Evaluation of GLiNER for Zero-Shot Named Entity Recognition on French

Anna MOSAKI
ENSAE Paris, Section FRD
anna.mosaki@ensae.fr

Abstract

Named Entity Recognition (NER) is a fundamental task in Natural Language Processing (NLP). Traditional supervised models often require costly retraining to adapt to new entity types, while Large Language Models (LLMs) offering open-domain flexibility are computationally expensive. GLiNER (Generalist Model for Named Entity Recognition using Bidirectional Transformer) provides a compact, efficient alternative using bidirectional encoders to match entity types and text spans. This report evaluates the zero-shot performance of pre-trained GLiNER models on a French NER task using the `CATIE-AQ/frenchNER_4entities` dataset. We analyze the characteristics of a custom train/test split derived from this dataset and present evaluation results for different GLiNER variants (`base`, `large`, `multi-v2.1`) on a subset (N=2000) of this custom test split. Our findings indicate that the multilingual GLiNER model achieves promising zero-shot results on French (0.525 F1 overall on the evaluated subset), with strong performance for common entity types like 'person' (0.68 F1) and 'location' (0.77 F1). However, performance varies significantly across types, with 'organization' proving challenging (0.40 F1) and 'misc' failing almost entirely (0.004 F1). While the overall French score on this subset appears higher than the multilingual average reported in the original GLiNER paper, we caution against direct comparison due to dataset and evaluation setup differences. We conclude that zero-shot GLiNER shows considerable cross-lingual potential for French, but supervised fine-tuning would be essential for robust, comprehensive performance, particularly for less common or ambiguous entity types.

1 Introduction

Named Entity Recognition (NER), the task of identifying and classifying named entities such as persons, organizations, and locations within text, is a cornerstone of Natural Language Processing (NLP). It serves as a foundational component for numerous downstream applications, including information extraction, question answering, and semantic search.

Historically, supervised NER models, evolving from rule-based systems to statistical methods like CRFs (Lafferty et al., 2001) and later to deep learning architectures like BiLSTM-CRFs (Lample et al., 2016; Huang et al., 2015) and transformer-based sequence labelers (Devlin et al., 2019), achieved high accuracy. However, a major limitation remains their reliance on a predefined set of entity categories established during training. Adapting these models to new domains or entity types necessitates significant annotation effort and complete retraining, hindering flexibility and scalability (Zaratiana et al., 2024).

The rise of Large Language Models (LLMs) like GPT-3 (Brown et al., 2020) introduced a paradigm shift towards "Open NER." LLMs can extract entities of arbitrary types specified through natural

Preprint. Under review.

language prompts, offering unprecedented adaptability (Wang et al., 2023; Zhou et al., 2023). Nevertheless, the immense size and computational requirements of state-of-the-art LLMs make them impractical for many resource-constrained scenarios. Furthermore, reliance on API-based access (e.g., to GPT-4 (OpenAI, 2023)) can lead to significant operational costs when deployed at scale.

Addressing this gap, Zaratiana et al. (2024) proposed GLiNER (Generalist Model for Named Entity Recognition using Bidirectional Transformer). GLiNER represents a compact and efficient alternative, designed to identify any entity type by reformulating NER. Instead of sequence labeling or generation, it frames the task as matching learned entity type embeddings against text span representations derived from a bidirectional transformer encoder (like BERT or DeBERTa (He et al., 2021)). This architecture leverages the rich contextual understanding of bidirectional models while enabling parallel entity extraction, making it faster than autoregressive LLMs. The original study demonstrated GLiNER's potent zero-shot performance on English benchmarks and provided initial multilingual evaluations.

This report extends the evaluation of GLiNER to French, assessing the zero-shot transfer capabilities of its pre-trained models on the `CATIE-AQ/frenchNER_4entities` dataset. Specifically, we aim to:

1) Analyze the statistical properties of a custom train/test split derived from the target French dataset.
2) Evaluate the zero-shot performance (Precision, Recall, F1-score) of publicly available GLiNER checkpoints (English-based and multilingual) on a defined subset of the French test data.
3) Analyze the performance variations across different models and entity types within the French context.
4) Compare these French zero-shot results, with appropriate caveats, to the multilingual performance reported in the original GLiNER paper.
5) Discuss the implications for practical application and the potential benefits of subsequent supervised fine-tuning.

Our investigation seeks to provide insights into GLiNER's effectiveness and limitations when applied zero-shot to French NER on the specific data and evaluation setup employed here.

2 Related Work

The landscape of NER research has seen considerable evolution, moving from rule-based methods to sophisticated deep learning models.

Traditional NER Approaches: Early systems often depended on handcrafted rules, gazetteers, and linguistic features. These methods, while effective in specific domains, struggled with scalability and robustness. Statistical machine learning models, such as Hidden Markov Models (HMMs) and Conditional Random Fields (CRFs) (Lafferty et al., 2001), offered improvements by learning from annotated data, typically framing NER as a sequence labeling task using schemes like BILOU (Ratinov and Roth, 2009).

Deep Learning for NER: Deep learning significantly advanced NER capabilities. Bidirectional Long Short-Term Memory networks coupled with a CRF layer (BiLSTM-CRF) (Lample et al., 2016; Huang et al., 2015) became a dominant architecture, effectively modeling sequential dependencies. Enhancements like character-level embeddings (Lample et al., 2016) and contextual string embeddings (e.g., Flair (Akbik et al., 2018)) further improved performance. The introduction of large pre-trained transformers like BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and DeBERTa (He et al., 2021) provided powerful contextual representations, leading to new state-of-the-art results when fine-tuned for sequence labeling (Li et al., 2021a).

Span-Based NER: An alternative approach treats NER as classifying candidate text spans (Shen et al., 2018; Fu et al., 2021). Models identify potential spans and assign an entity type (or none) to each. This paradigm naturally handles nested entities (Li et al., 2021b) and forms the basis for GLiNER's span-matching mechanism. Related methods include work by Zaratiana et al. (2022) and the MRC framework.

Open-Domain and Zero-Shot NER: To overcome the limitations of fixed entity sets, research explored open-domain NER. LLMs (Brown et al., 2020) have been pivotal, enabling zero-shot NER by specifying target types via prompts (Agrawal et al., 2022; Ashok and Lipton, 2023). Instruction-tuning, where LLMs are fine-tuned on diverse Information Extraction (IE) tasks formatted as instructions, further enhanced zero-shot capabilities, as seen in models like `InstructUIE` (Wang et al., 2023) and `GoLLIE` (Sainz et al., 2023) (see also Mishra et al., 2021). `UniNER` (Zhou et al., 2023) demonstrated that distilling knowledge from ChatGPT into a fine-tuned LLaMa model could yield exceptional zero-shot NER performance. However, these powerful LLM approaches generally entail high computational costs and inference latency (Zaratiana et al., 2024).

Compact Generalist NER Models: Models like GLiNER (Zaratiana et al., 2024) and `USM` (Lou et al., 2023) emerged to offer generalist NER capabilities with greater efficiency. They employ smaller bidirectional transformer backbones. GLiNER, the focus here, uses its BiLM backbone to compute representations for both entity type prompts and text spans, then matches them in latent space using dot-product similarity. This allows for parallel extraction and avoids autoregressive generation bottlenecks.

Multilingual NER: Applying NER across languages involves challenges like linguistic diversity and resource availability. Multilingual transformers (mBERT, XLM-R (Conneau et al., 2019), mDeBERTa-V3 (He et al., 2021)) provide foundations for cross-lingual transfer. Evaluating zero-shot cross-lingual performance, as done in this report for GLiNER on French, assesses a model's ability to generalize from its primary training language (mostly English) to a new target language without specific fine-tuning.

This work evaluates GLiNER, a compact generalist model, specifically focusing on its zero-shot cross-lingual transfer performance from its English-centric `Pile-NER` pre-training data to French.

3 Methodology

This section details the GLiNER model architecture, the dataset used, and the experimental procedure for evaluating its zero-shot performance on French.

3.1 GLiNER Model Architecture

We employ the GLiNER model architecture as described by Zaratiana et al. (2024). The model is designed to identify entities corresponding to arbitrary types provided at inference time. Key components include:

i) **Input Formatting:** The model takes a unified input sequence. This sequence begins with the target entity types (e.g., 'person', 'location'), each preceded by a special learned token `[ENT]`. These prompts are followed by a separator token `[SEP]`, and then the input text itself. Example: `[ENT] person [ENT] location [SEP] Marie Curie travaille à Paris .`

ii) **Bidirectional Encoder:** A pre-trained Bidirectional Language Model (BiLM), such as DeBERTa-V3 (He et al., 2021), processes the entire input sequence, generating contextualized embeddings for all tokens, including the `[ENT]` tokens and the text tokens.

iii) **Entity and Span Representations:**

- *Entity Type Embeddings (q_t):* The embeddings corresponding to the `[ENT]` tokens are passed through a FeedForward Network (FFN) to derive refined embeddings representing the target entity types.
- *Span Embeddings (S_{ij}):* For candidate text spans starting at token i and ending at token j (up to a maximum length K, typically 12), a span representation is computed. This is done by concatenating the BiLM's output embeddings for the start and end tokens (h_i, h_j) and passing them through another FFN: $S_{ij} = \text{FFN}(h_i \otimes h_j)$.

iv) **Matching Score (ϕ):** The core mechanism involves calculating a compatibility score between each entity type embedding (q_t) and each span embedding (S_{ij}). This is computed via a dot product followed by a sigmoid activation: $\phi(i, j, t) = \sigma(S_{ij}^T q_t)$. The score $\phi(i, j, t)$ represents the model's estimated probability that span (i, j) corresponds to entity type t.

v) **Decoding:** At inference time, a decoding algorithm processes these scores to select the final set of predicted entities. This typically involves applying a probability threshold (e.g., 0.5) and using a greedy strategy to handle overlapping spans, allowing for either flat or nested NER extraction depending on the configuration.

The model's training objective is to maximize the matching score for correct span-type pairs (positive examples) and minimize it for incorrect pairs (negative examples) using a binary cross-entropy loss, often incorporating techniques like negative entity type sampling during training (Zaratiana et al., 2024).

3.2 Dataset and Preprocessing

Dataset: We use the French NER dataset `CATIE-AQ/frenchNER_4entities` from the Hugging Face Hub[1]. It contains tokenized sentences annotated for four entity types: `PER` (Person), `ORG` (Organization), `LOC` (Location), and `MISC` (Miscellaneous), plus the `O` tag for non-entity tokens.

Data Split for Evaluation: The official dataset provides predefined train, validation, and test splits. However, for the experiments conducted in the accompanying notebook and reported here, a different setup was used: the original *train* split (328,757 examples) was shuffled (seed=42) and then re-split into a new training set (80%, 263,005 examples) and a new test set (20%, 65,752 examples). **Crucially, our zero-shot evaluation was performed only on the first 2000 examples of this custom test split**. This is a significant limitation, as the results may not generalize to the full official test set or other French NER data. The analysis presented in Section 4 pertains to the custom training split derived through this process.

Label Mapping: Since GLiNER requires natural language labels, we map the dataset tags to English labels for querying the model:

- `PER` → `person`
- `ORG` → `organization`
- `LOC` → `location`
- `MISC` → `misc`

These mapped labels (`person`, `organization`, etc.) were provided as the target entity types to GLiNER during inference.

3.3 Experimental Setup

Task: The core experiment is **zero-shot evaluation**. We apply pre-trained GLiNER models directly to the French test subset without any fine-tuning on French data. The models evaluated were primarily pre-trained on the English `Pile-NER` dataset (Zhou et al., 2023), allowing us to assess their cross-lingual transfer capability.

Models: We evaluate three checkpoints available from the Hugging Face Hub:

- `urchade/gliner_base`: DeBERTa-V3-Base backbone.
- `urchade/gliner_large`: DeBERTa-V3-Large backbone.
- `urchade/gliner_multi-v2.1`: mDeBERTa-V3-Base (multilingual) backbone.

Evaluation Metrics: We report standard micro-averaged Precision, Recall, and F1-score. A prediction is considered correct only if both the entity span boundaries and the entity type exactly match a ground truth entity (after mapping the ground truth label).

Implementation: We used the official `gliner` Python library[2] (version providing `GLiNER.from_pretrained`) and the Hugging Face `datasets` library (Lhoest et al., 2021). Model predictions were generated using `model.predict_entities()` with the mapped French labels (`person`, `organization`, `location`, `misc`) and a probability threshold of 0.5. True entities were extracted from the dataset's tag sequences.

[1] https://huggingface.co/datasets/CATIE-AQ/frenchNER_4entities

[2] https://github.com/urchade/GLiNER

4 Data Analysis

This section presents a statistical overview of the **custom training split** (N=263,005) derived from `CATIE-AQ/frenchNER_4entities`, which informed our understanding of the data distribution the models might implicitly generalize from (though they were not trained on it).

Sentence Length: The distribution of sentence lengths, measured in tokens, is depicted in Figure 1. The average sentence length in this training split is **27.57 tokens**. The histogram shows that while a significant number of sentences are relatively short (peaking around 15-30 tokens), there is a long tail of sentences with considerably more tokens.

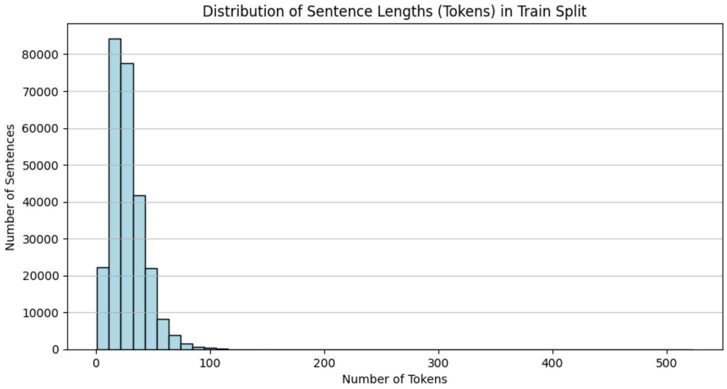

Figure 1: Distribution of sentence lengths (tokens) in the custom training split (N=263,005) derived from `CATIE-AQ/frenchNER_4entities`.

Entity Type Distribution: Figure 2 illustrates the frequency of annotated entity types within this custom training split. The counts are as follows:

- `MISC`: 208,469 instances
- `LOC`: 151,163 instances
- `PER`: 135,630 instances
- `ORG`: 42,031 instances

A notable characteristic of this specific split is the high prevalence of the `MISC` category, making it the most frequent entity type. `LOC` and `PER` are also very common, while `ORG` appears considerably less often. This distribution, particularly the dominance and potential heterogeneity of `MISC`, might influence the model's ability to generalize, especially in a zero-shot setting.

Most Frequent Entities: An examination of the most frequent entity texts (after stripping whitespace and excluding simple punctuation) revealed common geographical locations (`France`, `Paris`, `Europe`), mentions of countries/organizations (`États-Unis`), and terms likely originating from specific source domains within the dataset (`amphibiens`, `Female`, `Male`).

In summary, the custom training data split used for context analysis is characterized by relatively short average sentence length, a unique entity distribution dominated by `MISC`, followed by `LOC` and `PER`, with `ORG` being less frequent.

5 Results and Discussion

This section presents the results of the zero-shot evaluation of the three GLiNER models on the first **2000 examples** of the custom French test split derived from `CATIE-AQ/frenchNER_4entities`.

5.1 Overall Zero-Shot Performance

Table 1 summarizes the overall micro-averaged Precision, Recall, and F1-score achieved by each model on this specific French subset.

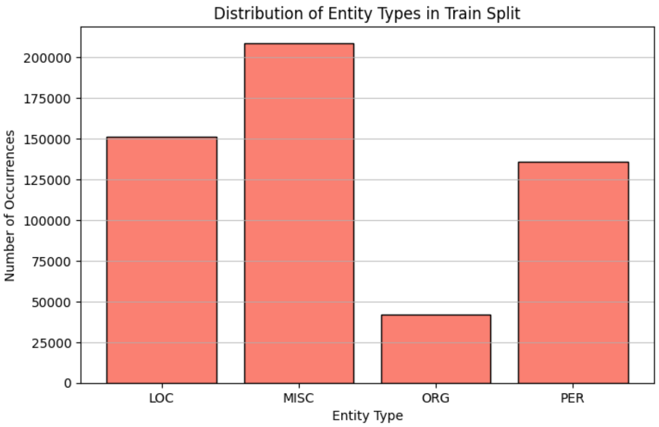

Figure 2: Distribution of annotated entity types in the custom training split (N=263,005). Note the high frequency of `MISC`.

Table 1: Overall Zero-Shot Performance on a Subset (N=2000) of the Custom French NER Test Split.

Model	Precision	Recall	F1-Score
`urchade/gliner_base`	0.4333	0.5257	0.4750
`urchade/gliner_large`	0.4493	**0.5338**	0.4879
`urchade/gliner_multi-v2.1`	**0.5481**	0.5041	**0.5251**

From these overall results on the limited test subset, we observe:

- **Effective Zero-Shot Transfer:** All models demonstrate a capacity for zero-shot NER on French, achieving F1 scores ranging from 0.475 to 0.525. This indicates that the underlying matching mechanism learned primarily from English (`Pile-NER`) transfers reasonably well to French syntax and entities.
- **Superior Multilingual Model:** The `gliner_multi-v2.1` model yields the highest F1-score (0.525). Its main advantage lies in significantly better precision (0.548) compared to the English-only models, even though its recall is slightly lower than the `large` English model. This aligns with expectations, as the multilingual mDeBERTa-V3 backbone is better equipped for cross-lingual understanding.
- **Impact of English Model Size:** Comparing the English-based models, the `large` version performs slightly better than the `base` version (0.488 vs 0.475 F1), primarily due to higher recall. However, the multilingual model surpasses both, suggesting the multilingual pre-training is more beneficial for this cross-lingual task than simply increasing the size of the English model.

5.2 Performance per Entity Type

To understand the nuances behind the overall scores, Table 2 provides the per-class performance metrics for the best-performing model in this evaluation, `gliner_multi-v2.1`. Figure 3 visualizes the per-class F1 scores.

The breakdown reveals significant disparities in performance across entity types:

- **Location (`LOC`):** This category shows the strongest performance with an F1-score of 0.767. The model achieves both high precision (0.719) and high recall (0.823), indicating reliable identification of locations in French.
- **Person (`PER`):** Person entities are also recognized relatively well (0.678 F1). Performance here is characterized by very high recall (0.895) but moderate precision (0.546). This

Table 2: Per-Class Zero-Shot Performance for `gliner_multi-v2.1` on the French Subset (N=2000). P/R/F1 based on exact match.

Entity Type	Precision	Recall	F1-Score	Predicted	True	Correct
person	0.5458	**0.8950**	0.6780	1530	933	835
organization	0.2731	0.7563	0.4013	875	316	239
location	**0.7189**	0.8225	**0.7672**	1373	1200	987
misc	0.0811	0.0018	0.0036	37	1643	3

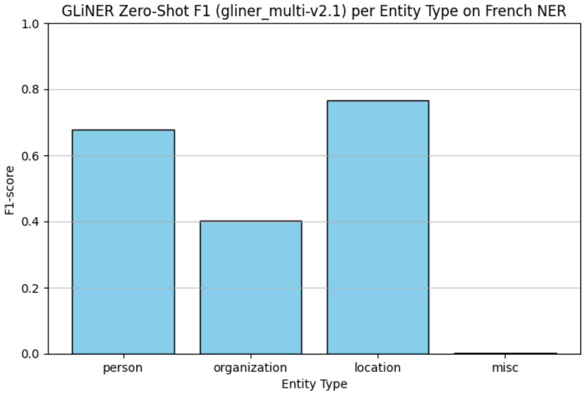

Figure 3: Zero-shot F1-score per entity type for `gliner_multi-v2.1` on the French NER subset (N=2000).

suggests the model is adept at identifying potential person names but may sometimes over-predict, classifying other spans incorrectly as persons.

- **Organization (`ORG`):** Performance on organization entities is considerably weaker (0.401 F1). Similar to the ’person’ type, recall is quite high (0.756), but precision is very low (0.273). The model frequently predicts spans as organizations that are not, leading to many false positives. This could be due to ambiguity in French organizational names or differences from English patterns learned during pre-training.
- **Miscellaneous (`MISC`):** The model almost completely fails to identify miscellaneous entities (0.004 F1). It correctly identified only 3 instances out of 1643 present in this subset. This highlights the inherent difficulty in transferring knowledge for heterogeneous, ill-defined, or dataset-specific categories like `MISC` in a zero-shot setting. The high frequency of `MISC` in the source data analysis further complicates this.

5.3 Comparison with Original GLiNER Multilingual Results

Zaratiana et al. (2024) report zero-shot results for a multilingual GLiNER model on the MultiCoNER dataset (Table 3 in their paper). The average F1-score across the 11 languages evaluated there was **32.9**. The highest F1 score for a single language in that evaluation was 41.7 (English).

Our zero-shot F1-score for French using `gliner_multi-v2.1` on our 2000-example subset is **52.5**. This score is markedly higher than both the average and the maximum score reported in the MultiCoNER evaluation from the original paper.

5.3.1 Interpretation and Caveats

Apparent Strong French Transfer: This result suggests that the multilingual GLiNER architecture might possess a particularly strong capability for zero-shot transfer to French compared to many other languages tested in the original paper. This could stem from factors like the effectiveness of mDeBERTa-V3 on French, linguistic similarities, or the nature of the entities in the `CATIE-AQ` dataset.

Crucial Limitations: This comparison requires significant caution. Our evaluation differs substantially from the paper's MultiCoNER evaluation in several key aspects:

1. *Different Dataset:* We use `CATIE-AQ/frenchNER_4entities`, not MultiCoNER (French wasn't present). Datasets vary in difficulty, annotation guidelines, and domain coverage.
2. *Custom and Limited Test Split:* We evaluated on only the first 2000 examples of a custom test split derived from the original training data. This small subset might not be representative of the full dataset's complexity or the official test split. The results may be inflated if this subset contains simpler examples.
3. *Entity Type Distribution:* The entity types and their distributions differ between the datasets. Our high overall score is driven by strong LOC/PER performance, while the model fails on MISC. The MultiCoNER evaluation likely involves a different mix of challenges.

Therefore, while the 52.5 F1 score is promising, it cannot be directly claimed as state-of-the-art French zero-shot performance without validation on standard benchmarks and full test sets.

Despite these necessary caveats, the result does indicate a noteworthy potential for zero-shot GLiNER application in French.

5.4 Discussion on Fine-tuning Potential

Consistent with the findings in the original GLiNER paper (Zaratiana et al., 2024, Table 4, Figure 5), we anticipate that **supervised fine-tuning** of the `gliner_multi-v2.1` model on the French training data would significantly enhance performance. Fine-tuning allows the model to adapt specifically to the nuances of French entities, potentially improving the precision for ORG and learning patterns for the MISC category (if consistent patterns exist). The `Pile-NER` pre-training provides a valuable initialization, likely leading to better results than training from scratch, especially with limited French data. Implementing this fine-tuning, however, requires a custom training setup beyond the scope of this evaluation.

5.5 Discussion on Backbones and Ablations

Our zero-shot comparison confirmed the benefit of the multilingual backbone (`multi-v2.1`) over English-specific ones (`base`, `large`) for this French task. Replicating the original paper's broader backbone comparisons (Zaratiana et al., 2024, Figure 4) or ablation studies on training techniques like negative sampling and entity dropping (Zaratiana et al., 2024, Section 5.3) was not feasible, as these pertain to the pre-training phase of the models we used. We assume the released models benefit from the optimized training strategies identified in the original work.

6 Conclusion

This report presented an evaluation of pre-trained GLiNER models for zero-shot Named Entity Recognition on French, using the `CATIE-AQ/frenchNER_4entities` dataset. Our experiments, conducted on a 2000-example subset of a custom test split, yielded the following key conclusions:

1) **Effective Zero-Shot Transfer:** GLiNER models, particularly the multilingual variant `gliner_multi-v2.1`, demonstrate substantial zero-shot NER capability on French. The multilingual model achieved an overall F1-score of **0.525** on our limited evaluation subset.
2) **Strong Performance Variation by Type:** The model performed well on common entity types like `Location` (0.77 F1) and `Person` (0.68 F1). However, it struggled with `Organization` (0.40 F1, low precision) and almost completely failed on the `Miscellaneous` category (0.004 F1), indicating challenges in transferring knowledge for ambiguous or heterogeneous types.
3) In conclusion, GLiNER stands as a valuable tool for efficient, generalist NER, exhibiting strong zero-shot potential for French, particularly with its multilingual architecture. Future work should focus on evaluation using standardized French benchmarks and exploring the impact of supervised fine-tuning to fully realize its capabilities for French NER applications.

References

M. Agrawal, S. Hegselmann, H. Lang, Y. Kim, and D. A. Sontag (2022). Large Language Models are Few-shot Clinical Information Extractors. (*arXiv preprint*). Available at: https://arxiv.org/abs/2205.12689

A. Akbik, D. Blythe, and R. Vollgraf (2018). Contextual string embeddings for sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING 2018)*, pages 1638–1649. [Online]. Available: https://aclanthology.org/C18-1139/

D. Ashok and Z. C. Lipton (2023). PromptNER: Prompting For Named Entity Recognition. (*arXiv preprint*). Available at: https://arxiv.org/abs/2305.15444

T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei (2020). Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, pages 1877–1901. [Online]. Available: https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html

A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov (2019). Unsupervised cross-lingual representation learning at scale. (*arXiv preprint*). Available at: https://arxiv.org/abs/1911.02116

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova (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)*, pages 4171–4186. [Online]. Available: https://aclanthology.org/N19-1423/

J. Fu, X. Huang, and P. Liu (2021). SpanNER: Named Entity Re-/Recognition as Span Prediction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7183–7195. [Online]. Available: https://aclanthology.org/2021.acl-long.558/

P. He, X. Liu, J. Gao, and W. Chen (2021). DeBERTa: Decoding-enhanced BERT with Disentangled Attention. In *International Conference on Learning Representations (ICLR)*. [Online]. Available: https://openreview.net/forum?id=XPZIaotutsD

S. Hochreiter and J. Schmidhuber (1997). Long short-term memory. *Neural computation*, 9(8), 1735–1780. Available: https://direct.mit.edu/neco/article-abstract/9/8/1735/6109/Long-Short-Term-Memory?redirectedFrom=fulltext

Z. Huang, W. Xu, and K. Yu (2015). Bidirectional LSTM-CRF models for sequence tagging. In *Proceedings of the 3rd Workshop on Continuous Vector Space Models and their Compositionality*, pages 126–131. [Online]. Available: https://arxiv.org/abs/1508.01991

J. D. Lafferty, A. McCallum, and F. C. Pereira (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning (ICML 2001)*, pages 282–289. Available: https://www.cs.columbia.edu/~jebara/6772/papers/crf.pdf

G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer (2016). Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of* *the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270. [Online]. Available: https://aclanthology.org/N16-1030/

Q. Lhoest, A. S. Villanova, P. von Platen, S. Patil, M. Drame, J. Chaumond, J. Plu, L. Tunstall, J. Davison, Y. Jernite, et al. (2021). Datasets: A Community Library for Natural Language Processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 175–184. [Online]. Available: https://aclanthology.org/2021.emnlp-demo.21/

Y. Li, L. Liu, and S. Shi (2021). An empirical analysis of unlabeled entity problem in named entity recognition. In *International Conference on Learning Representations (ICLR)*. [Online]. Available: https://arxiv.org/abs/2012.05426

X. Li, J. Feng, Y. Meng, Q. Han, F. Wu, and J. Li (2021). A Unified MRC Framework for Named Entity Recognition. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 135–147. [Online]. Available: https://aclanthology.org/2020.acl-main.519/

Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov (2019). Roberta: A robustly optimized bert pretraining approach. (*arXiv preprint*). Available at: https://arxiv.org/abs/1907.11692

Y. Lu, Q. Liu, D. Dai, X. Xiao, H. Lin, X. Han, L. Sun, and H. Wu (2023). Unified Structure Generation for Universal Information Extraction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8873–8889. [Online]. Available: https://aclanthology.org/2022.acl-long.395.pdf

S. Mishra, D. Khashabi, C. Baral, and H. Hajishirzi (2021). Cross-Task Generalization via Natural Language Crowdsourcing Instructions. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3830–3845. [Online]. Available: https://arxiv.org/abs/2104.08773

OpenAI (2023). *GPT-4 Technical Report*. (*arXiv preprint*). Available at: https://arxiv.org/abs/2303.08774

L.-A. Ratinov and D. Roth (2009). Design challenges and misconceptions in named entity recognition. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009)*, pages 147–155. [Online]. Available: https://aclanthology.org/W09-1119/

O. Sainz, O. Lopez de Lacalle, I. Serrano, A. Pérez, and E. Agirre (2023). GoLLIE: Guideline-following Large Language Model for Information Extraction. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2268–2281. [Online]. Available: https://arxiv.org/pdf/2310.03668

Y. Shen, H. Yun, Z. Lipton, Y. Kronrod, A. Anandkumar (2018). Deep Active Learning for Named Entity Recognition. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 374–379. [Online]. Available: https://aclanthology.org/W17-2630/

X. Wang, W. Zhou, C. Zu, H. Xia, T. Chen, Y. Zhang, R. Zheng, J. Ye, Q. Zhang, T. Gui, J. Kang, J. Yang, S. Li, C. Du (2023). InstructUIE: Multi-task Instruction Tuning for Unified Information Extraction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14563–14578. [Online]. Available: https://arxiv.org/abs/2304.08085

U. Zaratiana, N. Tomeh, P. Holat, and T. Charnois (2022). Named entity recognition as structured span prediction. In *Proceedings of the Workshop on Unimodal and Multimodal Induction of Linguistic Structures (UM-IoS)*, pages 1–10. [Online]. Available: https://aclanthology.org/2022.umios-1.1/

U. Zaratiana, N. Tomeh, P. Holat, and T. Charnois (2024). GLiNER: Generalist Model for Named Entity Recognition using Bidirectional Transformer. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5364–5376. [Online]. Available: https://aclanthology.org/2024.naacl-long.300/

W. Zhou, S. Zhang, Y. Gu, M. Chen, and H. Poon (2023). *UniversalNER: Targeted Distillation from Large Language Models for Open Named Entity Recognition*. (*arXiv preprint*). Available at: https://arxiv.org/abs/2308.03279