Evaluating GLiNER's Zero-Shot Cross-Lingual NER Capabilities in French: Architecture, Context, and Performance

Pierre Clayton
ENSAE Paris, Section DSSA
pierre.clayton@ensae.fr

Abstract

Named Entity Recognition (NER) remains a cornerstone task in NLP, yet traditional methods struggle with adaptability to new entity types, while large models impose significant computational overhead. GLiNER (Generalist Model for Named Entity Recognition using Bidirectional Transformer) presents an efficient architecture designed for flexible, open-schema NER by matching entity type embeddings against text span representations derived from bidirectional encoders. This report provides an in-depth look at the GLiNER model architecture and situates it within the broader context of NER research evolution. Subsequently, it presents an empirical evaluation of GLiNER's zero-shot cross-lingual transfer performance specifically for the French language. Using pre-trained checkpoints (`base`, `large`, `multi-v2.1`) primarily exposed to English data, we assess performance on a 2000-example subset of a custom test partition derived from the `CATIE-AQ/frenchNER_4entities` dataset. The multilingual GLiNER variant demonstrates notable transfer capabilities, achieving a 0.525 overall micro F1-score on this subset. However, performance varies significantly across entity types, excelling at 'location' (0.77 F1) and 'person' (0.68 F1) but struggling with 'organization' (0.40 F1) and failing on 'misc' (0.004 F1). While this score is promising relative to some baseline multilingual averages reported elsewhere (on different benchmarks), we emphasize the critical limitations imposed by our specific evaluation setup. The study concludes that GLiNER, especially its multilingual version, offers a valuable approach for efficient, flexible NER with demonstrable zero-shot potential in French, but achieving comprehensive, high-fidelity recognition likely requires supervised fine-tuning.

1 Introduction

Named Entity Recognition (NER), the task of identifying and classifying mentions of entities such as persons, organizations, locations, and other predefined categories within unstructured text, is a foundational component of numerous Natural Language Processing (NLP) applications. From information extraction and question answering systems to knowledge base population and semantic search, the ability to accurately pinpoint entities is crucial.

Despite significant progress, NER systems often face a trade-off between performance on predefined categories and adaptability to new domains or entity types. Traditional supervised models, while achieving high accuracy, are typically constrained to the entity schema defined during training, requiring costly re-annotation and retraining for any schema modifications. Conversely, recent Large Language Models (LLMs) offer remarkable "Open NER" capabilities, recognizing arbitrary entity types specified via prompts, but their immense size translates to substantial computational requirements and potential deployment costs.

Preprint. Under review.

This context motivates the development of more efficient yet flexible NER solutions. GLiNER (Generalist Model for Named Entity Recognition using Bidirectional Transformer) (Zaratiana et al., 2024) represents one such effort. It reformulates NER as a matching problem between entity type descriptions and text spans, leveraging the power of bidirectional transformer encoders while maintaining computational efficiency.

This report aims to provide a comprehensive assessment of GLiNER's capabilities, particularly focusing on its zero-shot cross-lingual transfer to French. We begin with a detailed exploration of the GLiNER model architecture (Section 2) and situate it within the evolution of NER methodologies by reviewing the relevant state-of-the-art (Section 3). Following this contextualization, we present an empirical evaluation of pre-trained GLiNER models on a specific French NER task (Sections 4 and 5). We analyze the overall performance and per-type variations on a subset of the `CATIE-AQ/frenchNER_4entities` dataset, discuss the implications and limitations of these zero-shot findings (Section 6), and conclude with an outlook on GLiNER's potential for French NLP tasks (Section 7).

2 The GLiNER Model Architecture

GLiNER (Zaratiana et al., 2024) introduces an innovative approach to NER, diverging from traditional sequence labeling and generative paradigms. It frames NER as a task of matching learned representations of entity types against representations of candidate text spans. This design prioritizes efficiency and flexibility, enabling the model to identify entities corresponding to arbitrary type descriptions provided at inference time. The core components and mechanics are detailed below.

2.1 Input Formulation

A key aspect of GLiNER is its unified input sequence format. For a given input text and a set of target entity types (e.g., "person", "location", "event"), the model constructs a single sequence for the underlying bidirectional encoder. This sequence typically follows the pattern: [ENT] type_1 [ENT] type_2 ... [ENT] type_n [SEP] text_token_1 text_token_2 ... text_token_m

Here, [ENT] is a special, learnable token prepended to each target entity type description (which are themselves treated as sequences of tokens). [SEP] is another special token separating the entity type prompts from the actual input text tokens. This format allows the encoder to process both the entity type semantics and the text context simultaneously.

2.2 Bidirectional Encoder Backbone

At the heart of GLiNER lies a pre-trained Bidirectional Language Model (BiLM), such as BERT (Devlin et al., 2019), RoBERTa, or, as used in the evaluated checkpoints, DeBERTa-V3 (He et al., 2021) or its multilingual variant mDeBERTa-V3. This encoder processes the entire formatted input sequence described above. Its primary function is to generate deep, contextualized embeddings for every token in the sequence, capturing complex dependencies between tokens, including those between the entity type prompts and the text content. Let $H = \{h_1, h_2, ..., h_N\}$ be the sequence of output embeddings from the final layer of the BiLM, where N is the total length of the formatted input sequence.

2.3 Entity Type Representation (q_t)

For each target entity type t provided in the input (e.g., "person"), GLiNER derives a representative embedding, q_t. This is typically achieved by taking the contextualized embedding h_{ent} corresponding to the special [ENT] token immediately preceding the type description t, and passing it through a dedicated FeedForward Network (FFN): $q_t = \text{FFN}_{\text{type}}(h_{ent})$. This allows the model to learn a refined semantic vector capturing the essence of the requested entity type within the context provided by the BiLM.

2.4 Candidate Span Representation (S_{ij})

GLiNER considers all possible contiguous text spans (i, j) within the input text portion of the sequence, where i is the start token index and j is the end token index, up to a predefined maximum span length K (e.g., $K = 12$). For each candidate span, it computes a representation S_{ij} designed to capture both its boundary information and internal semantics. This is commonly done by concatenating the contextualized embeddings of the start token (h_i) and the end token (h_j) from the BiLM output, and then passing this concatenated vector through another FFN: $S_{ij} = \text{FFN}_{\text{span}}(h_i \otimes h_j)$ where $\otimes$ denotes concatenation. This specific formulation gives prominence to the boundary tokens, which are often crucial for identifying entities accurately.

2.5 Span-Type Matching Score (ϕ)

The core mechanism of GLiNER involves calculating a compatibility score between each candidate span representation S_{ij} and each target entity type representation q_t. This score estimates the likelihood that the span (i, j) corresponds to the entity type t. It is computed using a dot product followed by a sigmoid activation function: $\phi(i, j, t) = \sigma(S_{ij}^T q_t)$ The dot product measures the similarity or alignment between the span and type representations in the learned latent space. The sigmoid function maps this score to a probability-like value between 0 and 1. This computation can be performed efficiently in parallel for all spans and all types.

2.6 Training Objective

GLiNER is trained using a supervised objective based on annotated data. The goal is to maximize the matching score $\phi(i, j, t)$ for "positive" examples (where span (i, j) truly represents an entity of type t in the training data) and minimize it for "negative" examples (incorrect span-type pairs). This is typically formulated using a binary cross-entropy (BCE) loss function summed over relevant positive and negative pairs. The original work (Zaratiana et al., 2024) also mentions techniques like negative entity type sampling during training to improve robustness and generalization, where the model is trained to distinguish the correct type from plausible but incorrect types for a given span.

2.7 Inference and Decoding

At inference time, given an input text and a list of target entity types, the model computes the matching scores $\phi(i, j, t)$ for all valid spans (i, j) and target types t. A decoding process then selects the final set of predicted entities. This typically involves:

1. **Thresholding:** Only considering span-type pairs where the matching score $\phi(i, j, t)$ exceeds a predefined probability threshold (e.g., 0.5).
2. **Handling Overlaps:** Applying a strategy to resolve conflicts where multiple spans overlap or are nested. A common approach is a greedy decoding strategy that iteratively selects the highest-scoring valid prediction and removes any overlapping predictions that conflict with it, depending on whether flat or nested NER is desired. GLiNER's architecture can naturally support nested entities if the decoding strategy allows it.

The output is a list of predicted entities, each defined by its start index, end index, text content, and predicted entity type label.

This architecture allows GLiNER to function as an efficient, generalist NER model capable of handling arbitrary entity types defined at inference time, leveraging the deep contextual understanding of bidirectional transformers through its span-type matching mechanism.

3 Evolution of NER Methodologies and GLiNER's Context

Understanding GLiNER's design and potential requires appreciating the landscape of NER research from which it emerged. NER techniques have evolved considerably, driven by advancements in machine learning and the increasing availability of data and computational power.

3.1 Early and Statistical Approaches

Initial NER systems often relied on hand-crafted rules, dictionaries (gazetteers), and linguistic features. While effective in constrained domains, these systems lacked robustness and scalability. The field then shifted towards statistical machine learning models. Hidden Markov Models (HMMs) and later Conditional Random Fields (CRFs) (Lafferty et al., 2001) became prominent. These models learn probabilistic sequence labeling patterns from annotated data, typically using encoding schemes like BIO or BILOU to tag each token based on its entity membership (Begin, Inside, Outside, Last, Unit). CRFs, in particular, excel at capturing dependencies between adjacent labels.

3.2 Deep Learning for Sequence Labeling

The advent of deep learning revolutionized NER. Recurrent Neural Networks (RNNs), especially Long Short-Term Memory networks (LSTMs), proved effective at modeling sequential dependencies in text. Bidirectional LSTMs (BiLSTMs), which process sequences in both forward and backward directions, combined with a CRF layer on top (BiLSTM-CRF) (Lample et al., 2016; Huang et al., 2015), became a standard and powerful architecture for sequence labeling NER. Performance was further enhanced by incorporating character-level embeddings (Lample et al., 2016) to handle out-of-vocabulary words and contextual string embeddings like Flair (Akbik et al., 2018).

The introduction of large pre-trained transformer models, exemplified by BERT (Devlin et al., 2019), marked another significant leap. Models like BERT, RoBERTa, and DeBERTa (He et al., 2021) learn deep bidirectional representations from massive unlabeled text corpora. Fine-tuning these models by adding a simple classification layer on top of the token embeddings for sequence labeling yielded state-of-the-art results on numerous NER benchmarks. However, a major limitation of these supervised sequence labeling approaches (both BiLSTM-CRF and transformer-based) remained their reliance on a fixed set of entity types defined during training. Adapting them to new types necessitates substantial re-annotation and retraining.

3.3 Span-Based NER Paradigms

An alternative paradigm formulates NER not as token-level labeling but as classifying candidate text spans. Models first identify potential spans (e.g., all n-grams up to a certain length) and then assign an entity type (or 'none') to each span. This approach can more naturally handle nested or overlapping entities. Techniques within this paradigm include formulating NER as a Question Answering task using Machine Reading Comprehension (MRC) frameworks, where questions query the presence of specific entity types, or as structured span prediction (Zaratiana et al., 2022). GLiNER's core mechanism of computing span representations and matching them against type representations clearly aligns with this span-based philosophy.

3.4 Open-Domain, Zero-Shot, and Few-Shot NER

The desire to move beyond fixed entity schemas spurred research into Open-Domain NER, where the model can identify entities of types not seen during training. Large Language Models (LLMs) like GPT-3 (Brown et al., 2020) and its successors proved particularly adept at this through prompting. By providing natural language descriptions of the target entity types in the prompt (zero-shot) or with a few examples (few-shot), LLMs can perform NER for arbitrary schemas. Instruction tuning, where models are fine-tuned on a diverse set of NLP tasks formatted as instructions, further enhances these zero-shot and few-shot capabilities for information extraction tasks, as seen in models like InstructUIE and GoLLIE. Knowledge distillation from powerful LLMs (like ChatGPT) into smaller, fine-tuned models like UniversalNER (Zhou et al., 2023) has also shown exceptional zero-shot NER performance. The primary drawback of these LLM-based approaches is their substantial computational footprint (model size, inference cost, latency), limiting their applicability in many scenarios (Zaratiana et al., 2024).

3.5 Efficient Generalist NER Models

GLiNER (Zaratiana et al., 2024) and similar efforts like USM aim to bridge the gap between traditional supervised models and large-scale LLMs. They seek to provide generalist NER capabilities

(handling arbitrary entity types) with significantly greater computational efficiency. They achieve this by building upon smaller, yet powerful, bidirectional transformer backbones (like DeBERTa-V3) and employing specialized architectures (like GLiNER's span-type matching) that avoid the costly autoregressive generation process inherent in many LLMs. GLiNER's design leverages the contextual power of BiLMs but focuses computation on identifying relevant spans and matching them to type descriptions in parallel.

3.6 Cross-Lingual NER

Applying NER across multiple languages introduces challenges related to linguistic diversity, resource availability, and transfer learning. Multilingual pre-trained transformers like mBERT, XLM-R (Conneau et al., 2019), and multilingual versions of DeBERTa (He et al., 2021) provide a foundation for cross-lingual transfer. Models like the multilingual variant of GLiNER build upon these backbones. Evaluating the zero-shot cross-lingual performance, as conducted in this study for GLiNER on French, directly assesses the model's ability to generalize its learned entity recognition capabilities from its primary training language(s) (mostly English) to a new target language without language-specific fine-tuning.

This study places GLiNER within this evolving landscape, specifically investigating its effectiveness as an efficient, generalist model when performing zero-shot cross-lingual transfer to French.

4 Experimental Setup for French Evaluation

To empirically assess GLiNER's zero-shot cross-lingual transfer capabilities to French, we designed the following experiment.

4.1 Target Task and Models

The core task is zero-shot French NER. We evaluate the performance of pre-trained GLiNER models, which have not been fine-tuned on any French data, on a French NER dataset. The evaluated models are the publicly released checkpoints detailed in Section 2:

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

These models were primarily pre-trained on the English `Pile-NER` dataset (Zhou et al., 2023), making this a test of direct cross-lingual transfer.

4.2 Dataset: CATIE-AQ/frenchNER_4entities

We utilized the French NER dataset `CATIE-AQ/frenchNER_4entities`, available from the Hugging Face Hub[1]. This dataset contains tokenized French sentences annotated with IOB2 tags for four entity types: Person (`PER`), Organization (`ORG`), Location (`LOC`), and Miscellaneous (`MISC`), plus the `O` tag for non-entity tokens.

4.3 Crucial Evaluation Subset Definition

While the official dataset provides standard train/validation/test splits, the experiments reported here (based on the accompanying code) employed a custom data partitioning strategy. The original *train* split (328,757 examples) was shuffled (seed=42) and re-split into a new training set (80%, 263,005 examples) and a new test set (20%, 65,752 examples). **Critically, our zero-shot performance evaluation was conducted only on the first 2000 examples of this custom-generated test set.** This represents a significant limitation, and the results reported must be interpreted as performance on this specific, potentially unrepresentative subset, not necessarily on the full dataset or standard French NER benchmarks.

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

4.4 Dataset Characteristics (Derived Training Split)

To provide context, we analyzed the statistical properties of the custom training split (N=263,005) from which the test subset was drawn.

- **Sentence Length:** The average sentence length is 27.57 tokens. The distribution (Figure 1) shows a concentration of sentences between 15-30 tokens, with a long tail extending to much longer sentences.
- **Entity Type Distribution:** The frequency of annotated entities (Figure 2) reveals `MISC` as the most common type (208,469 instances), followed by `LOC` (151,163), `PER` (135,630), and `ORG` (42,031). The high prevalence and potential heterogeneity of the `MISC` category are notable characteristics of this specific data split.

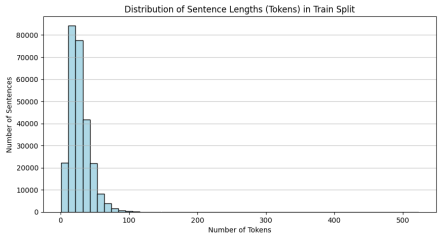

Figure 1: Sentence length distribution (tokens) in the custom French training split (N=263,005).

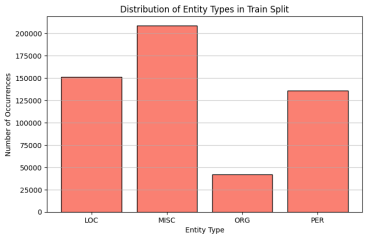

Figure 2: Entity type distribution in the custom French training split (N=263,005).

4.5 Inference and Evaluation

Label Mapping: For GLiNER inference, the dataset's tags were mapped to English natural language labels: `PER` → `person`, `ORG` → `organization`, `LOC` → `location`, `MISC` → `misc`. These English labels were provided as the target entity types to the models.

Prediction and Metrics: We used the official `gliner` Python library and the Hugging Face `datasets` library (Lhoest et al., 2021). Predictions were generated using `model.predict_entities()` with the mapped labels and a probability threshold of 0.5. Performance was measured using standard micro-averaged Precision, Recall, and F1-score, requiring an exact match of both span boundaries and entity type label for a prediction to be considered correct.

5 Zero-Shot Performance Results on French Subset

This section presents the empirical results obtained from evaluating the pre-trained GLiNER models on the designated 2000-example French test subset in a zero-shot setting.

5.1 Overall Performance Comparison

Table 1 aggregates the micro-averaged Precision, Recall, and F1-scores across all four entity types for the three GLiNER model variants tested.

Table 1: Overall Zero-Shot Performance on the French NER Subset (N=2000 from custom split). Micro-averaged P/R/F1.

Model Checkpoint	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**

The results indicate that:

- All models demonstrate non-trivial zero-shot NER capability on French, with F1-scores ranging from 0.475 to 0.525, confirming the potential for cross-lingual transfer using the GLiNER architecture.
- The multilingual model, `gliner_multi-v2.1`, based on mDeBERTa-V3-Base, achieves the highest overall F1-score (0.525). Its advantage stems primarily from significantly higher precision (0.548) compared to the English-only models, even though its recall is slightly lower than the `large` model. This highlights the importance of the multilingual backbone for this cross-lingual task.
- Increasing the size of the English-only model from `base` to `large` yields only a marginal improvement (0.475 to 0.488 F1). The benefit of the multilingual pre-training appears more substantial than model scaling within the English-specific checkpoints for this particular zero-shot French evaluation.

5.2 Performance Breakdown by Entity Type

Analyzing the performance for each entity type reveals significant variations, as shown in Table 2 and visualized in Figure 3 for the best-performing `gliner_multi-v2.1` model.

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

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

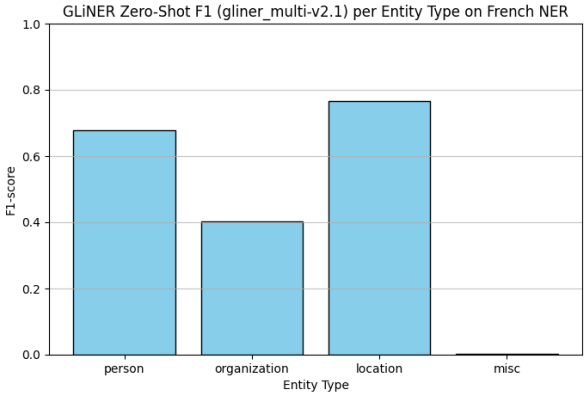

Figure 3: Zero-shot F1-score per entity type for `gliner_multi-v2.1` on the French NER subset (N=2000).

The per-type analysis highlights:

- **Strongest Performance for Location (`LOC`):** Location entities were identified most effectively, achieving the highest F1-score (0.767) with a good balance of precision (0.719) and recall (0.823).
- **High Recall but Moderate Precision for Person (`PER`):** Person entities were also recognized relatively well (0.678 F1). The model exhibited very high recall (0.895), indicating it successfully found most person names, but its precision was only moderate (0.546), suggesting a tendency to over-predict this category.

- **Significant Challenge for Organization (`ORG`):** Recognizing organization entities proved difficult (0.401 F1). Similar to persons, recall was quite high (0.756), but precision was notably low (0.273). This indicates the model frequently proposed spans as organizations incorrectly, leading to many false positives. This could stem from greater cross-lingual variation or ambiguity in organizational names.

- **Near-Complete Failure for Miscellaneous (`MISC`):** The model almost entirely failed to identify entities in the `MISC` category (0.004 F1), correctly identifying only 3 out of 1643 instances in the evaluated subset. This underscores the extreme difficulty of zero-shot transfer for heterogeneous, ill-defined, or potentially dataset-specific categories like `MISC`. Given its high frequency in the source data split (Figure 2), this failure significantly impacts potential overall utility if this category is important.

6 Discussion

The empirical results, situated within the context of GLiNER's architecture and the broader NER landscape, warrant further discussion regarding the implications of zero-shot transfer, the observed performance variations, and the path forward.

6.1 Effectiveness of Zero-Shot Cross-Lingual Transfer

The overall F1-score of 0.525 achieved by `gliner_multi-v2.1` on our specific French subset is noteworthy. It demonstrates that the combination of a powerful multilingual encoder (mDeBERTa-V3) and GLiNER's span-type matching mechanism enables substantial transfer of NER capabilities from English pre-training to French without any target-language supervision. This score, while evaluated on a limited and custom subset, appears considerably higher than the average multilingual F1 (32.9) reported for GLiNER on the MultiCoNER benchmark in the original paper (Zaratiana et al., 2024).

However, as strongly emphasized previously, this comparison is fraught with caveats. The difference in datasets (`CATIE-AQ` vs. MultiCoNER), the lack of French in MultiCoNER, the use of only 2000 examples from a custom split derived from training data – all these factors prevent drawing definitive conclusions about relative performance. The `CATIE-AQ` subset might be easier, or the entity distribution might favor types where GLiNER transfers well (LOC, PER). Nonetheless, the result provides a positive signal about GLiNER's potential for French.

6.2 Factors Influencing Per-Type Performance

The stark performance differences across entity types are arguably the most revealing aspect of the evaluation.

- The success with **LOC** and **PER** suggests that the features distinguishing these entities (e.g., capitalization patterns, typical contexts, common name structures) are sufficiently captured by the multilingual encoder and the span representations, allowing for effective zero-shot transfer. The high recall indicates the model is sensitive to potential candidates, while the reasonable precision (especially for LOC) shows it can often correctly classify them.

- The poor precision for **ORG** highlights a common challenge in NER. Organizational names are diverse, can overlap with common nouns or other entity types, and naming conventions may differ significantly across languages. The model seems to identify many potential candidates (high recall) but struggles to reliably filter out false positives without specific French linguistic or domain knowledge.

- The failure on **MISC** is indicative of the limits of zero-shot learning for highly heterogeneous or ill-defined categories. Without explicit examples or a clearer semantic definition transferable across languages, the model has little basis for identifying such entities. This category often acts as a repository for dataset-specific annotations, making zero-shot transfer particularly arduous.

6.3 The Inevitability of Fine-tuning for Robustness

While the zero-shot performance offers a valuable baseline and might suffice for applications focused solely on well-defined types like locations or persons, achieving high-quality, comprehensive NER across all relevant categories necessitates adaptation. The results strongly suggest that supervised fine-tuning of the pre-trained multilingual GLiNER model (`gliner_multi-v2.1`) on annotated French data would be crucial for practical deployment in demanding scenarios.

Fine-tuning would allow the model to:

- Adapt its representations to French-specific linguistic patterns.
- Improve precision, particularly for ambiguous categories like ORG, by learning discriminative features relevant to the French context.
- Potentially learn to recognize consistent patterns within the MISC category, if such patterns exist in the target French dataset.

As demonstrated in the original GLiNER paper (Zaratiana et al., 2024, Table 4, Figure 5), fine-tuning yields substantial performance gains over zero-shot results. The pre-trained checkpoints provide a strong initialization, likely enabling effective learning even with moderately sized French datasets.

7 Conclusion

This report provided an in-depth examination of the GLiNER model architecture and its place within the evolving field of Named Entity Recognition, followed by an empirical evaluation of its zero-shot cross-lingual performance on a French NER task. Our assessment, conducted on a 2000-example subset of a custom partition of the `CATIE-AQ/frenchNER_4entities` dataset, yielded several key insights:

1. **GLiNER Architecture:** GLiNER offers an efficient and flexible alternative to traditional NER models and large LLMs by framing NER as a span-type matching problem built upon powerful bidirectional encoders.
2. **Substantial Zero-Shot Transfer Potential:** The multilingual GLiNER variant (`gliner_multi-v2.1`) demonstrated significant zero-shot transfer capability to French, achieving an overall micro F1-score of 0.525 on our specific evaluation subset. This highlights the effectiveness of its architecture combined with a multilingual backbone.
3. **Performance Variability:** Zero-shot performance varied dramatically across entity types. While common types like 'Location' (0.77 F1) and 'Person' (0.68 F1) were recognized effectively, 'Organization' proved challenging (0.40 F1 due to low precision), and the 'Miscellaneous' category was almost entirely missed (0.004 F1).
4. **Contextual Performance and Limitations:** While the French F1 score appears high relative to some prior multilingual benchmarks (on different data), direct comparisons are unreliable due to critical differences in datasets and the limited, custom nature of our evaluation subset.
5. **Fine-tuning Recommendation:** For robust and comprehensive French NER across diverse entity types, particularly challenging ones like ORG and MISC, supervised fine-tuning of the pre-trained multilingual GLiNER model on French annotated data is strongly recommended.

In summary, GLiNER stands as a promising model for efficient, generalist NER. Its multilingual version shows considerable aptitude for zero-shot application to French, especially for well-defined entity types. However, realizing its full potential for high-accuracy, broad-coverage French NER will likely depend on targeted fine-tuning to adapt the model to the specific nuances of the language and target domain. Future work should focus on evaluation using standardized French benchmarks and exploring optimal fine-tuning strategies.

References

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. Available: https://aclanthology.org/C18-1139/

T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, et al. (2020). Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, pages 1877–1901. Available: https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html

A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, et al. (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. Available: https://aclanthology.org/N19-1423/

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

Z. Huang, W. Xu, and K. Yu (2015). Bidirectional LSTM-CRF models for sequence tagging. (*arXiv preprint*). 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.

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. Available: https://aclanthology.org/N16-1030/

Q. Lhoest, A. S. Villanova, P. von Platen, S. Patil, 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. Available: https://aclanthology.org/2021.emnlp-demo.21/

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. 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. 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