Twenty's Plenty: Semantic Scaffolding and Span Architecture for 19-Label NER in Medieval Latin Charters

Tamás Kovács **Giuseppe Consolo** **Georg Vogeler**
Department of Digital Humanities
University of Graz
{tamas.kovacs, giuseppe.consolo, georg.vogeler}@uni-graz.at

Abstract

We ask whether a few hundred annotated sentences are sufficient to build a high-quality named entity recogniser for medieval Latin charters spanning 19 fine-grained entity types — including complex legal clauses, property boundary descriptions, and rare fiscal categories. We evaluate three tiers of approach: zero-shot GLiNER baselines, fine-tuned GLiNER variants (including a domain-pretrained medieval Latin model), and a custom span-based architecture pairing XLM-RoBERTa-large with an Asymmetric Focal–Dice loss and BGE-M3 label encodings. Our results show that *semantic scaffolding* — providing richly descriptive English label phrases — activates latent multilingual knowledge and enables fine-tuned GLiNER to reach 80.8% overlap F1 with no oversampling. Critically, domain pre-training on medieval Latin offers no advantage once task-specific fine-tuning is applied (77.1% vs. 77.2%). A custom span model further surpasses all GLiNER variants at 83.4% overlap F1, with persistent challenges on rare and structurally complex categories.

1 Introduction

Medieval Latin charters are the core sources for Central European history — recording land transfers, judicial decisions, and social hierarchies in a formulaic Latin that is computationally challenging: non-standardised orthography, heavy abbreviation, and dense entity structures in which a single identified person (ACTOR) simultaneously encodes a personal name (PER), an office (TITLE), and an institutional affiliation (INS).

Annotating such material at scale is expensive. Expert medievalists are few, and gold-standard annotation for fine-grained NER requires months to stabilise. The practical question is therefore: *how much annotated data do we actually need, and can architectural choices compensate for data scarcity?*

We exploit a property of GLiNER (Zaratiana et al., 2024): the *label-as-prompt* mechanism, where the semantic content of an entity type description becomes part of the model input and directly shapes span detection. By crafting English descriptions that encode domain expertise — *"legal clause declaring rights, conditions, penalties, or papal commands"* rather than LEG — we activate knowledge latent in the model's multilingual space, routing Latin surface forms through English conceptual anchors. We call this strategy **semantic scaffolding**.

Existing systems for historical Latin NER cover at most two entity types (Chastang et al., 2021; Ehrmann et al., 2023) and none address the full documentary complexity of charters. We contribute: (1) a systematic three-tier comparison of zero-shot, fine-tuned, and custom span-based NER across 19 entity categories on a new charter corpus; (2) the first empirical demonstration that domain pre-training on medieval Latin provides *no* downstream advantage after task-specific fine-tuning; (3) a custom span architecture that surpasses all GLiNER variants without oversampling, together with a per-label analysis identifying the structural sources of remaining failures.

2 Data and Annotation Schema

Corpus. Our dataset consists of 458 sentences drawn from twenty 13th-century Austrian and Central European charters from the Monasterium.net (Vogeler, 2019; Consolo et al., 2026), manually annotated with 19 entity categories spanning four documentary layers: *persons and roles* (PER, ACTOR, TITLE, REL), *places and landscape* (LOC, INS, NAT), *property and legal content* (EST, PROP, LEG, TRANS), and *time and value* (TIM, DAT, MON, TAX, COM, NUM, MEA, RELIC).

A key design decision is the separation of EST (*estate* a short physical plot — field, meadow, vine-

Proceedings of the 6th International Conference on Natural Language Processing for the Digital Humanities, pages 236–241
July 4, 2026 ©2026 Association for Computational Linguistics

yard) from PROP (*property*, a full boundary description including past owners and movable goods). This distinction reflects genuine diplomatic practice but produces a heavily imbalanced distribution: PROP has only 37 corpus instances. The full list of label descriptions used as prompts, with test-set frequencies, is provided in Appendix A.

The semantic scaffold in practice. We use English descriptive phrases for all 19 labels to activate cross-lingual knowledge without extra training data — for example, *"unit of measurement for land, volume, or weight such as mansus, carratas, or talentum"* for MEA. All descriptions use English regardless of Latin surface forms, consistent with the scaffold hypothesis in Section 1.

Splits. We apply stratified splitting preserving per-label distribution: 65% training (298 sentences), 15% development (69), 20% test (91). Fixed seed 42; **no oversampling** in any experiment.

3 Models

3.1 Zero-Shot and Fine-Tuned GLiNER

GLiNER (Zaratiana et al., 2024) scores each span (i, j) against label l via $s(i, j, l) = \sigma(h_{ij}^{\top} e_l)$, where both span representation h_{ij} and label embedding e_l are produced by the same encoder. The label string is an encoder input, not metadata: its semantic content determines e_l and what the model searches for. We evaluate three checkpoints: **gliner-bi-large-v2.0** (Stepanov et al.), a large bidirectional model; **gliner-multi-v2.1** (Zaratiana et al., 2024), a general multilingual model; and **gliner-multi-v2.1-medieval-latin** (Medieval Data), further pretrained on medieval Latin text. Each checkpoint is evaluated zero-shot and fine-tuned for up to 20 epochs with early stopping (patience 5); hyperparameters are held constant.

3.2 Custom Span-NER Architecture

Our model enumerates all spans up to width 80 tokens. We set this limit higher than the 55-token default used in standard NER benchmarks: inspection of the training corpus revealed that PROP boundary descriptions routinely exceed 60 tokens. Empirical search over widths confirmed 80 as the sweet spot: wider windows (e.g. 85) admit excessive noise into the attention layer, increasing ACTOR false positives substantially. Span representations concatenate the start token, end token, and a **4-head attention-pooled** interior from **XLM-RoBERTa-large** (Liu et al., 2019; Conneau et al., 2020) plus a learned width embedding, projected to 512 dimensions via an MLP with LayerNorm and dropout. Attention pooling prevents the semantic dilution of long boundary descriptions caused by simple mean-pooling. To fully leverage semantic scaffolding, we replace the standard classification head with a **dot-product bi-encoder** space. Frozen **BGE-M3** (Chen et al., 2024) encodings of the 19 label phrases serve as label vectors; both span and label representations are L2-normalised, and their dot product is scaled by a learned temperature (initialised to $\log(100.0)$).

Loss and training. A **Dynamic Per-Class Asymmetric Focal Loss** (Lin et al.) suppresses false positives (γ_{neg}=2.0) while scaling γ_{pos} from 1.0 (frequent labels) to 4.5 (classes with <8 training instances), preventing the network from ignoring rare categories. **Dice Loss** (Sudre et al.) (weight 1.0) optimises an F1-like objective; an **InfoNCE**(Oord et al., 2019; Rusak et al., 2025) contrastive term (weight 0.6, prediction temperature T=1.35) pushes spans toward their English label anchors. We also apply **90% hard negative mining**: sampled negatives are required to partially overlap with gold entities, penalising partial-span false positives in nested structures. Training: $\text{lr}_{\text{enc}}=10^{-5}$, $\text{lr}_{\text{head}}=2\times10^{-4}$; batch size 4, gradient accumulation 8 (effective batch 32); 30 epochs, patience 10; no oversampling.

4 Results

We report *overlap* F1 (span overlaps a gold entity of the same label) and *exact* F1 (exact boundaries and label match). Thresholds $t \in \{0.40, 0.45, 0.50\}$; best test result reported per model. For the Custom Bi-Encoder, the best threshold is t=0.50.

Zero-shot limits. Table 1 shows severe zero-shot limitations. Only gliner-bi-large exceeds 40%; the medieval pretrained model reaches 21.2%, barely above the general multilingual model at 9.2%. Both the medieval and multi models predict almost exclusively high-frequency categories (PER, LOC, INS), yielding zero recall on most specialist labels. Without fine-tuning, semantic scaffolding is insufficient to bridge the gap between general pretraining and the documentary realities of charter NER.

The scaffold activates fine-tuning. Fine-tuning with our descriptive label prompts yields gains of

System	Overlap F1	Exact F1
Zero-shot		
GLiNER bi-large	46.2	—
GLiNER medieval	21.2	—
GLiNER multi	9.2	—
Fine-tuned GLiNER (English label prompts)		
GLiNER bi-large	80.8	61.9
GLiNER multi	77.2	—
GLiNER medieval	77.1	—
Custom Span-NER (this work)		
Custom Bi-Encoder (Ours)	**83.4**	**67.7**

Table 1: Test overlap F1 at best threshold per system. Exact F1 reported for top models only.

$+35$ to $+71$ percentage points. The scaffold transforms sparse training signal into rich constraints: the model is not learning the semantics of fiscal obligation or land measurement from 298 sentences — it is being pointed toward knowledge already encoded during multilingual pretraining.

Domain pre-training: a null result. The medieval pretrained model (77.1%) converges to near-identical performance to the general multilingual model (77.2%) after fine-tuning — a gap of 0.1 pp within evaluation variance. This is a substantive negative finding with a practical implication: effort is better invested in task annotation than in domain-adaptive pre-training.

Custom Span-NER. The custom model achieves 83.4% overlap F1 at t=0.50, exceeding the best fine-tuned GLiNER by 2.6 pp; the difference is statistically significant (McNemar $p < 0.01$). Frequent categories achieve high F1 (PER 95.7%, TITLE 92.4%, LOC 93.5%), while the Dynamic Per-Class Focal Loss substantially recovers recall on mid-frequency rare categories (TAX 75.0%, TIM 73.2%, COM 72.7%). A notable span-width effect is visible in our per-label error analysis (Appendix B): at width 80, ACTOR false positives total 82 at t=0.50, compared to 105 at width 85 and over 500 in earlier single-encoder configurations. This confirms 80 tokens as the sweet spot for this corpus: long enough to contain full PROP and LEG spans, but short enough to limit attention noise in nested person phrases. Position-dependent legal categories remain challenging (LEG 53.1%, TRANS 52.6%), though both improved substantially over the width-85 configuration (48.1% and 49.2%). PROP reaches 0.0% despite γ_{pos}=4.5: PROP and EST share near-identical surface forms, and 3 of 8 test instances exhibit direct boundary overlap with EST spans — a guideline-level ambiguity rather than a modelling failure that further training cannot resolve without annotation guideline revision. Full per-label results are in Appendix B.

5 Discussion

5.1 Why Domain Pre-Training Does Not Help

Zero-shot performance of the medieval model (21.2%) confirms that in-domain vocabulary alone is insufficient without task adaptation. After fine-tuning, task annotations provide a more relevant signal than unlabelled charter text: the model learns the exact span boundaries and label distinctions of our schema, which simply do not occur in raw diplomatic text. This result challenges the intuition that historical domain pre-training is a prerequisite for historical NER.

5.2 The Span Architecture Advantage

Components. The custom model's gain over fine-tuned GLiNER stems from several complementary factors. The 4-head attention pooling prevents semantic dilution in exceptionally long property and legal descriptions, where simple mean-pooling would average away the boundary cues that distinguish PROP from surrounding text. The Dynamic Per-Class Focal Loss ensures the network does not collapse into predicting only the dominant classes (PER, LOC) to minimise global loss: extreme penalty weights (γ_{pos}=4.5) force the gradient to treat a missed *lucrum camere* (TAX) as orders of magnitude more costly than a missed *Stephanus* (PER). The Dice Loss component optimises F1-like overlap directly, rescuing rare positive classes further suppressed under BCE-based losses. Finally, the InfoNCE term realises semantic scaffolding at the gradient level: span representations are explicitly pushed toward their English label anchors during training, extending the scaffolding effect beyond the label-as-prompt surface into the learning signal itself.

5.3 Positional, Not Semantic, Failures

The most challenging labels (LEG, TRANS, PROP) share a structural property: their boundaries depend on the charter's formulaic zone rather than lexical content alone. The same Latin phrase may constitute a LEG clause in the *corroboratio* but serve as connective tissue in the *dispositio*; similarly, an ACTOR phrase in the *intitulatio* follows a rigid formula that differs structurally from an actor reference em-

Configuration	Ov. F1	ΔOv.	Ex. F1	ΔEx.
Full model (Ours)	**83.4**		**67.7**	
w/o InfoNCE loss	61.1	−22.3	41.8	−25.9
w/o Multi-Head Attn (H=1)	78.1	−5.4	53.6	−14.2
w/o Prediction Temp.	81.7	−1.7	64.0	−3.7
w/o Dynamic Focal Loss	82.9	−0.5	64.8	−2.9
w/o Hard Negative Mining	83.6	+0.2	64.1	−3.7

Table 2: Leave-one-out ablation on the test set. Overlap F1 measures span-level coverage; Exact F1 measures boundary precision. Both reveal distinct failure modes.

bedded in the *dispositio*. The model understands the semantics of fiscal exemption and land transfer correctly — it fails because it lacks positional awareness within the document's formulaic structure.

A key methodological caveat is that our training data was compiled at *sentence level*: each sentence was annotated independently, without access to its position within the full charter. This means the model never observed the structural signal that would allow it to disambiguate zone-dependent labels. Processing complete charters as documents — with explicit section markers for *arenga*, *dispositio*, *sanctio*, and *corroboratio* — would provide precisely this missing context, and we expect it to substantially improve LEG, TRANS, and ACTOR recognition simultaneously. Document-level encoding is therefore the single most impactful next step for this corpus, beyond any further architectural refinement.

PROP's complete failure (0.0%) is a distinct problem: the EST/PROP boundary is an annotation-level ambiguity that document structure alone cannot resolve, and requires guideline revision.

5.4 Ablation Study

To quantify the contribution of each architectural component, we perform a leave-one-out ablation: each variant removes a single component from the full model, keeping all other settings fixed (Table 2). We report both overlap and exact F1, since the two metrics expose complementary failure modes.

Semantic alignment (InfoNCE). Removing InfoNCE causes the largest drop (−22.3 pp overlap, −25.9 pp exact), confirming it is the mechanism through which semantic scaffolding operates architecturally: without it, Latin span representations cannot align with English label anchors in the dot-product space.

Long-span processing and boundary precision. Reverting to single-head mean-pooling reduces overlap F1 by 5.4 pp but exact F1 by 14.2 pp; LEG overlap recall falls from 53.1% to 23.6%, showing the model detects clause regions but loses precise boundaries. Hard negative mining shows the converse pattern: its removal marginally *increases* overlap F1 (+0.2 pp) while reducing exact F1 by 3.7 pp, as ACTOR false positives rise from 82 to 152 — the model covers more spans but their boundaries deteriorate.

Rare-entity rescue and calibration. The −0.5 pp overlap drop from removing Dynamic Focal Loss understates its importance: TAX, COM, and MEA all collapse to 0.0% recall without γ_{pos}=4.5, as dominant classes absorb the gradient budget entirely. Disabling prediction temperature (−1.7 pp overlap, −3.7 pp exact) confirms the contrastive space requires calibration before thresholding.

6 Conclusion

Twenty charters are enough — if you write the right labels and use the right architecture. Semantic scaffolding enables fine-tuned GLiNER to reach 80.8% overlap F1 on 19-label charter NER from 298 sentences; domain pre-training adds nothing after task fine-tuning. A custom bi-encoder with 4-head attention, Dynamic Per-Class Focal–Dice loss, and InfoNCE contrastive training achieves 83.4%, confirmed by ablation to depend critically on semantic alignment (−22.3 pp without InfoNCE) and boundary supervision (−14.2 pp exact F1 without multi-head attention). Remaining failures trace to positional ambiguity (LEG 53.1%, TRANS 52.6%) and annotation-level boundary confusion (PROP 0.0% vs. EST); both point to document-zone-aware encoding as the primary next step.

Limitations

The corpus covers 13th-century Austrian/Central European charters; generalisability elsewhere requires study. Results on rare categories ($n \leq 11$) exhibit high variance; multi-seed averaging will be used in future to ensure robustness. Single-expert annotation without inter-annotator agreement is a limitation; a formal study is planned.

Acknowledgments

The work presented in this paper has been supported by the ERC Advanced Grant project

(101019327) “From Digital to Distant Diplomatics”.

References

Pierre Chastang, Sergio Torres Aguilar, and Xavier Tannier. 2021. A Named Entity Recognition Model for Medieval Latin Charters. 015(4).

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *Preprint*, arXiv:2402.03216.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Giuseppe Consolo, Tamás Kovács, and Georg Vogeler. 2026. Named entity recognition (NER) dataset for medieval latin charters from Monasterium.Net.

Maud Ehrmann, Ahmed Hamdi, Elvys Linhares Pontes, Matteo Romanello, and Antoine Doucet. 2023. Named Entity Recognition and Classification in Historical Documents: A Survey. 56(2):27:1–27:47.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal Loss for Dense Object Detection. *Preprint*, arXiv:1708.02002.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *Preprint*, arXiv:1907.11692.

Medieval Data. medieval-data/gliner_multi-v2.1-medieval-latin.

Aaron Oord, Yazhe Li, and Oriol Vinyals. 2019. Representation learning with contrastive predictive coding. *Preprint*, arXiv:1807.03748.

Evgenia Rusak, Patrik Reizinger, Attila Juhos, Oliver Bringmann, Roland S. Zimmermann, and Wieland Brendel. 2025. InfoNCE: Identifying the Gap Between Theory and Practice. *Preprint*, arXiv:2407.00143.

Ihor Stepanov, Mykhailo Shtopko, Dmytro Vodianytskyi, and Oleksandr Lukashov. The Million-Label NER: Breaking Scale Barriers with GLiNER bi-encoder. *Preprint*, arXiv:2602.18487.

Carole H. Sudre, Wenqi Li, Tom Vercauteren, Sébastien Ourselin, and M. Jorge Cardoso. Generalised Dice overlap as a deep learning loss function for highly unbalanced segmentations. volume 10553, pages 240–248.

Georg Vogeler. 2019. ‚Monasterium.net‘ – eine Infrastruktur für diplomatische Forschung. 24(1):247–252.

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)*, pages 5364–5376, Mexico City, Mexico. Association for Computational Linguistics.

A Label Dictionary

B Per-Label Results

Label	Definition (used as label prompt)	n
Persons and roles		
PER	Individual person name, without titles	157
ACTOR	Person phrase incl. name, noble title, profession, or geographic origin	146
TITLE	Social rank, noble title, ecclesiastical office, or papal rank	155
REL	Word indicating family, kinship, marriage, or social relationship	30
Places and landscape		
LOC	Geographical place, settlement, city, diocese, or region	178
INS	Monastery, abbey, church, or religious order as a corporate body	73
NAT	River, stream, forest, mountain, or natural landscape feature	15
Property and legal content		
EST	Short physical plot of land, estate, farm, meadows, woods, or vineyards	25
PROP	Detailed property description incl. boundaries, movable goods, past owners	8
LEG	Legal clause declaring rights, conditions, penalties, or papal commands	100
TRANS	Verb or phrase denoting a transaction, confirmation, transfer, or donation	26
Time and value		
TIM	Time period, duration, dating formula, indiction, or regnal year	22
DAT	Specific calendar date, year of incarnation, or liturgical feast day	14
MON	Money, currency, or monetary value (libra, solidus, denarius, marca)	6
TAX	Customary toll, tax, tithe, exaction, lucrum camere, or tribute	7
COM	Crops, food, goods, salt, wine, wax, gold, wood, or animals	6
NUM	Number written as a word or Roman numeral, including fractions	17
MEA	Unit of measurement for land, volume, or weight (mansus, carratas)	6
RELIC	Holy relic, cross, or sacred object of veneration	0
Total		**1054**

Table 3: All 19 entity categories with full label prompt definitions and test-set instance counts (n). RELIC has no test instances; retained for schema completeness.

Label	P	R	F1	n
PER	92.3	99.4	95.7	157
LOC	90.1	97.2	93.5	178
TITLE	90.7	94.2	92.4	155
NUM	100.0	82.4	90.3	17
ACTOR	84.5	85.6	85.0	146
INS	82.4	83.6	83.0	73
REL	70.0	93.3	80.0	30
EST	81.8	72.0	76.6	25
TAX	66.7	85.7	75.0	7
TIM	78.9	68.2	73.2	22
DAT	100.0	57.1	72.7	14
COM	80.0	66.7	72.7	6
NAT	90.0	60.0	72.0	15
MON	60.0	50.0	54.5	6
LEG	50.5	56.0	53.1	100
TRANS	48.4	57.7	52.6	26
MEA	66.7	33.3	44.4	6
PROP	0.0	0.0	0.0	8
Overall	82.2	84.7	**83.4**	1054

Table 4: Per-label overlap F1, Custom Bi-Encoder (width-80) at t=0.50. Overall: P 82.2%, R 84.7%, F1 83.4%.