The Effect of Label Only vs. Label Plus Description: Comparing GLiNER1 and GLiNER2 for Clinical Entity Recognition Across Five Languages

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

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

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

Abstract

We study whether zero-shot multilingual clinical Named Entity Recognition (NER) benefits from a longer, descriptive entity label rather than a short name. Using the five-language European Clinical Case Corpus (E3C: English, French, Italian, Spanish, Basque), we fine-tune two model families — GLiNER1 (small) and GLiNER2 (base) — under two label conditions, a short name ("clinical entity") and the same name paired with an explanatory description, and evaluate with both exact-match and IoU-based overlap (Intersection over Union) F1 to separate boundary errors from missed entities entirely. To go beyond aggregate scores, we apply SHAP with a paired ablation check to a span-detection scoring function and present a case study in which a configuration with a higher aggregate F1 nonetheless misses a clinically salient entity that a lower-scoring configuration detects correctly. These results suggest that model-family choice, evaluation granularity, and example-level behavior each carry information a single F1 number cannot, and that longer label descriptions are not a safe default for zero-shot clinical NER.

Keywords

clinical named entity recognition, zero-shot NER, GLiNER, E3C corpus, explainability

1 Introduction

Zero-shot and few-shot NER models such as GLiNER [10] recognize arbitrary entity types by matching free-text label descriptions against spans in the input, rather than relying on a fixed label set learned at training time. This is attractive for clinical NLP in particular, where the entities of interest (symptoms, findings, diagnoses) vary by use case, by institution, and by language, and where collecting enough annotated data to train a conventional fixed-label classifier for every new setting is rarely feasible. One line of prior work addresses this scarcity directly by generating synthetic non-English clinical text [3, 7]; we instead ask whether better use of the label interface itself — rather than more training data — can close some of that gap. In practice, this raises a simple but understudied question: does giving the model a longer, more explicit description of what counts as the entity actually improve extraction, or is a short label name sufficient?

We study this question in a clinical setting — extracting a single generic "clinical entity" span type (symptoms, diagnoses, findings) — across five languages of the multilingual E3C corpus, comparing two model families (GLiNER1 and the newer GLiNER2) under two label conditions (short name vs. name+description). We report both strict exact-match F1 and a more forgiving overlap-based F1, since in a clinical review workflow a span that is off by one token at a boundary is a very different kind of error from a span that is missed entirely, and an exact-match-only evaluation conflates the two. We further use SHAP to explain what individual predictions are driven by, at the level of single words, giving a qualitative complement to the aggregate numbers and a way to sanity-check whether an aggregate F1 improvement actually reflects more robust, generalizable behaviour or is instead concentrated in a handful of easy examples.

This paper contributes (1) a controlled comparison of short-name versus name+description labels for two GLiNER families across five languages, with per-language and per-model tuning, and (2) a multi-angle evaluation — exact-match and overlap F1, manual error inspection, and word-level SHAP with a paired ablation check — applied to a case where aggregate and example-level conclusions diverge.

2 Related Work

2.1 Zero-shot and few-shot NER

GLiNER reframes NER as matching text spans against natural-language label descriptions within a shared encoder, recognizing entity types never seen in training — unlike fixed-inventory pipelines such as spaCy or Stanza. Extensions include GLiNER2 [9], multi-task NER/classification/structured extraction, GLiNER multi-task [8], relation extraction, synthetic-data-augmented pre-training, and biomedical adaptation (GLiNER-BioMed). None of this prior work, to our knowledge, studies whether label phrasing itself — a bare name versus an expanded description — changes extraction quality, the question we isolate here.

2.2 Multilingual clinical NER and E3C corpus

Most clinical NER benchmarks are English-only, limiting cross-lingual generalization. The European Clinical Case Corpus [6] addresses this by annotating clinical case narratives in five European languages with clinical-entity, temporal, and factuality information. Closest to our use, [2] evaluate prompt-based and fine-tuned LLMs on MACCROBAT2020 and extend to further biomedical benchmarks including E3C, but under a single fixed labelling scheme rather than a name-vs-description comparison. Prior E3C work has largely focused on a single language or entity subtype; our use of all five languages under one generic "clinical entity" label treats language itself as an experimental variable alongside model family and label condition.

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.114

Table 1: Document counts per split and language.

Language	Train	Validation	Test	Total
English	248	43	327	618
French	266	47	380	693
Italian	152	27	259	438
Spanish	238	43	422	703
Basque	475	84	719	1,278
Total	1,379	244	2,107	3,730

2.3 Explaining span-detection predictions with SHAP

SHAP (SHapley Additive exPlanations) attributes a model's output to its input features via a game-theoretic Shapley-value decomposition and is standard for explaining tabular and text classifiers [5]. Applying it to a span-detection model like GLiNER is less standard, since GLiNER has no fixed classification head — a prediction is the outcome of matching spans against a label description rather than a single output logit [4]. Our approach to adapting SHAP to this setting is described in Section 3.4.

3 Method

3.1 Dataset

We use the E3C clinical corpus in five languages (English, French, Italian, Spanish, Basque). Each languages origial train file is further split into a stratified train/validation set (85% / 15%, stratified by entity count per document); the original test file is left completely untouched until final scoring. Table 1 gives document counts. This three-way split was adopted after an earlier pipeline iteration reused one file for both tuning and reporting — a leakage risk discussed in Section 5. Here, the validation split is used exclusively for hyper-parameter tuning, and the test split is touched only once, at the end, to compute every number in Section 4.

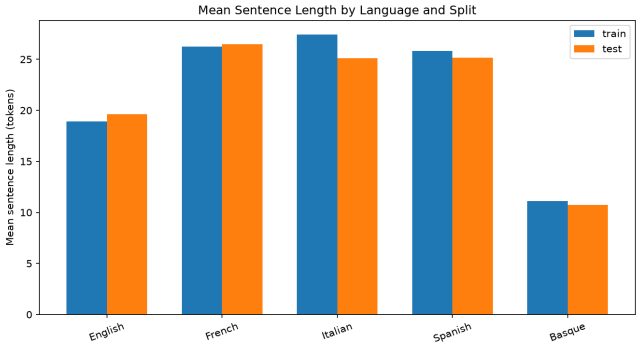


Figure 1: Mean sentence length in tokens for the train and test splits of each E3C language.

3.2 Model and label conditions

We fine-tune two model families: GLiNER1 (G1, small variant) and GLiNER2 (G2, base variant). For each model and language, we train under two label conditions: name (a short label, e.g., "clinical entity") and desc (the same name plus a longer natural language description of what qualifies as the entity — covering symptoms, diagnoses, and clinical findings and excluding, for instance, purely administrative or demographic mentions). Hyperparameters (learning rate, batch size, training steps/epochs, and, for G2, encoder/task learning-rate split and warm-up) are tuned per language and per model with Optuna [1] (20 trials, median pruning) against the validation split only; the test split is never used for tuning. We train final models with the tuned hyperparameters across three random seeds per language x condition to obtain a mean and standard deviation.

3.3 Evaluation

We report exact-match F1 (spans must coincide exactly, IoU = 1.0) and overlap F1 (IoU $\geq$ 0.5,), the latter giving partial credit for spans in roughly the right place but slightly mis-bounded. Both metrics use the same greedy one-to-one matching in descending order of IoU, so a single predicted span cannot be credited against more than one gold span. All numbers are computed on the held-out test split only, after confirming that reloaded checkpoints reproduce their training-time scores.

3.4 Confidence-threshold sweep

GLiNER assigns each predicted span a confidence score; the F1 numbers in Section 4 use a single fixed threshold. To see how each model behaves away from that point, we reuse each language×condition's first-seed checkpoint (no retraining) to sweep the threshold from 0.3 to 0.8 and recompute precision, recall, and F1. This is diagnostic rather than a headline result: it shows where the precision/recall trade-off sits for a given language and model — information a single-threshold F1 cannot convey — useful for choosing between a high-recall screening use and a high-precision autofill use. Full per-language curves are kept in supplementary logs.

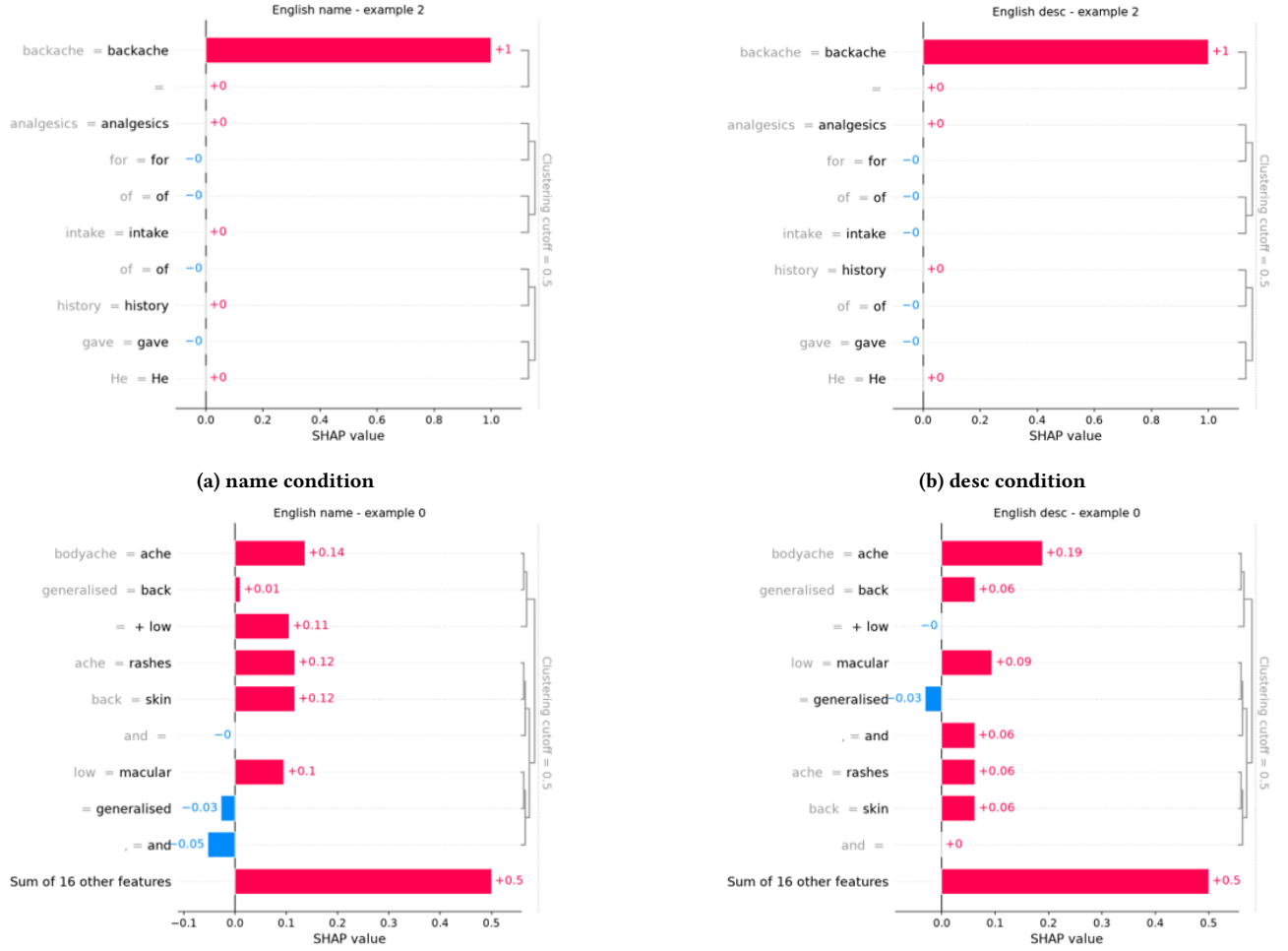
3.5 Qualitative error inspection and interpretability (SHAP)

Alongside aggregate metrics, we manually inspect false positives and false negatives on real test documents for both label conditions and model families; this surfaced the concrete divergence analysed in Section 4.3, where an aggregate score moves one way while a clinically salient entity moves the other. To probe individual predictions further, we define a scoring function equal to the model's single most-confident predicted span for the target label and apply SHAP to it, yielding a signed per-word attribution (positive = pushes toward detection, negative = pushes away). As a sanity check beyond the additive Shapley approximation, we run a paired before/after ablation: removing the single highest-attribution word and re-running the model, to confirm the attribution reflects a genuinely load-bearing word rather than a masking artifact.

4 Results

4.1 Exact-match and overlap F1 (test set)

Table 2 reports exact-match F1 and Table 3 reports overlap F1 across all language x model x condition combinations. Overlap F1 exceeds exact F1 for every language, model, and condition (e.g. English G1-name: 0.861 overlap vs. 0.765 exact), confirming that most misses are near-misses on span boundaries rather than missed entities entirely; G2's overlap-exact gap (7–12 points) is consistently larger than G1's (8–9 points), meaning G2 finds roughly the right region slightly more often than it gets the exact boundary. Across all ten language x condition pairs, G1 wins



(a) name condition **(b) desc condition**

(c) name condition **(d) desc condition**

Figure 2: Word-level SHAP attributions for two further correctly-predicted English test sentences under both label conditions. Leftmost pair: nearly all context words carry near-zero attribution under both conditions, and the target entity word alone (*backache*) drives the prediction (+1), indicating a case where the model's decision is robust to the choice of label description. Rightmost pair: attributions are decomposed into individual clustered word-pair contributions plus the aggregated contribution of the remaining 16 lower-ranked features; the dominant contributor shifts from *low* → *ache* (+0.25 under name) to *bodyache* → *ache* (+0.19 under desc), and the sign of *generalised* → *skin/back* flips between conditions, illustrating that the two label conditions can rely on different words even when both correctly detect the entity.

Table 2: Exact-match F1, mean ± std. across 3 seeds, held-out test set. Higher of G1/G2 per row shown for reference in text.

Language	Cond.	G1 F1 (±sd)	G2 F1 (±sd)
English	name	0.7654 ± .0025	0.7283 ± .0116
English	desc	0.7616 ± .0021	0.7322 ± .0069
French	name	0.6337 ± .0075	0.6426 ± .0106
French	desc	0.6297 ± .0049	0.6386 ± .0077
Italian	name	0.6479 ± .0052	0.6778 ± .0026
Italian	desc	0.6658 ± .0078	0.6776 ± .0053
Spanish	name	0.6616 ± .0082	0.6410 ± .0034
Spanish	desc	0.6499 ± .0031	0.6462 ± .0023
Basque	name	0.7127 ± .0066	0.7043 ± .0063
Basque	desc	0.7265 ± .0021	0.7121 ± .0037

exact-match F1 in 6/10 (English x2, Spanish x2, Basque x2) and G2 wins in 4/10 (French x2, Italian x2); on overlap F1, G1 wins 9/10, losing only a near-tie on Italian-name. G1 is therefore the stronger and more consistent model in this setting, with G2's advantage concentrated in French and Italian — a difference that may partly reflect the small/base size mismatch between the two model families rather than the fine-tuning process alone.

4.2 Effect of the label description

The description's effect is small and inconsistent. For G1, name beats desc in all five languages by a small, consistent margin. For G2, it flips by language: desc wins for English and Spanish, name wins for French, Italian, and Basque — so the answer depends on model family, not just language.

Table 3: Overlap F1 (IoU ≥ 0.5), held-out test set.

Language	Cond.	G1 overlap F1	G2 overlap F1
English	name	0.8610	0.8152
English	desc	0.8510	0.8173
French	name	0.7550	0.7323
French	desc	0.7614	0.7308
Italian	name	0.7451	0.7415
Italian	desc	0.7587	0.7415
Spanish	name	0.8050	0.7282
Spanish	desc	0.7940	0.7292
Basque	name	0.8131	0.7747
Basque	desc	0.8316	0.7800

Table 4: Same sentence, four model×condition combinations.

Config.	Predicted entities	Result
G1 —- name	fever, backache, splenomegaly, Brucellosis, skin rash	5/6, misses gi bleed
G1 —- desc	fever, backache, splenomegaly, Brucellosis, skin rash	5/6, misses gi bleed
G2 —- name	fever, skin rash, backache, splenomegaly, Brucellosis, gi bleed	6/6, perfect
G2 —- desc	fever, backache, skin rash, splenomegaly, Brucellosis, “infection”	misses gi bleed + 1 false positive

4.3 SHAP case study

For the English test sentence "Residing in Middle East with background history of fever, backache, skin rash, splenomegaly and a low platelet count possibility of atypical infection like Brucellosis with NSAID induced gi bleed was considered," gold contains six entities including "gi bleed" (Table 4). Only G2 with the short name recovers all six; G1 (both conditions) and G2-desc all miss "gi bleed," and G2-desc additionally hallucinates "infection." Notably, G2-desc wins on aggregate English F1 (0.7322 vs. 0.7283) — a small aggregate gain can still lose on individual, clinically salient examples.

At the word level, SHAP attributions are intuitive: for a separate sentence, "rashes" carries the largest positive attribution (+0.33) while the negator "non" (as in "non pruritic") pushes negatively (−0.11). Sentences carried by several moderate positive bars appear more robust to small wording changes than single-dominant-bar cases. A paired ablation removing "rashes" confirms the direction: detection confidence drops, matching the SHAP attribution.

5 Discussion and Limitations

Three takeaways follow from Section 4. Model-family choice matters more than label engineering: the G1/G2 gap (up to 5–6 F1 points on English and Spanish) exceeds the name/desc gap within either model. A longer description is not a safe default — it helped G2 on two of five languages but never G1. And aggregate F1 can mask example-level failures that matter clinically, as the "gi bleed" case shows; word-level SHAP is a practical way to audit for this. Limitations: G1 and G2 are not size-matched, so part of the gap may reflect scale rather than architecture; Italian's small validation set (27 documents) meant validation-time ranking did not match final test ranking, so its conclusions should be treated cautiously; SHAP proxies detection confidence, not boundary correctness; and the threshold sweep uses a single seed's checkpoint, so its operating points are illustrative. Finally, this study uses a single generic "clinical entity" label rather than fine-grained subtypes — testing whether the label-length trade-off changes for narrower entity types is a natural next step.

6 Conclusion

Across five languages, GLiNER1 is the more consistent clinical NER model, with GLiNER2's advantage concentrated in French and Italian; a descriptive label helps GLiNER2 on some languages but never GLiNER1. Overlap-based evaluation and word-level SHAP add information exact-match F1 alone does not — most errors are near-miss boundaries, and SHAP shows concretely which words drive or undermine individual predictions, as the "gi bleed" case illustrates. These results argue for evaluating zero-shot NER label choices with more than a single aggregate F1 number. Future work includes fine-grained entity subtypes, additional model families under the same leakage-free protocol, and a human evaluation of whether SHAP explanations help clinical annotators audit predictions faster than raw output alone.

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*, 2623–2631.

[2] Erik Calcina, Erik Novak, Dunja Mladenić, and PREPARE Project Group. 2026. Extracting biomedical entities from clinical records: a comparative study of large language model approaches. *Applied Artificial Intelligence*, 40, 1, 2700905.

[3] Lenart Dolinar, Erik Calcina, and Erik Novak. 2025. Evaluating open-source large language models for synthetic non-english medical data generation using prompt-based techniques. *Informatica*, 49, 27, 125–134.

[4] Xueguang Li, Di Lin, Xue Jiang, Yanxi Li, and Yugang Chi. 2026. Explanation-guided medical named entity recognition with stability and boundary awareness for atopic dermatitis. *arXiv preprint arXiv:2606.22886*.

[5] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

[6] Bernardo Magnini, Begoña Altuna, Alberto Lavelli, Manuela Speranza, and Roberto Zanoli. 2020. The e3c project: collection and annotation of a multilingual corpus of clinical cases. In *Proceedings of the Seventh Italian Conference on Computational Linguistics (CLiC-it 2020)*, 190–196.

[7] Abdul Sittar, Mateja Smiljanic, Alenka Guček, and Marko Grobelnik. 2026. Fake news detection through llm-driven text augmentation across media and languages. *Machine Learning and Knowledge Extraction*, 8, 4, 103.

[8] Ihor Stepanov and Mykhailo Shtopko. 2024. Gliner multi-task: generalist lightweight model for various information extraction tasks. *arXiv preprint arXiv:2406.12925*.

[9] Urchade Zaratiana, Gil Pasternak, Oliver Boyd, George Hurn-Maloney, and Ash Lewis. 2025. Gliner2: schema-driven multi-task learning for structured information extraction. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 130–140.

[10] 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)*, 5364–5376.