity recognition using bidirectional transformer. In *NAACL*, p. 5364–5376.

ZHANG K., WU P., YU B., WU K., ZHENG A., HUANG X., ZHU C., PENG M., ZAN H. & SONG Y. (2025). CaDRL : Document-level relation extraction via context-aware differentiable rule learning. In *COLING*.

ZHANG Y. & KANG Z. (2024). Tpn : Transferable proto-learning network towards few-shot document-level relation extraction. In *IJCNN*, p. 1–9.

ZHOU W., HUANG K., MA T. & HUANG J. (2021). Document-level relation extraction with adaptive thresholding and localized context pooling. In *AAAI*.

ZHOU W., ZHANG S., GU Y., CHEN M. & POON H. (2023). Universalner : Targeted distillation from large language models for open named entity recognition. *arXiv preprint arXiv :2308.03279*.

A Annexes

A.1 Détails d'implémentation

Toutes les expérimentations sont réalisées sur un GPU NVIDIA H100 avec une taille de *batch* de 16. Pour la phase de pré-entraînement, le modèle est entraîné durant 50 000 étapes. Pour l'affinage, le modèle est entraîné pendant 10 000 étapes. Nous utilisons deux taux d'apprentissage : 1×10^{-5} pour les encodeurs du modèle et 1×10^{-4} pour toutes les autres couches. Pour les expériences dans un contexte avec peu de ressources et entièrement supervisées, nous rapportons la moyenne et l'écart-type sur 5 expériences différentes. La version du modèle obtenant la meilleure performance sur l'ensemble de développement, ainsi qu'un seuil de décision de 0.5, sont utilisés pour l'évaluation sur l'ensemble de test. Le modèle utilisé est fondé sur deux encodeurs pré-entraînés sur l'anglais : DebertaV3 Large (He *et al.*, 2021) pour l'encodeur de documents et BGE-Large V1.5 (Xiao *et al.*, 2023) pour l'encodeur d'étiquettes, conformément à la version bi-encodeur de GLiNER. Le modèle final comprend environ 800 millions de paramètres. La phase de pré-entraînement dure environ 24 heures, tandis que l'affinage sur le jeu de données Re-DocRED nécessite 3,5 heures. De plus, l'évaluation du modèle sur le jeu de données Re-FReDo est particulièrement exigeante en calcul en raison du grand nombres d'épisodes, qui impose des entraînements répétés et la ré-initialisation des poids du modèle pour chaque épisode. Ce processus représente environ 35 heures de calcul pour la tâche intra-domaine et 5 heures pour la tâche hors-domaine.

A.2 Détails des modèles utilisés dans les expériences

Nous détaillons dans cette section les modèles de l'état de l'art auxquels nous nous comparons dans nos expériences :

Modèles évalués en contexte de données limitées

DREEAM (Ma *et al.*, 2023) représente l'état de l'art sur le jeu de données dans le scénario entièrement supervisé. Ce modèle s'appuie sur le guidage de l'attention via les preuves utilisées comme signaux d'apprentissage, et peut employer une stratégie d'auto-entraînement pour apprendre l'extraction de preuves sans annotations explicites.

ATLOP (Zhou *et al.*, 2021) formule l'ER-DOC comme une tâche de segmentation sémantique, introduit le module de contexte local pour capturer des informations locales grâce à l'attention, ainsi qu'une fonction de perte à seuil adaptatif pour apprendre des seuils de décision dynamiques. Ce modèle est plus comparable à GLiDRE car il ne nécessite pas d'annotations sur les preuves.

Modèles évalués en contexte few-shot épisodique

DL-MNAV (Popovic & Färber, 2022) adapte la méthode MNAV (Sabo *et al.*, 2021), initialement conçue pour l'extraction de relations au niveau de la phrase, aux documents en effectuant l'agrégation des représentations de mentions et en modélisant explicitement les cas ou aucun des types de relations ne correspond (NOTA pour *none-of-the-above*) via des vecteurs NOTA appris, une fonction de perte à seuil adaptatif et un échantillonnage des exemples NOTA s'appuyant sur le support lors de l'inférence.

RAPL (Meng *et al.*, 2023) affine les prototypes par l'agrégation au niveau des instances et un apprentissage contrastif pondéré par les relations, tout en construisant des prototypes NOTA spécifiques à la tâche.

TPN (Zhang & Kang, 2024) améliore le transfert inter-domaines grâce à un encodeur hybride, un apprentissage de prototypes NOTA transférables et un module de calibration pour atténuer le biais envers les NOTA.

Prototype Tuning (Pan *et al.*, 2025) est une approche de méta-apprentissage pour les LLM qui intègre directement les prototypes de relations dans l'affinage.

Modèles évalués en contexte supervisé

KD-DocRE (Tan *et al.*, 2022a) exploite un module d'attention axiale pour modéliser l'interdépendance des paires d'entités entre les phrases, utilise une fonction de perte focale adaptative pour atténuer le déséquilibre entre les classes et distille de la connaissance de données annotées par supervision distante.

TTM-RE (Gao *et al.*, 2024) introduit un module de mémoire fondé sur les machines de Turing (*Token Turing Machine*) qui enrichit les représentations des documents avec des *tokens* de mémoire externe, associée à une fonction de perte robuste au bruit, spécifiquement conçue pour le cas des exemples positifs non étiquetés.

A.3 Analyses complémentaires

A.3.1 LogSumExp vs. Mean Pooling

Comme indiqué dans la section 3.2, l'agrégation des représentations des différentes mentions est effectuée par moyennage. D'autres stratégies d'agrégation peuvent être utilisées, notamment la fonction LogSumExp (LSE), appliquée avec succès à l'ER-DOC par (Jia *et al.*, 2019) puis adoptée par ATLOP, où elle a montré des performances légèrement supérieures à une moyenne conventionnelle. La fonction LSE est une approximation lissée de l'opérateur maximum et est définie comme suit :

$$\mathrm{LSE}(\mathbf{x_1}, ..., \mathbf{x_n}) = \log\left(\sum_{i=1}^{n} \exp(x_i)\right)$$

Cependant nos résultats présentés dans le Tableau 5, montrent un comportement différent au sein de l'architecture étudiée. Pour GLiDRE, la moyenne classique surpasse la fonction LSE de près d'un point de F1 sur l'ensemble de test. Nous attribuons cet écart aux différences architecturales : contrairement au classifieur bilinéaire d'ATLOP, GLiDRE emploie un cadre bi-encodeur qui compare les plongements lexicaux des relations et des étiquettes de relations. Cette divergence structurelle suggère que les optimisations ne sont pas toujours directement transférables d'un modèle à l'autre. Étant donné que l'article original d'ATLOP ne rapportait que des gains mineurs avec la fonction LSE, nos conclusions confirment que la moyenne est un choix plus efficace et robuste pour ce modèle. Pour économiser les ressources de calcul, cette expérience a été menée avec la variante du modèle sans pré-entraînement.

Méthode d'agrégation	Test F1	Test Ign F1
LogSumExp	$76.34_{\pm 0.16}$	$74.98_{\pm 0.17}$
Mean	$\mathbf{77.15}_{\pm 0.42}$	$\mathbf{75.96}_{\pm 0.49}$

TABLE 5 – Comparaison des scores F1 sur l'ensemble de test de Re-DocRED en utilisant la moyenne et la fonction LogSumExp pour l'agrégation des plongements lexicaux.

A.3.2 Effets du pré-entraînement et du module de contexte local

Nous menons une étude d'ablation dans le cadre entièrement supervisé pour isoler l'influence de la phase de pré-entraînement synthétique et du mécanisme de contexte local (noté LOP dans le modèle ATLOP). Deux variantes du modèle sont évaluées : l'une affinée sans pré-entraînement et l'autre sans le module LOP.

Les résultats du Tableau 6 confirment que les deux composants contribuent positivement à la performance finale de GLiDRE. Le retrait de la phase de pré-entraînement entraîne la baisse de performance la plus significative, avec une chute de près de $0,5$ point de F1, montrant l'efficacité de la génération de données synthétiques pour le pré-entraînement du modèle. La désactivation du LOP entraîne une baisse supplémentaire de $0,2$ point de F1, ce qui valide son rôle dans l'amélioration des représentations des relations. Ces résultats justifient l'inclusion des deux techniques dans l'architecture finale.

Configuration	Test F1	Test Ign F1
GLiDRE	$\mathbf{77.64}_{\pm 0.05}$	$\mathbf{76.54}_{\pm 0.04}$
sans pré-entraînement	$77.15_{\pm 0.42}$	$75.96_{\pm 0.49}$
sans pré-entraînement et LOP	$76.94_{\pm 0.19}$	$76.01_{\pm 0.15}$

TABLE 6 – Etude d'ablation sur l'ensemble de test de Re-DocRED. Nous reportons les scores F1 après avoir enlevé la phase de pré-entraînement et le module de contexte local.

A.3.3 Adaptation de la fonction de perte à seuil adaptatif

ATLOP a introduit un mécanisme de seuil adaptatif pour apprendre un seuil de décision dynamique par relation, évitant ainsi un seuil global fixe sous-optimal. Il est implémenté en ajoutant une classe spéciale au classifieur, notée `TH` et une relation n'est prédite que si la confiance du modèle sur cette relation dépasse celui de la classe `TH`.

L'adaptation de ce mécanisme à GLiDRE n'est pas triviale en raison de l'architecture bi-encodeur, qui ne possède pas de tête de classification fixe. Pour reproduire ce mécanisme, un petit perceptron multicouche est utilisé pour imiter ce seuil adaptatif. L'entrée du réseau est la concaténation du plongement lexical de la relation candidate et d'un vecteur de contexte global, formé par la moyenne de tous les plongements des types de relations possibles. Le modèle est ensuite entraîné en utilisant la fonction de perte originale d'ATLOP.

Pour économiser des ressources de calcul, cette expérience est menée sur la variante du modèle sans pré-entraînement. Comme le montre le Tableau 7, cette implémentation dégrade les performances

d'environ $1,4$ point de score F1. Nous émettons plusieurs hypothèses pour expliquer ce résultat négatif :

(1) la méthode de calcul de seuil via un MLP séparé est fondamentalement différente de l'approche par classifieur intégré d'ATLOP

(2) la fonction de perte ATLOP pourrait être moins efficace pour gérer le sévère déséquilibre de répartition entre les classes dans DocRED par rapport à la fonction de perte focale déjà utilisée dans le modèle principal.

Configuration	**Test F1**	**Test Ign F1**
Focal Loss	$\mathbf{77.15}_{\pm 0.42}$	$\mathbf{75.96}_{\pm 0.49}$
Adaptive Threshold	$75.74_{\pm 1.05}$	$74.84_{\pm 1.05}$

TABLE 7 – Comparaison entre l'entraînement standard avec fonction de perte focale et l'adaptation de la méthode de seuil adaptatif d'ATLOP. Les expériences sont menées sans la phase de pré-entraînement.