Data Efficiency of GLiNER: A Multilingual and Multi-Scale Learning Curve Study

Nina Kokalj
Jožef Stefan Institute
Ljubljana, Slovenia
nina.kokalj123@gmail.com

Erik Calcina
Jožef Stefan Institute
Ljubljana, Slovenia
erik.calcina@ijs.si

Abdul Sittar
Jožef Stefan Institute
Ljubljana, Slovenia
abdul.sittar@ijs.si

Erik Novak
Jožef Stefan Institute
Ljubljana, Slovenia
erik.novak@ijs.si

Abstract

Manually annotated clinical data are expensive to obtain, making data-efficient adaptation of named entity recognition models particularly important. We study how GLiNER performance changes with increasing amounts of task-specific training data across three model sizes, four clinical datasets, and multiple languages. Models were fine-tuned on nested subsets containing 10%, 30%, 60%, 80%, and 100% of the available training data and evaluated across three training seeds. The largest gains occurred early: with only 10% of the training data, models already captured approximately 52–76% of the total improvement from zero-shot to full-data performance, increasing to 88–97% at 60%. Larger models often provided the greatest advantage in low-data settings, although this was not consistent across datasets. Learning dynamics also varied across languages, while persistent span-boundary errors remained even after overall performance began to plateau. These results highlight the potential of learning curves to guide annotation effort in iterative clinical NER workflows.

Keywords

named entity recognition, gliner, learning curves, data efficiency

1 Introduction

Named Entity Recognition (NER) is a key component of many clinical text processing pipelines, but adapting NER systems to a specific domain, language, or annotation scheme typically requires manually annotated data. In practice, such annotations are costly to obtain, making it important to understand how much training data is actually needed to achieve useful performance. This question is particularly relevant for models such as GLiNER, which support flexible entity schemas and can be fine-tuned for task-specific extraction [7].

This work was motivated by a human-in-the-loop workflow developed within the European PREPARE project. An initial model identifies entities in clinical documents, which users review and correct. The corrected examples are added to the training data, and the model is retrained as more annotated documents become available. Ideally, this progressively reduces the amount of manual correction required. However, it remains unclear how quickly these gains occur and whether they differ across model sizes and experimental settings.

We evaluate three GLiNER model sizes across five nested training-data fractions, multiple seeds, datasets, and languages to compare how performance evolves with increasing annotation.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
Information Society 2026, Ljubljana, Slovenia
© 2026 Copyright held by the owner/author(s).
https://doi.org/10.70314/is.2026.sikdd.118

2 Related Work

Recent approaches to NER aim to combine the flexibility of large language models with the efficiency of encoder-based architectures. Zaratiana et al. introduced **GLiNER**, a compact bidirectional transformer for flexible NER using natural-language entity labels [7]. While GLiNER showed strong zero-shot performance across multiple benchmarks, the original work did not systematically examine how performance changes with the amount of task-specific fine-tuning data.

In clinical NER, Kormilitzin et al. introduced Med7 for extracting medication-related concepts from electronic health records and showed that performance dropped when the model was transferred to a different clinical setting, but improved substantially after fine-tuning on a small amount of target-domain data [3].

The amount of annotated data available for fine-tuning is particularly important in low-resource and domain-specific NER. Seeha et al. investigated this effect for German medical NER by fine-tuning BERT- and LLaMA-based models on increasing fractions of two annotated datasets, observing large initial gains followed by more gradual improvements as additional data were introduced [5]. More broadly, Urbizu et al. studied how model size and training data interact in low-resource settings using three BERT model sizes across four languages. They found that scaling behavior in such settings differs from that observed for large language models, with smaller-scale models generally requiring relatively more training data to achieve efficient performance [6]. While previous work has examined training-data requirements in NER and scaling in BERT-based models, GLiNER learning curves across model sizes remain underexplored.

3 Methodology

3.1 Dataset characteristics

The experiments were conducted on four clinical NER datasets with different characteristics. **SMMN (Synthetic Multi-Med Notes)** is a multilingual synthetic clinical NER dataset released by E3-JSI [2]. **E3C (European Clinical Case Corpus)** is a multilingual corpus of clinical case narratives collected from heterogeneous, freely licensed sources [4]. **AHR-NL and AHR-IT (Annotated Health Records in Dutch and Italian; study-specific names)** are non-public de-identified clinical datasets

provided by project partners, originating from hospital settings. In the present experiments, E3C uses a single broad clinical entity label, whereas SMMN, AHR-NL, and AHR-IT contain multiple entity types. Table 1 summarizes the main dataset characteristics.

Table 1: Dataset characteristics and size of the full training partition. For multilingual datasets, sizes are reported as the minimum–maximum range across languages.

Dataset	Lang.	Source	Entities	Docs
SMMN	9	Synthetic	Multiple	447-582
E3C	5	Clinical cases	Single broad	798-816
AHR-NL	NL	Hospital	Multiple	488
AHR-IT	IT	Hospital	Multiple	1623

3.2 Data preparation

For each dataset-language combination, the original documents were divided into **training (72%), validation (8%), and test (20%) sets** using stratified splitting to preserve the entity-label distribution across subsets. Long documents were chunked after splitting at approximately 384 tokens while preserving sentence and entity boundaries whenever possible. The validation and test sets were fixed throughout all subsequent experiments.

3.3 Hyperparameter optimization

Before the learning-curve experiment, hyperparameters were optimized separately for each dataset–language–model combination using **60 Optuna trials** [1] on the full training partition. Models were trained on the training set and evaluated exclusively on the fixed validation set. The best-performing hyperparameter configuration was then used for the subsequent learning-curve experiments.

3.4 Learning-curve experiment

To evaluate the effect of training data size, we created five **nested, stratified subsets** containing **10%, 30%, 60%, 80%, and 100%** of the training documents. We denote the subset containing x% of the training data as D_x, such that $D_{10} \subset D_{30} \subset D_{60} \subset D_{80} \subset D_{100}$. Each larger subset retained all documents from the preceding fraction.

For each fraction, models were trained using the selected hyperparameters. Prediction thresholds were optimized on the fixed validation set by maximizing exact micro-F1 and then applied once to the fixed test set. Experiments were repeated with **three seeds (42, 43, and 44)** using identical training subsets. The zero-shot baseline used no task-specific fine-tuning, with its threshold calibrated on the same validation set.

We evaluated three GLiNER model sizes: `gliner_small-v2`[1], `gliner_medium-v2`[2], and `gliner_large-v2`[3] [7]. Performance was measured using exact entity-level micro-averaged precision, recall, and F1. We also calculated relaxed F1, counting predictions as correct when they overlapped the gold span by at least one token; the relaxed–exact F1 difference is referred to as the *boundary gap*.

For cross-dataset summaries, seeds were averaged within each language–model configuration and configurations were weighted equally. Language-specific curves were averaged equally across model sizes after seed averaging. Seed variability was summarized by the sample SD of the three seed-specific dataset-level mean exact micro-F1 scores. Cumulative gain was computed as $(F1_f - F1_0)/(F1_{100} - F1_0)$, where $F1_f$ is the exact micro-F1 at training fraction f, $F1_0$ the zero-shot score, and $F1_{100}$ the full-data score. Marginal gain was the F1 increase between consecutive fractions, normalized per additional 10% of training data. For AHR-IT, the large model was excluded from aggregate learning-curve analyses due to high seed instability.

[1] https://huggingface.co/urchade/gliner_small-v2
[2] https://huggingface.co/urchade/gliner_medium-v2
[3] https://huggingface.co/urchade/gliner_large-v2

4 Results

4.1 Early gains and diminishing returns

Across all four datasets, the largest performance improvement occurred after introducing the first supervised training fraction, followed by progressively smaller gains as more training data were added. Figure 1a shows the cumulative share of the total exact micro-F1 improvement between the zero-shot baseline and training on 100% of the available data. The corresponding mean exact micro-F1 (zero-shot→full-data) was 0.144→0.624 for SMMN, 0.339→0.713 for E3C, 0.032→0.457 for AHR-NL, and 0.190→0.662 for AHR-IT. With only 10% of the training data, the models already captured approximately **52–76%** of this total improvement. At 30%, this increased to approximately **78–91%**, and by 60% all datasets had reached roughly **88–97%** of the improvement observed at full data.

The same pattern is visible in the marginal gains in Figure 1b. The transition from zero-shot inference to training on the first 10% yielded an improvement of approximately 22–37 exact-F1 points across datasets. Increasing the training fraction from 10% to 30% provided only about 3.5–5.6 F1 points per additional 10% of training data, while later increments generally contributed less than 2 points per additional 10%. SMMN approached its full-data performance particularly early, whereas AHR-NL retained comparatively larger gains at later fractions. AHR-IT showed minor non-monotonic variation between 60% and 80%. Across the aggregate learning curves, seed variability was generally low (median SD = 0.004; 90th percentile = 0.015).

4.2 Model-size effects across training fractions

Figure 2 compares the exact micro-F1 of the large and small GLiNER models across training fractions. For SMMN, E3C, and AHR-IT, the largest advantage of the large model occurred at 10% of the training data, reaching +2.4, +8.8, and +16.9 F1 points, respectively. At 30%, these differences decreased to +1.5, +4.6, and +7.9 points. At higher fractions, the gap remained small for SMMN and considerably below its 10% value for E3C.

The pattern was different for AHR-NL, where the small model outperformed the large model at four of the five fractions, including by 1.7 F1 points at full data. AHR-IT also showed substantial variation at larger fractions. At 100%, the large AHR-IT model ranged from 0.274 to 0.652 exact micro-F1 across seeds; its apparent 20.9-point disadvantage should therefore be interpreted cautiously. Overall, larger models tended to benefit most when supervision was limited, while increasing training data reduced the performance gap between model sizes. At full data, large-model training was **1.5–4.4**× slower than small-model training in our setup.

4.3 Cross-language learning dynamics

The two multilingual datasets showed different language-specific learning trajectories (Figure 3). In SMMN, the nine languages

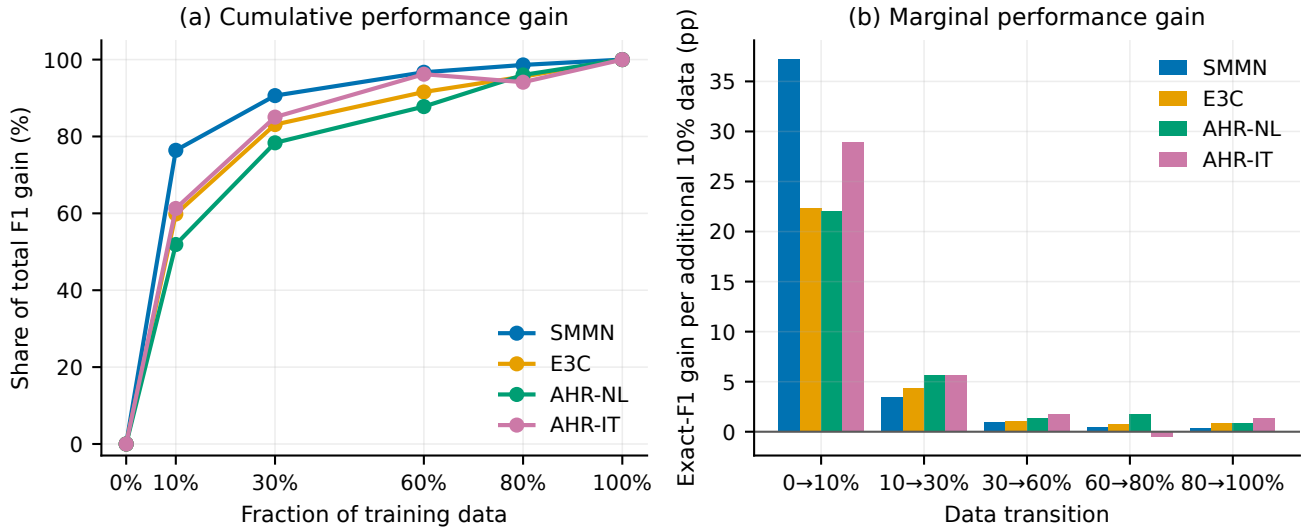

Figure 1: Cumulative share of the total zero-shot-to-100% exact micro-F1 gain (a) and marginal gain per additional 10% of training data (b), measured in percentage points (pp). AHR-IT aggregates include only the stable small and medium models.

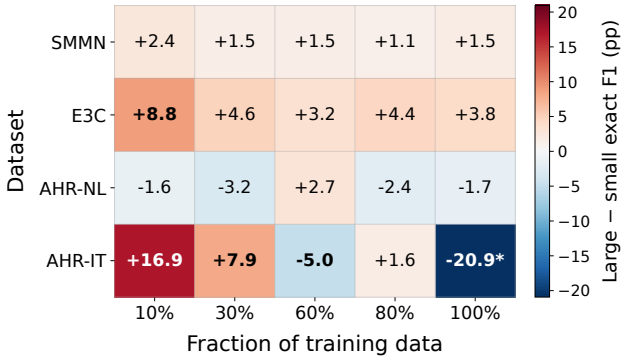

Figure 2: Exact micro-F1 difference between the large and small GLiNER models across training fractions. Positive values favour the large model. *AHR-IT large at 100% showed high seed variability (SD = 0.191).

followed similar curves and were already closely clustered after training on 10% of the data. The remaining differences decreased further as additional supervision was introduced.

E3C showed greater cross-language variation at low training fractions, followed by stronger convergence as more data became available. Basque provides the clearest example: it started well below the other languages in the zero-shot and 10% settings but improved substantially with additional supervision, eventually reaching a similar performance range. English followed a different trajectory, starting from a much stronger baseline and showing smaller subsequent gains.

4.4 Global plateaus mask persistent boundary errors

The gap between relaxed and exact-F1 remained substantial even at high training fractions for some datasets. The mean boundary gap decreased only from approximately 11.9 F1 points at 10% to 10.7 at 100% in SMMN and from 16.3 to 14.3 points in E3C. At full data, the corresponding gaps were smaller for the stable AHR-IT models and AHR-NL, at approximately 4.9 and 3.1 points, respectively.

Label-level results from SMMN further showed that similar exact-F1 performance can reflect different error profiles. *Observation*, the consistently most difficult entity type, reached only about 0.293 exact-F1 at full data but had a boundary gap of approximately 2.1 points. In contrast, *Drug dose*, *Event*, *Rehabilitation*, and *Operation* showed much larger gaps, ranging from approximately 22.7 to 29.9 F1 points. For these entity types, a considerably larger share of the exact-match error was associated with span-boundary mismatches.

5 Discussion

5.1 Annotation and model trade-offs

For iterative annotation workflows, our results suggest that **annotation provides its highest return early in the adaptation process**. A model trained after an initial annotation round can already capture a large share of the improvement achievable with the full training set. However, the point at which further annotation becomes less beneficial is dataset-dependent. The results therefore do not support a single stopping fraction across datasets.

Model selection should similarly depend on the available annotation budget. Where a large-model advantage was present, it was concentrated in the low-data regime, while smaller models generally became more competitive as supervision increased. This pattern was not universal, however. AHR-NL showed no consistent benefit from increased model size, and the large AHR-IT model also exhibited substantial instability at high training fractions. Smaller models may therefore offer a better performance–training-time trade-off with more training data.

5.2 Dataset and language effects

The multilingual results show that the same annotation fraction can yield different gains across languages. This contrast may partly reflect the more homogeneous synthetic construction of SMMN versus the heterogeneous clinical content of E3C, although causality is not established.

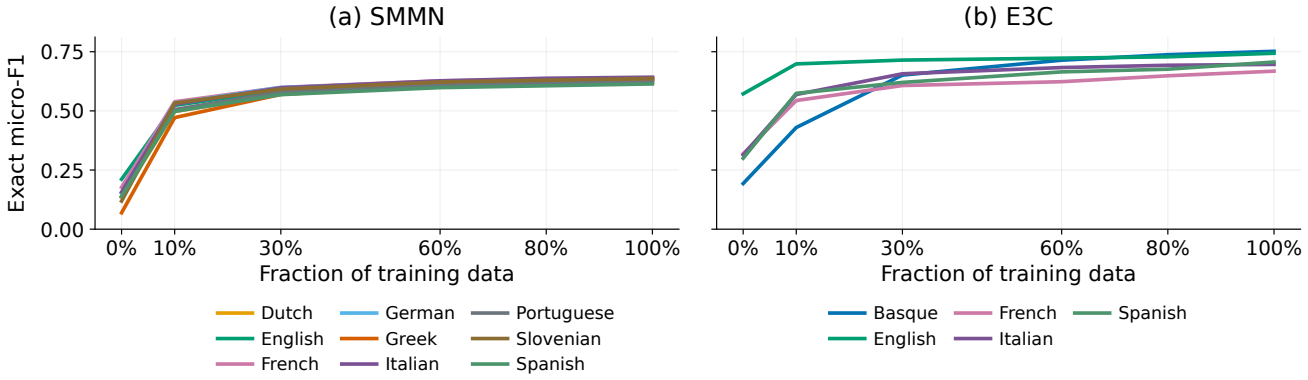

Figure 3: Cross-language exact micro-F1 learning curves for the multilingual SMMN and E3C datasets.

The E3C results suggest that a uniform multilingual annotation budget may be inefficient. Stronger languages may reach diminishing returns earlier, while weaker ones can continue to benefit from additional supervision. Annotation effort could therefore be guided by language-specific learning progress rather than fixed fractions.

5.3 Implications for targeted annotation

Once overall performance begins to plateau, the remaining errors may provide a more useful guide for annotation than the global F1 score alone. Later annotation rounds could therefore shift from broad data expansion toward more targeted examples, prioritizing weaker languages, difficult entity types, or recurring span-boundary errors. Entity types with large boundary gaps may benefit from examples that clarify span boundaries, whereas labels with low exact-F1 but small boundary gaps may require greater focus on recognition or classification errors. Although our experiments do not directly evaluate such a targeted strategy, the results suggest that uniformly adding more data may become increasingly inefficient as the learning curve flattens.

6 Limitations

Several limitations should be considered when interpreting the results. First, the training fractions represent **dataset-specific proportions rather than equal absolute annotation budgets**. A 10% subset may therefore contain different numbers of documents and annotated entities across datasets and languages. The observed saturation points should consequently not be interpreted as a universal amount of training data required by GLiNER.

Second, each configuration was repeated with three training seeds, which limits the precision with which training variability can be estimated. This is particularly relevant for unstable configurations such as the large AHR-IT model.

Finally, hyperparameters were optimized before the learning-curve experiment and then kept fixed across training fractions. This enables a consistent comparison of training-set size, but the optimal configuration may itself depend on the amount of available training data. Future work could examine fraction-specific optimization and evaluate whether the observed learning patterns generalize to additional datasets and annotation settings.

7 Conclusion

The largest gains occurred at low training fractions, followed by diminishing returns. Larger models tended to perform better in the low-data regime, while their advantage generally narrowed as more training data became available. Learning dynamics also varied across datasets and languages, while span-boundary errors persisted after overall performance began to plateau. These findings suggest that annotation effort should follow observed learning progress, with later rounds targeting weaker languages, difficult entity types, and persistent error patterns.

Acknowledgements

This work was supported by the Slovenian Research Agency. Funded by the European Union. UK participants in Horizon Europe Project PREPARE are supported by UKRI grant number 10086219 (Trilateral Research). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Health and Digital Executive Agency (HADEA) or UKRI. Neither the European Union nor the granting authority nor UKRI can be held responsible for them. Grant Agreement 101080288 PREPARE HORIZON-HLTH-2022-TOOL-12-01.

References

[1] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: a next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, 2623–2631. doi:10.1145/3292500.3330701.

[2] E3-JSI. 2026. Synthetic multi-med notes for ner. Hugging Face. (2026). https://huggingface.co/datasets/E3-JSI/synthetic-multi-med-notes-ner-v1.

[3] Andrey Kormilitzin, Nemanja Vaci, Qiang Liu, and Alejo Nevado-Holgado. 2021. Med7: a transferable clinical natural language processing model for electronic health records. *Artificial Intelligence in Medicine*, 118, 102086. doi:10.1016/j.artmed.2021.102086.

[4] Bernardo Magnini, Begoña Altuna, Alberto Lavelli, Anne-Lyse Minard, Manuela Speranza, and Roberto Zanoli. 2023. European clinical case corpus. In *European Language Grid*. Springer, Cham, 283–288. doi:10.1007/978-3-031-17258-8_17.

[5] Suteera Seeha, Sihan Wu, Justin Hofenbitzer, Claudio Benzoni, Peter Pallaoro, Raphael Scheible, Martin Boeker, and Luise Modersohn. 2025. German medical NER with BERT and LLMs: the impact of training data size. In *Intelligent Health Systems – From Technology to Data and Knowledge, Proceedings of MIE 2025* (Studies in Health Technology and Informatics). Vol. 327. IOS Press, (May 2025), 798–802. ISBN: 9781643685960. doi:10.3233/SHTI250469.

[6] Gorka Urbizu, Iñaki San Vicente, Xabier Saralegi, Rodrigo Agerri, and Aitor Soroa. 2023. Scaling laws for BERT in low-resource settings. In *Findings of the Association for Computational Linguistics: ACL 2023*. Association for Computational Linguistics, Toronto, Canada, (July 2023), 7771–7789. doi:10.18653/v1/2023.findings-acl.492.

[7] Urchade Zaratiana, Nadi Tomeh, Pierre Holat, and Thierry 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)*. Association for Computational Linguistics, Mexico City, Mexico, (June 2024), 5364–5376. doi:10.18653/v1/2024.naacl-long.300.