Data Augmentation for Historical NER: A Systematic Comparison of Lexical and LLM-based Approaches

Léa Blinière[1], Maud Ehrmann[1], Emanuela Boros[1], Simon Clematide[2], Frédéric Kaplan[1]

[1]École Polytechnique Fédérale de Lausanne, Switzerland
[2]University of Zurich, Switzerland

lea@bliniere.com {first.last}@epfl.ch simon.clematide@cl.uzh.ch

Abstract

Named Entity Recognition (NER) on historical materials suffers significant performance degradation compared with modern text, owing to optical character recognition (OCR) errors, language evolution, and scarce annotated training data. Although various remedies have been explored to increase robustness and generalization, data augmentation techniques, despite their proven effectiveness on modern NER benchmarks, remain largely unexplored in the historical setting. This article investigates data augmentation strategies for historical NER through a systematic comparison of two complementary approaches: intrinsic augmentation via mention replacement and extrinsic augmentation through large language model (LLM)-based corpus annotation. We experiment with different augmentation variants and corpus sizes on French and German Swiss historical newspapers. Our results show contrasting patterns: mention replacement yields stable improvements across settings, whereas LLM-based silver data is most useful at moderate scale and when quality-filtered, but its effectiveness degrades as additional pseudo-labeled data is introduced. Overall, simple lexical augmentation emerges as the more robust strategy for historical NER, while LLM-based approaches remain sensitive to annotation noise and data shift.

1 Introduction

Over the past two decades, large-scale digitization efforts by cultural heritage institutions across Europe and beyond have made vast collections of newspapers, books, and archival documents available in machine-readable form (Balk and Conteh, 2011; Neudecker and Antonacopoulos, 2016), opening new possibilities for large-scale information extraction and exploration of historical sources (Neudecker, 2022; Ehrmann et al., 2023a; Bunout et al., 2023). Among the semantic enrichments applied to such collections, named entities – people, places, and organizations – stand out as particularly valuable: their automatic recognition provides key entry points for document retrieval, semantic indexing, and the tracing of historical actors, events, and geographies across large corpora (Gooding, 2016; Fokkens et al., 2018; Düring et al., 2023; Koolen et al., 2023).

Despite this potential, Named Entity Recognition (NER) on historical texts remains substantially more challenging than on modern data (van Strien et al., 2020). OCR noise, diachronic language variation, and domain-specific genre and layout conventions all complicate the identification of entity boundaries and types (Hamdi et al., 2020; Ehrmann et al., 2023b). Moreover, annotated training data for historical NER is scarce and expensive to produce, especially in multilingual settings and under fine-grained annotation schemes. As a result, even strong NER models often show marked performance degradation when applied to historical newspaper corpora (Schweter and Baiter, 2019; Todorov and Colavizza, 2020).

Data augmentation is a natural way to address this limitation, with techniques successfully applied across NLP tasks and, more specifically, to NER across a range of domains and settings (Feng et al., 2021; Huang et al., 2025). In the historical setting, however, such techniques remain underexplored and it is unclear which forms of synthetic training data are most useful. In particular, there is still little evidence on the relative utility of two plausible strategies: lexical augmentation based on existing gold annotations, and LLM-based pseudo-labeling of additional in-domain text to create silver-standard training material. This raises a practical question: can data augmentation improve historical NER, and if so, which strategy proves most effective?

This article addresses this question through a controlled comparison of these two augmentation strategies for historical NER on French and Ger-

man Swiss historical newspapers. We compare intrinsic augmentation via mention replacement, which preserves gold labels while varying entity surface forms, with extrinsic augmentation based on LLM-generated silver annotations for additional in-domain sentences. Our results show that the two approaches behave differently: mention replacement yields stable improvements across settings, whereas LLM-based silver data is most useful at moderate scale and becomes less effective as more pseudo-labeled data is added. These findings suggest that simple mention replacement augmentation is currently the more robust strategy for historical NER, while LLM-based augmentation remains more sensitive to annotation noise and data shift.

Our contributions are as follows:

(1) We present a systematic comparison of lexical and LLM-based data augmentation for historical NER, evaluated on French and German Swiss historical newspapers across multiple augmentation variants and corpus sizes.

(2) We show that mention replacement yields stable gains, whereas LLM-based silver data is most useful at moderate scale and degrades with additional pseudo-labeled data.

(3) We discuss the implications of this contrast for the practical design of augmentation pipelines in low-resource historical NER.

The following sections describe the related work, data, methods, experiments, and results in turn.

2 Related Work

2.1 Historical NER

The HIPE shared task series[1] has played an important role in documenting the challenges of historical NE processing and establishing standardized evaluation protocols across multiple languages and document types, with editions in 2020 and 2022 (Ehrmann et al., 2020, 2022) and a forthcoming 2026 edition focusing on person-place relation extraction (Opitz et al., 2026). Participating teams developed both data-centered approaches, notably transfer learning through historical domain pretraining (Schweter et al., 2022), and model-centered ones, such as architectural modifications to improve OCR robustness (Boros et al., 2020). While these strategies yield consistent gains, performance on historical documents remains below that achieved on modern benchmarks.

[1]Identifying Historical People, Places and other Entities: https://hipe-eval.github.io

A core limiting factor is the scarcity of annotated training data. Historical document collections span vast topical, linguistic, and temporal diversity — from 17th-century administrative records to 20th-century multilingual newspapers — making it difficult to build annotated resources with sufficient coverage. Producing gold-standard annotations is further complicated by OCR noise, evolving naming conventions, and period-specific entities, rendering the process both slow and costly (Ehrmann et al., 2023b). Despite continuous efforts to expand available resources – including recent work by Schneider et al. (2025) – dataset scale and diversity remain limited, highlighting the need for more cost-effective strategies to expand training data.

2.2 Data Augmentation for NER

Data augmentation (DA) refers to strategies that increase the size and diversity of training examples without explicitly collecting new data. While its application to NLP is complicated by language's compositional nature – naive transformations risk producing incoherent or semantically inconsistent samples – augmentation techniques have nonetheless been successfully deployed across a range of NLP tasks addressing low-resource settings, corpus bias, and class imbalance (Feng et al., 2021). For NER specifically, DA has emerged as a valuable strategy for tackling data scarcity, with approaches spanning simple rule-based transformations to generative prompt-based methods (Huang et al., 2025).

Among rule-based methods, mention replacement (MR) has emerged as a widely used baseline. MR substitutes entity mentions in annotated sentences with alternative mentions of the same type drawn from the training corpus, preserving both labels and the distributional properties of the original data. Dai and Adel (2020) demonstrate consistent improvements in low-resource biomedical and materials science settings. Subsequent work has proposed semantically and contextually informed extensions: Phan and Nguyen (2022) introduce semantic neighbor replacement, which constrains substitutions to similar entity mentions using embedding-based filtering, yielding gains over random MR on biomedical corpora, while Bartolini et al. (2023) further refine this with dynamic context-aware selection, improving performance in few-shot scenarios. A shared limitation of these approaches is their closed-world assumption: augmentation draws exclusively from existing training entities, amplifying rather than correcting

any dataset biases.

LLMs offer complementary augmentation strategies, though with varying degrees of reliability. A first line of work leverages LLMs to assist annotation: Naraki et al. (2024) show that LLMs can identify and correct errors in existing annotations, including missing entities and label switches, with hybrid human-LLM labels outperforming purely manual ones on CoNLL-2003, though the authors caution that LLMs tend toward label imbalance. A second line generates entirely new annotated training instances: Dao et al. (2025) demonstrate gains on biomedical NER across multiple languages using LLaMA-3.2, while Kamath and Vajjala (2025) evaluate GPT-4-generated data for low-resource medical NER across eleven languages. Both studies, however, reach consistent conclusions regarding synthetic data limitations: synthetic annotations tend to introduce hallucinated entities, small amounts of gold data consistently outperform larger synthetic datasets, and LLMs struggle with domain-specific terminology — limitations likely compounded in the historical setting, where OCR artifacts and time-specific language fall outside the distribution of models trained primarily on contemporary text.

Finally, direct LLM inference for NER has also been explored, but consistently falls short of supervised models: even sophisticated few-shot strategies combining kNN-based demonstration selection and self-verification cannot match supervised performance on CoNLL-2003 (Wang et al., 2025), and this gap persists on historical documents, where Zhang and Colavizza (2025) find that DeepSeek-V3 with retrieval-based few-shot prompting remains below state-of-the-art fine-tuned models on HIPE-2022. This further motivates the use of LLMs as annotation tools rather than inference engines.

To our knowledge, this is the first systematic study of DA for NER on historical newspaper data.

3 Data

Our experiments draw on two historical newspaper datasets: the `hipe-2020` NE-annotated dataset for model training and evaluation, and `historical-corpus` as the external corpus for LLM-based corpus annotation.

Coarse tag set	Fine tag set
PERS	PERS.IND
	PERS.COLL
	PERS.IND.ARTICLEAUTHOR
ORG	ORG.ADM
	ORG.ENT
	ORG.ENT.PRESSAGENCY
PROD	PROD.MEDIA
	PROD.DOCTR
TIME	TIME.DATE.ABS
LOC	LOC.ADM.TOWN
	LOC.ADM.REG
	LOC.ADM.NAT
	LOC.ADM.SUP
	LOC.PHYS.GEO
	LOC.PHYS.HYDRO
	LOC.PHYS.ASTRO
	LOC.ORO
	LOC.FAC
	LOC.ADD.PHYS
	LOC.ADD.ELEC
	LOC.UNK

Table 1: HIPE dataset entity types for NERC.

3.1 HIPE NER Dataset

Developed for the HIPE-2020 and 2022 evaluation campaigns on NE recognition and linking in historical documents, the HIPE dataset comprises several NE-annotated collections, primarily historical newspapers and classical commentaries. We use the French and German portions of the `hipe-2020` subset, consisting of articles from Swiss and Luxembourgish newspapers (19th–20th century)[2]. Annotations cover two tasks: entity recognition and classification, and entity linking, the latter not considered here. For NER, the dataset provides two levels of granularity: `NER-Coarse` with five top-level entity types, and `NER-Fine` with 21 specific subtypes (Table 1). Metonymic readings are additionally annotated for PERS, ORG, and LOC under `NER-Coarse-Meto`. Annotation followed detailed guidelines covering entity mention form (mainly proper names), type definitions and coverage, and annotation rules (Ehrmann et al., 2020). Corpus statistics and entity type distributions across splits are provided in Appendix A (Tables 6–8).

3.2 Corpus of Historical Newspapers

The `historical-corpus` dataset serves as the external source for the LLM-based annotation experiments. Sampled from the same digitized Swiss and Luxembourgish historical newspaper collection as

[2]Specifically, release v2.1 of the 2022 campaign: https://github.com/hipe-eval/HIPE-2022-data/releases/tag/v2.1-test-all-unmasked

hipe-2020, it comprises yearly editorial content samples from four German and seven French newspapers spanning 1876–1945, already segmented into sentences. It therefore shares the same domain, time period, and textual characteristics, including OCR noise.

4 Methods

Our approach uses fine-tuned BERT models as baselines and compares two data augmentation approaches: intrinsic augmentation via mention replacement, which recombines existing gold annotations from hipe-2020, and extrinsic LLM corpus annotation, which generates pseudo-labeled data from the external historical-corpus.

4.1 Baseline Models

We adopt an extended BERT architecture that, following Boros et al. (2020), adds two transformer encoder layers between the pretrained backbone and the classification head. NER is performed at the sentence level, with each sentence processed independently as input to the model. As pretrained backbone, we use HMBERT (Schweter et al., 2022; Schweter, 2022), a multilingual model trained on digitized historical documents in five languages – including French and German – drawn from Europeana newspaper collections and British Library books (17th–20th century). Unlike standard multilingual BERT pretrained on contemporary Wikipedia, its historical pretraining corpus exposes the model to OCR noise and historical language variation, making it a well-suited choice for historical NER.

This baseline configuration – extended BERT with the HMBERT backbone – is referred to as HIST-base. It was selected based on an earlier French-only version of this study, in which 12 model configurations (3 pretrained models $\times$ 2 architectures $\times$ 2 sequence lengths) were systematically evaluated across 5 random seeds. HIST-base emerged as both a top-performing baseline and the configuration most responsive to the MR and LCA augmentation strategies. Further details are provided in Blinière (2026) (Sections 4.2, 5.1, and 5.4). The same pretrained weights and architecture are used for both languages, ensuring that performance differences across languages are attributable to augmentation strategies rather than model configuration.

4.2 Augmentation Strategies

4.2.1 Mention Replacement (MR)

MR substitutes entity mentions in annotated sentences with alternative mentions of the same type drawn from the training corpus. Given a sentence containing an entity (e.g., *Charles de Gaulle*), the method replaces it with another entity of the same type (e.g., *Winston Churchill*), generating a new training example while preserving sentence structure and annotation labels. This approach has two key advantages: label consistency is guaranteed since replacement mentions originate from gold-annotated data, and the distributional properties of the original corpus – including OCR noise and domain-specific vocabulary – are preserved in the augmented sentences.

MR candidate pool selection. For each sentence s in the hipe-2020 training set containing an entity mention m_i, a candidate pool $\mathcal{C}(m_i)$ is constructed by filtering all other entities in the training set through three successive stages: (1) type filtering, retaining only entities matching the coarse type, metonymic status, and fine-grained type of m_i;[3] (2) quality filtering, excluding mentions with annotation issues such as mentions spanning sentence boundaries; (3) deduplication filtering, discarding near-identical mentions via OCR-robust sentence embeddings (Michail et al., 2025), removing mention candidates with cosine similarity ≥ 0.99. For each candidate c_j in $\mathcal{C}(m_i)$, an augmented sentence s' is generated by substituting m_i with c_j while leaving all other mentions unchanged. Up to two candidates are sampled per mention, yielding up to $2 \times N$ augmented sentences for a sentence containing N mentions.

MR augmentation variants. We evaluate four candidate-selection variants:

1. Random: no additional constraints.
2. Semantic: retains augmented sentences with sentence similarity between s and s' of at least 0.85 (same OCR-robust model), ensuring contextual coherence between the original and augmented sentences.
3. Temporal: restricts replacements to mentions from documents published within ± 10 years,

[3] Candidate selection does not enforce grammatical correctness; the impact of this limitation was not quantified or evaluated.

preserving period-specific naming conventions.

4. `Semantic+Temporal`: combines both constraints.

4.2.2 LLM Corpus Annotation (LCA)

LLM corpus annotation generates new pseudo-labeled training data by annotating sentences from `historical-corpus`.

Annotation. Three LLMs serve as annotators: GPT-5-Mini (OpenAI), Mistral-Small-3.2 (Mistral AI), and Qwen3-Next-80B (Alibaba Cloud). Both the model set and the prompting strategy were selected through preliminary experiments on the `hipe-2020` test set (FR), in which multiple LLMs and prompt configurations were compared based on NER performance; Qwen3-Next-80B achieved the highest overall performance, followed by Mistral-Small-3.2 and GPT-5-Mini (see Blinière (2026), Section 4.4.2).

For each language, 25,000 sentences are sampled from the `historical-corpus` (seed = 42), restricted to editorial content (no ads) and to sentences with lengths in $[20, 500]$ tokens. Each sentence is independently annotated by all three models using an identical system prompt – in French and German – describing the HIPE annotation scheme and guidelines (see Appendix C).

Ensembling. LLM predictions are validated using three strategies with increasing levels of quality control, trading off recall for precision through stricter agreement and validation constraints[4]:

1. `LLM-Strict`: all three LLMs must detect an overlapping span with matching `Coarse` and `Fine` types; Qwen boundaries are retained. This strategy maximizes precision at the expense of recall.

2. `LLM-Majority`: at least two of three LLMs must predict spans that overlap on at least one token and have matching types; boundaries follow Qwen > Mistral > GPT.

3. `LLM-Majority+BERT`: LLM majority vote confirmed by `HIST-base` (confidence ≥ 0.5) on an overlapping span with matching types; TIME entities reintroduced from BERT for validated sentences. This strategy combines LLM consensus with BERT verification for higher annotation reliability.

Only sentences with at least one validated mention are retained, ensuring that augmented data contributes an NER signal, though this increases the proportion of entity-bearing sentences relative to the original corpus.

4.2.3 Augmented HIPE NER Dataset Construction

Both strategies produce pools of NE-annotated sentences that are appended to the original `hipe-2020` training data. To study the effect of augmented training data size, we construct eight augmented datasets ranging from $+25\%$ to $+200\%$ of `hipe-2020` sentence count, in steps of 25%. All augmented datasets include the full HIPE training set. Augmented sentences are added using nested sampling, such that lower augmentation levels are strict subsets of higher ones (e.g., $25\% \subset 50\% \subset \ldots \subset 200\%$), enabling controlled comparison of scaling effects.

The two augmentation methods differ in their sentence selection strategies. Mention replacement uses a round-robin diversification scheme that maximizes both sentence and entity replacement diversity across augmentation levels. LCA applies filtered random selection, retaining only sentences containing at least one validated entity. Each resulting dataset is used to independently fine-tune `HIST-base` under the configuration described in Section 5. Table 2 reports entity volume growth and mean entity density per sentence for French at $+25\%$ and $+200\%$. German exhibits comparable trends. Full per-level statistics for both languages – including entity density, volume growth, type distributions, and rates of novel entity introduction – are provided in Figures 2–5 (Appendix E).

In total, two augmentation paradigms, seven variants, and eight dataset scales are evaluated across two languages.

5 Experimental Setup

Training configuration. `HIST-base` models are fine-tuned for 3 epochs using AdamW optimization with a constant learning rate and a maximum sequence length of 512 tokens. Full hyperparameter details are provided in Appendix B.

Evaluation metrics. We evaluate at entity level, treating each entity mention as a single unit regardless of token length. Evaluation follows the

[4]Mentions of type TIME are excluded from ensembling due to poor zero-shot performance observed in preliminary experiments on the `hipe-2020` test set.

	+25%		+200%	
Method	**Vol.**	**Dens.**	**Vol.**	**Dens.**
Baseline	—	2.17	—	2.17
MR-Random	+45%	2.19	+519%	2.99
MR-Temporal	+45%	2.19	+519%	2.99
MR-Semantic	+51%	2.28	+928%	4.97
MR-Sem+Temp	+51%	2.28	+928%	4.97
LLM-Strict	+48%	2.23	+361%	2.33
LLM-Majority	+53%	2.31	+424%	2.53
LLM-Maj+BERT	+47%	2.22	+373%	2.29

Table 2: Entity volume growth (Δ% vs. baseline) and mean entity density (ent./sent.) at +25% and +200% augmentation levels, for French (baseline: 7,138 entities). Full per-level evolution in Figures 2 and 3 (Appendix E).

HIPE scorer protocol[5], with one adaptation: entity mentions spanning automatic sentence boundaries are split into two distinct gold entities, one per sentence. This is necessary because automatic sentence segmentation is not always reliable and models process text sentence by sentence. The HIPE scorer supports two evaluation settings: *strict*, requiring exact boundary and type match, and *fuzzy*, allowing partial boundary overlap. We adopt strict evaluation throughout. Performance is reported as micro-averaged F1 across all entity types. To account for training stochasticity, all models are trained with five random seeds and results are reported as mean F1 ± standard deviation.

This work is designed to be fully replicable. All code and augmented datasets are publicly available[6]. The `hipe-2020` dataset is distributed publicly and our baseline replicates the architecture of Boros et al. (2020) on the same benchmark.

6 Results

6.1 Overall NER Performance

Figure 1 shows F1 as a function of augmentation level for coarse- and fine-grained NER, from which three patterns emerge.

First, all four MR variants match or improve on the baseline at every augmentation level, and their trajectories are nearly indistinguishable, suggesting that augmentation volume matters more than candidate selection strategy. By contrast, `LLM-Strict` and `LLM-Majority` fall below the baseline from the first augmentation step and decline monotonically as more pseudo-labeled data is added. `LLM-Majority+BERT` occupies an intermediate position: it remains near-neutral or slightly positive on `NER-Coarse`, while consistently improving over the baseline on `NER-Fine`, peaking at early augmentation levels (+25% in French, approximately +75% in German) before stabilizing. Second, gains are generally larger in German than in French across all methods and tasks, consistent with the lower German baseline leaving more room for improvement. Third, MR gains are larger on `NER-Fine` than on `NER-Coarse`, a pattern we examine further at the entity type level in Section 6.2.

		French		German	
Task	**System**	**F1**	**Δ**	**F1**	**Δ**
Coarse	HIST-base	76.2 ± 0.5	—	71.7 ± 0.8	—
	Best MR	77.2 ± 0.6	+0.9	74.9 ± 1.0	+3.2
	Best LCA	76.7 ± 0.6	+0.5	71.8 ± 1.1	+0.1
Fine	HIST-base	69.3 ± 0.7	—	63.2 ± 0.7	—
	Best MR	72.4 ± 0.7	+3.2	69.1 ± 0.6	+5.9
	Best LCA	70.7 ± 0.6	+1.4	65.2 ± 0.8	+2.0

Table 3: F1 scores (mean ± std over 5 seeds) of best MR and LCA variants for Coarse and Fine NER in French and German. Δ = absolute improvement over baseline (pp). Best configurations for each cell are listed in Table 10.

	HIST-base		Δ MR		Δ LCA	
Type	**FR**	**DE**	**FR**	**DE**	**FR**	**DE**
LOC	82.6	82.4	–0.1	+2.2	+0.1	+0.1
PERS	75.9	67.9	–0.2	+4.0	–0.1	–0.3
ORG	50.3	44.0	+8.0	+5.8	+3.4	+3.0
PROD	64.3	50.7	+10.1	+0.1	+2.5	–6.5
TIME	60.5	62.7	+1.8	+11.1	+0.4	+3.0

Table 4: Baseline F1 and ΔF1 (pp) per coarse entity type. MR = best per language (FR: Semantic +150%; DE: Sem+Temporal +125%); LCA = `LLM-Majority+BERT` +25% (both languages).

Table 3 summarizes peak performance across settings for both granularity levels. The best MR configurations (detailed in Table 10) yield gains ranging from +0.9 pp (FR, coarse) to +5.9 pp (DE, fine), whereas the best LCA configurations reach +1.4 pp (FR, fine) and +2.0 pp (DE, fine) but remain close to the baseline on coarse NER. Full per-configuration results are reported in Tables 11–16 (Appendix D).

6.2 Performance by Entity Type

Table 4 reports ΔF1 by coarse entity type for the best MR and LCA configurations. In French, gains

[5] https://github.com/hipe-eval/HIPE-scorer

[6] https://github.com/impresso/impresso-named-entity-data-augmentation

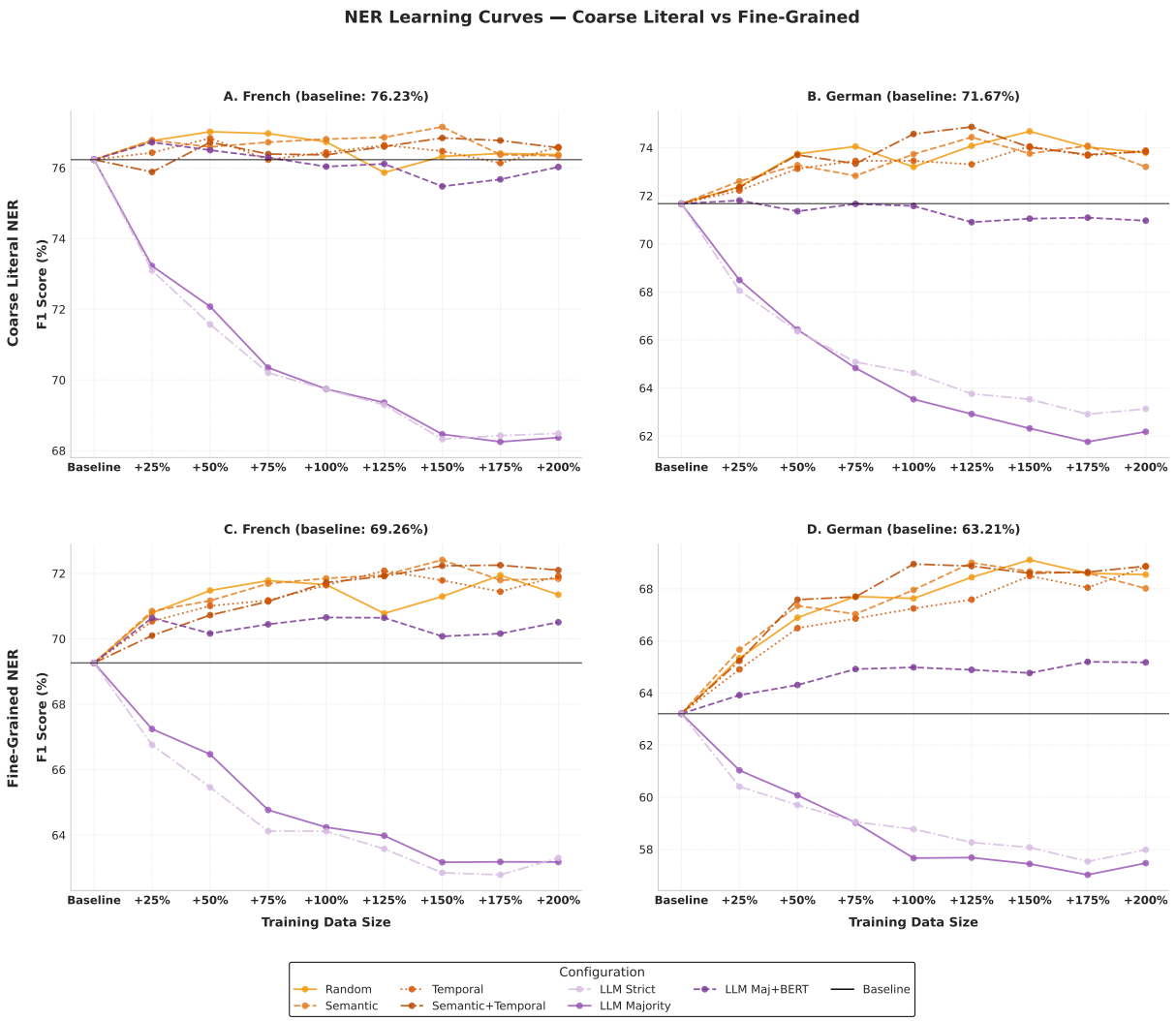

Figure 1: NER F1 as a function of augmentation level for French (left) and German (right). Top: coarse-grained NER; bottom: fine-grained NER. The horizontal line indicates the baseline.

from both methods concentrate on the types with the lowest baseline performance, that is, ORG (baseline 50.3; Δ best MR: +8.0 pp, Δ best LCA: +3.4 pp) and PROD (baseline 64.3; Δ best MR: +10.1 pp, Δ best LCA: +2.5 pp), while LOC and PERS remain largely unchanged. In German, PROD, despite its lower baseline, is essentially unchanged under MR (+0.1 pp) but is severely degraded by LCA (−6.5 pp). The largest gains in German concentrate on ORG and TIME, the latter showing the strongest MR improvement (+11.1 pp).

At the fine-grained level (Table 5), MR consistently maintains or improves performance across all subtypes in both languages. The largest gains occur for rare subtypes, including ADM.SUP (+33.8 pp FR, +77.6 pp DE) and ENT.PRESSAGENCY (+15.4 pp FR, +49.8 pp DE). By contrast, the most frequent subtypes, which already showed the strongest baseline performance, change little: PERS.IND and ADM.TOWN remain stable or improve only marginally in both languages.

LCA results show a more uneven pattern. Some rare subtypes benefit substantially: PHYS.HYDRO gains +21.8 pp in French and +10.2 pp in German, while ADM.SUP improves by +15.4 pp and +6.1 pp, respectively. However, other subtypes degrade sharply, including LOC.FAC in French (−5.5 pp) and PROD.MEDIA in German (−10.1 pp). This disparity suggests that LLM annotation quality varies across entity types, with some categories being more prone to systematic errors.

7 Discussion

Why mention replacement is consistently effective. The robustness of MR across task granularity levels, languages, and augmentation levels

		n		HIST-base		Δ MR		Δ LCA	
Parent	**Subtype**	**FR**	**DE**	**FR**	**DE**	**FR**	**DE**	**FR**	**DE**
LOC	ADM.TOWN	450	257	78.6	76.7	+0.7	+3.5	–0.1	+0.8
	ADM.NAT	151	161	81.7	82.3	+3.5	+3.6	+2.7	+2.1
	ADM.REG	147	84	38.3	35.9	+14.2	+19.1	+5.5	+7.4
	PHYS.GEO	28	14	37.5	0.0	+11.7	+31.4	–5.4	+7.9
	PHYS.HYDRO	23	29	38.0	30.5	+31.6	+16.3	+21.8	+10.2
	ADM.SUP	19	21	46.3	3.6	+33.8	+77.6	+15.4	+6.1
	ORO†	19	5	66.3	56.6	+0.4	+6.0	–0.2	–24.2
	FAC	18	14	16.1	2.2	+10.9	+29.7	–5.5	–1.3
	PHYS.ASTRO†	—	10	—	0.0	—	0.0	—	0.0
	UNK†	3	1	0.0	0.0	0.0	0.0	0.0	0.0
ORG	ENT	69	85	48.0	39.9	+5.6	+5.9	+2.8	+7.5
	ADM	43	29	47.5	37.1	+10.0	+10.0	+4.0	+10.0
	ENT.PRESSAGENCY	20	18	64.0	25.0	+15.4	+49.8	+0.1	+24.0
PERS	IND	519	330	76.3	68.3	–0.3	+2.9	–0.2	–0.2
	IND.ARTICLEAUTHOR†	13	3	13.0	0.0	+19.8	0.0	–1.6	0.0
	COLL†	5	—	0.0	—	0.0	—	0.0	—
PROD	MEDIA	58	61	66.3	47.9	+9.6	+2.0	+2.8	–10.1
	DOCTR†	3	5	28.6	10.0	+26.0	+56.2	+0.7	+14.7

Table 5: ΔF1 (pp) per fine-grained subtype for all subtypes, sorted by n (test set entity mention count, descending) within each parent type. MR = best per language (FR: Semantic $+150\%$; DE: Random $+150\%$); LCA = best `LLM-Majority+BERT` per language (FR: $+100\%$; DE: $+175\%$). †Test set too small for reliable evaluation.

can be attributed to two complementary properties. First, label consistency and distributional faithfulness to the original corpus are guaranteed by construction. Second, the near-identical performance of all MR variants (Section 6) suggests that surface form diversity alone drives the improvement, with semantic or temporal filtering providing no measurable benefit. This simplicity is practically significant: random mention replacement is sufficient and requires no additional resources beyond the training set itself.

Why LLM annotation degrades with scale. The monotonic degradation of `LLM-Strict` and `LLM-Majority` points to systematic annotation errors rather than random noise. We attribute this primarily to the complexity of the HIPE annotation guidelines: 21 fine-grained types and metonymic annotations make it difficult for LLMs to produce consistent annotations, leading to systematic type assignment errors and missed entities. The BERT validation step in `LLM-Majority+BERT` partially mitigates this effect by filtering pseudo-labels through a domain-aware model trained on gold data, but does not eliminate it entirely. Crucially, as more pseudo-labeled data is introduced, these residual errors accumulate in the training signal, explaining the continuous performance decline observed beyond moderate augmentation levels.

Practical implications and limitations. These results carry clear practical implications for historical NER practitioners: mention replacement is currently the more reliable augmentation strategy, requiring only a gold-annotated training set and yielding consistent gains across all experimental conditions. LLM-based augmentation may complement MR at moderate scale when validated by BERT, but should not be relied upon as a primary augmentation strategy.

This work is subject to several limitations: results are obtained with a single model architecture and dataset (HIPE-2020, French and German), and may not generalize to other historical languages, periods, or annotation schemes. The closed-world assumption of MR means it cannot introduce entity surface forms absent from the training data, potentially limiting its effectiveness in very low-resource settings.

8 Conclusion

This article presented a systematic comparison of two data augmentation strategies for historical NER on French and German historical newspapers. Mention replacement (MR) proved consistently effective and robust across all experimental conditions, yielding gains of +0.93 to +5.90 F1 points depending on the language and task granularity. LLM-based corpus annotation (LCA) was beneficial only

under strict validation and at moderate scale, with smaller gains of up to +1.99 F1 points; performance deteriorated as more pseudo-labeled data was added. These results show that simple label-preserving augmentation is currently the more reliable strategy, while LLM-based approaches remain limited by annotation noise, guideline complexity, and distributional mismatch with historical text.

Several directions for future work emerge from these findings. First, mention replacement could be improved through frequency-weighted candidate sampling, prioritizing rare entities and entity forms underrepresented in the training data. Second, LLM-based annotation would benefit from more sophisticated approaches, including prompt revision to reduce entity hallucination and improve handling of OCR noise, dynamic few-shot selection, and per-type inference. Third, hybrid strategies combining the robustness of mention replacement with the broader entity coverage of externally annotated silver data warrant investigation. Finally, both strategies should be evaluated on broader historical collections covering additional languages, periods, and annotation schemes, to assess the generalizability of the findings beyond the `hipe-2020` subset of the HIPE benchmark.

References

Hildelies Balk and Aly Conteh. 2011. IMPACT: Centre of Competence in Text Digitisation. In *Proceedings of the 2011 Workshop on Historical Document Imaging and Processing*, HIP '11, pages 155–160, Beijing, China, USA. ACM.

Ilaria Bartolini, Vincenzo Moscato, Marco Postiglione, Giancarlo Sperlì, and Andrea Vignali. 2023. Data augmentation via context similarity: An application to biomedical Named Entity Recognition. *Information Systems*, 119:102291.

Léa Blinière. 2026. Data augmentation strategies for historical named entity recognition. Master's thesis, École Polytechnique Fédérale de Lausanne, February.

Emanuela Boros, Ahmed Hamdi, Elvys Linhares Pontes, Luis Adrián Cabrera-Diego, Jose G. Moreno, Nicolas Sidere, and Antoine Doucet. 2020. Alleviating digitization errors in named entity recognition for historical documents. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 431–441, Online. Association for Computational Linguistics.

Estelle Bunout, Maud Ehrmann, and Frédéric Clavert, editors. 2023. *Digitized Newspapers - A New Eldorado for Historians? Reflections on Tools, Methods and Epistemology*. Studies in Digital History and Hermeneutics. De Gruyter Oldenbourg, Berlin, Germany.

Xiang Dai and Heike Adel. 2020. An Analysis of Simple Data Augmentation for Named Entity Recognition. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3861–3867, Barcelona, Spain (Online). International Committee on Computational Linguistics.

An Dao, Hiroki Teranishi, Yuji Matsumoto, Florian Boudin, and Akiko Aizawa. 2025. Overcoming data scarcity in named entity recognition: Synthetic data generation with large language models. In *Proceedings of the 24th Workshop on Biomedical Language Processing*, pages 328–340, Vienna, Austria. Association for Computational Linguistics.

Marten Düring, Matteo Romanello, Maud Ehrmann, Kaspar Beelen, Daniele Guido, Brecht Deseure, Estelle Bunout, Jana Keck, and Petros Apostolopoulos. 2023. Impresso Text Reuse at Scale. An interface for the exploration of text reuse data in semantically enriched historical newspapers. *Frontiers in Big Data*, 6.

Ehrmann, Watter, Romanello, Clematide, and Flückiger. 2020. *Impresso Named Entity Annotation Guidelines*.

Maud Ehrmann, Marten Düring, Clemens Neudecker, and Antoine Doucet. 2023a. Computational Approaches to Digitised Historical Newspapers (Dagstuhl Seminar 22292). *Dagstuhl Reports*, 12(7):112–179.

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

Maud Ehrmann, Matteo Romanello, Alex Flückiger, and Simon Clematide. 2020. Extended overview of CLEF HIPE 2020: Named entity processing on historical newspapers. In *CEUR Workshop Proceedings*, 2696. CEUR-WS.

Maud Ehrmann, Matteo Romanello, Sven Najem-Meyer, Antoine Doucet, and Simon Clematide. 2022. Extended Overview of HIPE-2022: Named Entity Recognition and Linking in Multilingual Historical Documents. In *Proceedings of the Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum*. CEUR-WS.

Steven Y. Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. 2021. A Survey of Data Augmentation Approaches for NLP. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 968–988, Online. Association for Computational Linguistics.

Antske Fokkens, Serge ter Braake, Ronald Sluijter, Paul Longley Arthur, and Eveline Wandl-Vogt, editors. 2018. *Proceedings of the Second Conference on Biographical Data in a Digital World 2017*, volume 2119 of *CEUR Workshop Proceedings*. CEUR-WS.org, Linz, Austria.

Paul Gooding. 2016. Exploring the information behaviour of users of Welsh Newspapers Online through web log analysis. *Journal of Documentation*, 72(2):232–246.

Ahmed Hamdi, Axel Jean-Caurant, Nicolas Sidère, Mickaël Coustaty, and Antoine Doucet. 2020. Assessing and Minimizing the Impact of OCR Quality on Named Entity Recognition. In *Digital Libraries for Open Knowledge*, Lecture Notes in Computer Science, pages 87–101, Cham. Springer International Publishing.

Yi Huang, Yuhan Gao, and Chengjuan Ren. 2025. A survey of data augmentation in named entity recognition. *Neurocomputing*, 651:130856.

Gaurav Kamath and Sowmya Vajjala. 2025. Does Synthetic Data Help Named Entity Recognition for Low-Resource Languages? In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 159–167, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

Marijn Koolen, Rik Hoekstra, Joris Oddens, Ronald Sluijter, Rutger Van Koert, Gijsjan Brouwer, and Hennie Brugman. 2023. The Value of Preexisting Structures for Digital Access: Modelling the Resolutions of the Dutch States General. *J. Comput. Cult. Herit.*, 16(1):1:1–1:24.

Andrianos Michail, Juri Opitz, Yining Wang, Robin Meister, Rico Sennrich, and Simon Clematide. 2025. Cheap Character Noise for OCR-Robust Multilingual Embeddings. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 11705–11716, Vienna, Austria. Association for Computational Linguistics.

Yuji Naraki, Ryosuke Yamaki, Yoshikazu Ikeda, Takafumi Horie, Kotaro Yoshida, Ryotaro Shimizu, and Hiroki Naganuma. 2024. Augmenting NER Datasets with LLMs: Towards Automated and Refined Annotation. *Preprint*, arXiv:2404.01334.

Clemens Neudecker. 2022. Cultural Heritage as Data: Digital Curation and Artificial Intelligence in Libraries. In *Proceedings of the Third Conference on Digital Curation Technologies (Qurator 2022)*, volume 3234 of *CEUR Workshop Proceedings*, Berlin, Germany. CEUR.

Clemens Neudecker and Apostolos Antonacopoulos. 2016. Making Europe's Historical Newspapers Searchable. In *Proceedings of the 12th IAPR Workshop on Document Analysis Systems (DAS)*, pages 405–410, Santorini, Greece. IEEE.

Juri Opitz, Corina Raclé, Emanuela Boros, Andrianos Michail, Matteo Romanello, Maud Ehrmann, and Simon Clematide. 2026. CLEF HIPE-2026: Evaluating Accurate and Efficient Person-Place Relation Extraction from Multilingual Historical Texts. *Preprint*, arXiv:2602.17663.

Uyen Phan and Nhung Nguyen. 2022. Simple semantic-based data augmentation for named entity recognition in biomedical texts. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 123–129, Dublin, Ireland. Association for Computational Linguistics.

Sophie Schneider, Ulrike Förstel, Kai Labusch, Jörg Lehmann, and Clemens Neudecker. 2025. ZEFYS2025: A German Historical Newspaper Dataset for Named Entity Recognition and Entity Linking. In *Proceedings of the 21st Conference on Natural Language Processing (KONVENS 2025): Long and Short Papers*, pages 48–58, Hannover, Germany. HsH Applied Academics.

Stefan Schweter. 2022. Hugging Face model - dbmdz/bert-base-historic-multilingual-cased.

Stefan Schweter and Johannes Baiter. 2019. Towards Robust Named Entity Recognition for Historic German. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 96–103, Florence, Italy. Association for Computational Linguistics.

Stefan Schweter, Luisa März, Katharina Schmid, and Erion Çano. 2022. hmBERT: Historical Multilingual Language Models for Named Entity Recognition. In *Proceedings of the Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum*, volume 3180 of *CEUR Workshop Proceedings*, pages 1109–1129, Bologna, Italy. CEUR.

Konstantin Todorov and Giovanni Colavizza. 2020. Transfer learning for named entity recognition in historical corpora. In *Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum*, volume 2696, pages 1–12, Thessaloniki, Greece. CEUR-WS.

Daniel van Strien, Kaspar Beelen, Mariona Ardanuy, Kasra Hosseini, Barbara McGillivray, and Giovanni Colavizza. 2020. Assessing the Impact of OCR Quality on Downstream NLP Tasks. In *Proceedings of the 12th International Conference on Agents and Artificial Intelligence*, pages 484–496, Valletta, Malta. SCITEPRESS - Science and Technology Publications.

Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, Guoyin Wang, and Chen Guo. 2025. GPT-NER: Named Entity Recognition via Large Language Models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4257–4275, Albuquerque, New Mexico. Association for Computational Linguistics.

Shibingfeng Zhang and Giovanni Colavizza. 2025. Named Entity Recognition of Historical Text via Large Language Model. *Preprint*, arXiv:2508.18090.

A HIPE Dataset Statistics

Tables 6–8 provide an overview of the `hipe-2020` dataset used in our experiments. Table 6 reports document, sentence, and entity mention counts per split and language. Table 7 shows the coarse-grained entity type distribution as a percentage of total mentions. Table 8 details the fine-grained subtype distribution, reporting mention counts and corpus-wide ranks across languages and splits.

Lang.	Split	Docs	Sents	Mentions
FR	Train	158	5,743	7,138
	Dev	43	1,244	1,746
	Test	43	1,462	1,642
DE	Train	103	3,472	3,655
	Dev	33	1,202	1,279
	Test	49	1,217	1,176

Table 6: Document, sentence, and entity mention counts per split for French and German `hipe-2020`.

Lang.	Split	LOC	PERS	ORG	PROD	TIME
FR	Train	43.5%	37.6%	12.2%	2.8%	4.0%
	Dev	44.4%	39.7%	9.2%	2.8%	3.9%
	Test	52.3%	32.7%	8.0%	3.7%	3.3%
DE	Train	47.6%	35.6%	9.9%	3.5%	3.4%
	Dev	46.0%	31.5%	13.1%	4.0%	5.5%
	Test	50.7%	28.3%	11.2%	5.6%	4.2%

Table 7: Coarse entity type distribution (% of total entities) per split for French and German `hipe-2020`.

B Training Configuration

`HIST-base` models are trained using the following hyperparameters, selected based on preliminary experiments (Boros et al., 2020):

Table 9: Training hyperparameters

Hyperparameter	Value
Batch size	32
Learning rate	5e-5
Optimizer	AdamW
ϵ (AdamW)	1e-8
Weight decay	0.0
Epochs	3
Gradient clipping	Max norm 1.0
Warmup steps	0
Learning rate schedule	Constant

Models are trained on NVIDIA A100 GPUs

		FR				DE			
		Train		Test		Train		Test	
Parent	**Subtype**	n	Rank	n	Rank	n	Rank	n	Rank
LOC FR: 3,106 / 858 DE: 1,741 / 596	ADM.TOWN	1,695	#2	450	#2	687	#2	257	#2
	ADM.NAT	648	#3	151	#3	564	#3	161	#3
	ADM.REG	382	#5	147	#4	199	#5	84	#5
	ORO	91	#10	19	#12	52	#11	5	#15
	PHYS.GEO	85	#11	28	#9	58	#10	14	#12
	PHYS.HYDRO	73	#12	23	#10	88	#9	29	#8
	ADM.SUP	63	#13	19	#12	42	#13	21	#10
	FAC	59	#14	18	#14	46	#12	14	#12
	PHYS.ASTRO	—	—	—	—	2	#17	10	#14
	ADD.PHYS	4	#18	—	—	2	#17	—	—
	ADD.ELEC	3	#19	—	—	—	—	—	—
	UNK	3	#19	3	#17	1	#19	1	#18
ORG FR: 868 / 132 DE: 362 / 132	ENT	599	#4	69	#5	200	#4	85	#4
	ADM	212	#7	43	#8	136	#6	29	#8
	ENT.PRESSAGENCY	57	#15	20	#11	26	#14	18	#11
PERS FR: 2,682 / 537 DE: 1,302 / 333	IND	2,553	#1	519	#1	1,288	#1	330	#1
	IND.ARTICLEAUTHOR	109	#9	13	#15	—	—	3	#17
	COLL	20	#17	5	#16	14	#16	—	—
PROD FR: 200 / 61 DE: 127 / 66	MEDIA	152	#8	58	#6	108	#8	61	#6
	DOCTR	48	#16	3	#17	19	#15	5	#15
TIME FR: 282 / 54 DE: 123 / 49	DATE.ABS	282	#6	54	#7	123	#7	49	#7

Table 8: Entity type distribution across train and test splits for French (FR) and German (DE) `hipe-2020`. For each fine-grained subtype, mention count n and corpus-wide rank are reported (by descending frequency, per language and split). Coarse parent totals are shown as train / test counts. "—" indicates subtypes absent from a given split.

(40GB VRAM). Training duration per configuration ranges from 15–20 minutes.

C LLM Annotation Prompt

All three LLMs received an identical system prompt specifying the annotation task and guidelines. The prompt covers: (1) task definition and entity types; (2) entity boundary rules, including modifiers, appositions, and coordination; (3) handling of OCR errors (annotate text as-is, without correction); (4) metonymic annotation criteria; and (5) output format requirements (JSON with character-level offsets).

The full prompt is available at https://github.com/impresso/impresso-named-entity-data-augmentation/tree/main/prompts.

The following excerpt illustrates the OCR handling instructions, which are particularly critical for historical text:

> **8. GESTION DES ERREURS OCR Principe fondamental : TOUJOURS annoter le texte TEL QUEL, SANS CORRECTION.**
>
> Les offsets (start/end) doivent pointer exactement sur les chaînes bruitées dans le texte.
>
> *Exemples :* "Léo Blanchard" → annoter la chaîne entière ; "L . Bridel" → annoter avec espaces ; "Ocan Atlantique" pour "Océan Atlantique" → annoter tel quel.

D Full Results per Configuration

Table 10 identifies the best augmentation variant and level for each task and language, corresponding to the peak performances reported in Table 3. Tables 11–16 then report ΔF1 (pp) vs. baseline for every configuration, across all tasks and languages. **Bold** indicates the best Mention Replacement and best LCA configuration in each table, as listed in Table 10.

Task	Lang.	Best MR	Best LCA
Coarse	FR	Semantic +150%	LLM-Maj+BERT +25%
	DE	Sem+Temp +125%	LLM-Maj+BERT +25%
Fine	FR	Semantic +150%	LLM-Maj+BERT +100%
	DE	Random +150%	LLM-Maj+BERT +175%
Meto	FR	Semantic +125%	LLM-Maj+BERT +25%
	DE	Random +150%	LLM-Maj+BERT +75%

Table 10: Best augmentation configuration (variant and level) for each cell in Table 3.

E Augmentation Data Statistics

Figures 2–5 report augmentation data statistics across all methods, levels, and languages. Figure 2 shows mean entity density (entities per sentence) and Figure 3 total entity volume ($\Delta\%$ vs. baseline) for all augmentation levels. Figure 4 shows entity type distribution evolution across augmentation levels, and Figure 5 reports the proportion of novel entities (surface forms absent from the original training data) introduced by each LCA configuration.

Level	Mention Replacement				LLM Corpus Annotation		
	Rand.	Sem.	Temp.	S+T	LLM-S	LLM-M	LLM-M+B
+25%	+0.53	+0.55	+0.20	−0.35	−3.14	−3.01	**+0.49**
+50%	+0.79	+0.36	+0.60	+0.49	−4.66	−4.16	+0.27
+75%	+0.74	+0.49	0.00	+0.16	−6.03	−5.89	+0.06
+100%	+0.50	+0.58	+0.20	+0.13	−6.50	−6.49	−0.20
+125%	−0.37	+0.63	+0.41	+0.37	−6.94	−6.87	−0.12
+150%	+0.09	**+0.93**	+0.24	+0.61	−7.91	−7.77	−0.76
+175%	+0.17	+0.14	−0.10	+0.54	−7.81	−7.98	−0.56
+200%	+0.14	+0.11	+0.36	+0.34	−7.75	−7.86	−0.21

Table 11: ΔF1 (pp) per configuration — **Coarse, French**. Baseline: 76.23 ± 0.53.

Level	Mention Replacement				LLM Corpus Annotation		
	Rand.	Sem.	Temp.	S+T	LLM-S	LLM-M	LLM-M+B
+25%	+0.71	+0.93	+0.55	+0.68	−3.62	−3.18	**+0.13**
+50%	+2.08	+1.60	+1.46	+2.03	−5.31	−5.24	−0.32
+75%	+2.38	+1.16	+1.78	+1.66	−6.59	−6.84	−0.01
+100%	+1.53	+2.06	+1.79	+2.90	−7.04	−8.14	−0.10
+125%	+2.41	+2.77	+1.64	**+3.20**	−7.91	−8.75	−0.77
+150%	+3.01	+2.09	+2.39	+2.35	−8.14	−9.35	−0.62
+175%	+2.36	+2.41	+2.00	+2.04	−8.76	−9.91	−0.58
+200%	+2.12	+1.54	+2.21	+2.17	−8.54	−9.49	−0.71

Table 12: ΔF1 (pp) per configuration — **Coarse, German**. Baseline: 71.67 ± 0.82.

Level	Mention Replacement				LLM Corpus Annotation		
	Rand.	Sem.	Temp.	S+T	LLM-S	LLM-M	LLM-M+B
+25%	+1.53	+1.59	+1.27	+0.84	−2.51	−2.02	+1.38
+50%	+2.21	+1.90	+1.74	+1.46	−3.80	−2.79	+0.90
+75%	+2.52	+2.43	+1.91	+1.88	−5.14	−4.50	+1.18
+100%	+2.39	+2.59	+2.38	+2.46	−5.15	−5.02	**+1.39**
+125%	+1.51	+2.68	+2.82	+2.65	−5.68	−5.28	+1.38
+150%	+2.03	**+3.15**	+2.52	+2.97	−6.41	−6.09	+0.81
+175%	+2.68	+2.53	+2.18	+2.99	−6.48	−6.08	+0.90
+200%	+2.09	+2.57	+2.65	+2.84	−5.96	−6.09	+1.24

Table 13: ΔF1 (pp) per configuration — **Fine-Grained, French**. Baseline: 69.26 ± 0.72.

Level	Mention Replacement				LLM Corpus Annotation		
	Rand.	Sem.	Temp.	S+T	LLM-S	LLM-M	LLM-M+B
+25%	+2.13	+2.46	+1.70	+2.03	−2.80	−2.17	+0.71
+50%	+3.69	+4.14	+3.28	+4.38	−3.50	−3.13	+1.10
+75%	+4.50	+3.82	+3.65	+4.48	−4.16	−4.18	+1.71
+100%	+4.42	+4.75	+4.04	+5.74	−4.43	−5.54	+1.78
+125%	+5.23	+5.80	+4.38	+5.67	−4.94	−5.52	+1.68
+150%	**+5.90**	+5.46	+5.28	+5.39	−5.13	−5.77	+1.56
+175%	+5.39	+5.40	+4.84	+5.44	−5.67	−6.18	**+1.99**
+200%	+5.34	+4.81	+5.65	+5.66	−5.22	−5.74	+1.97

Table 14: ΔF1 (pp) per configuration — **Fine-Grained, German**. Baseline: 63.21 ± 0.70.

Level	Mention Replacement				LLM Corpus Annotation		
	Rand.	Sem.	Temp.	S+T	LLM-S	LLM-M	LLM-M+B
+25%	+0.76	+0.68	+0.27	−0.23	−2.90	−2.81	**+0.79**
+50%	+0.94	+0.53	+0.80	+0.37	−4.54	−3.95	+0.46
+75%	+0.87	+0.72	+0.11	+0.23	−5.91	−5.70	+0.40
+100%	+0.72	+0.89	+0.33	+0.43	−6.41	−6.33	+0.08
+125%	−0.17	**+1.04**	+0.44	+0.61	−6.89	−6.82	+0.18
+150%	+0.17	+1.04	+0.37	+0.91	−7.89	−7.52	−0.50
+175%	+0.30	+0.18	+0.04	+0.85	−8.03	−7.95	−0.16
+200%	+0.17	+0.26	+0.48	+0.49	−7.73	−7.79	+0.07

Table 15: ΔF1 (pp) per configuration — **Metonymic, French**. Baseline: 73.59 ± 0.28.

Level	Mention Replacement				LLM Corpus Annotation		
	Rand.	Sem.	Temp.	S+T	LLM-S	LLM-M	LLM-M+B
+25%	+0.77	+1.17	−0.04	+0.81	−3.28	−3.06	+0.17
+50%	+2.02	+1.38	+1.39	+1.83	−5.10	−5.13	−0.21
+75%	+2.11	+0.83	+1.53	+1.46	−6.28	−6.87	**+0.27**
+100%	+1.09	+1.66	+1.66	+2.75	−6.71	−8.03	+0.11
+125%	+2.38	+2.45	+1.71	+2.78	−7.56	−9.05	−0.35
+150%	**+3.16**	+1.89	+2.42	+2.06	−7.80	−9.45	−0.39
+175%	+2.49	+2.25	+1.86	+1.80	−8.51	−9.87	−0.27
+200%	+2.14	+1.38	+2.22	+1.81	−8.22	−9.55	−0.52

Table 16: ΔF1 (pp) per configuration — **Metonymic, German**. Baseline: 67.62 ± 0.42.

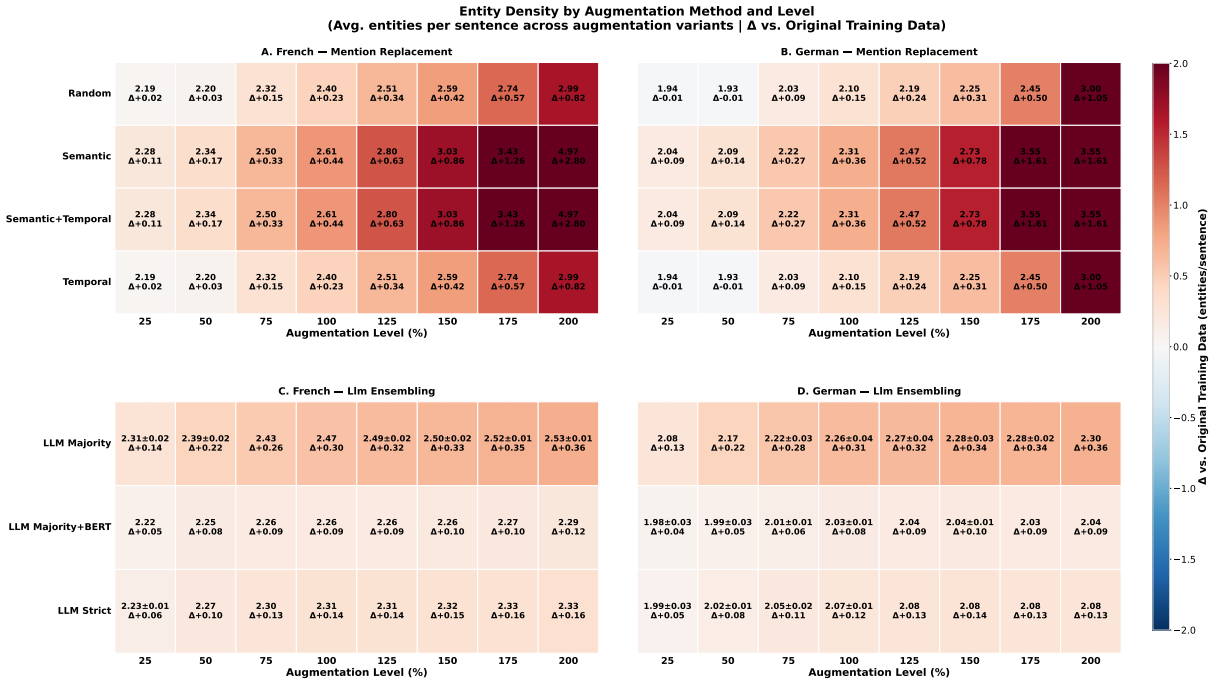

Figure 2: Mean entity density (entities/sentence, Δ vs. baseline) per augmentation method and level, for French (A, C) and German (B, D). Top: Mention Replacement; bottom: LLM Ensembling.

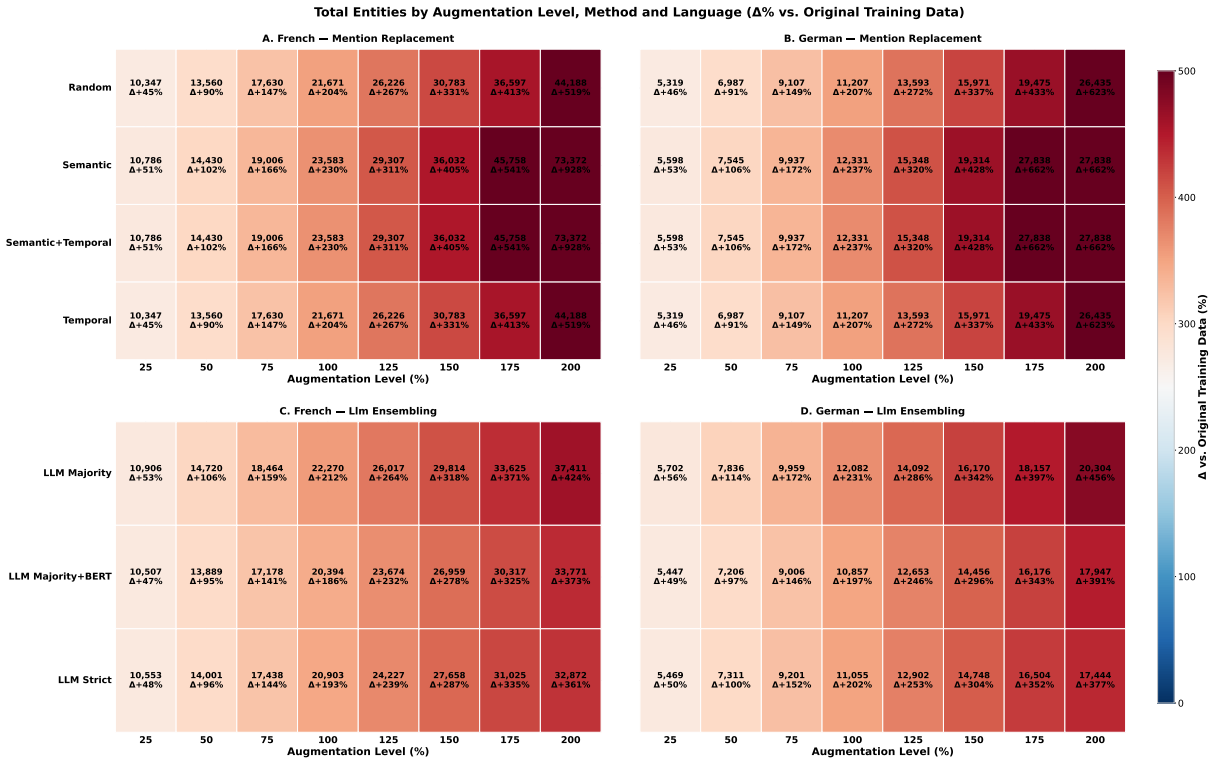

Figure 3: Total entity volume (Δ% vs. baseline) per augmentation method and level, for French (A, C) and German (B, D). Top: Mention Replacement; bottom: LLM Ensembling.

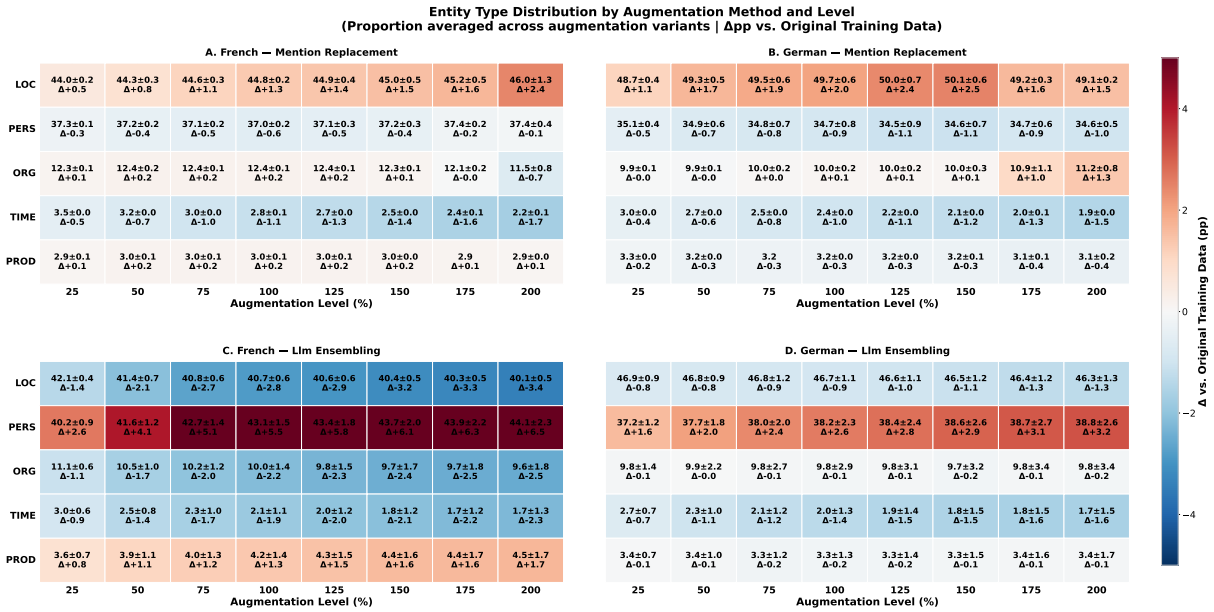

Figure 4: Entity type distribution evolution (Δpp vs. baseline) per augmentation method and level, for French (A, C) and German (B, D). Top: Mention Replacement; bottom: LLM Ensembling.

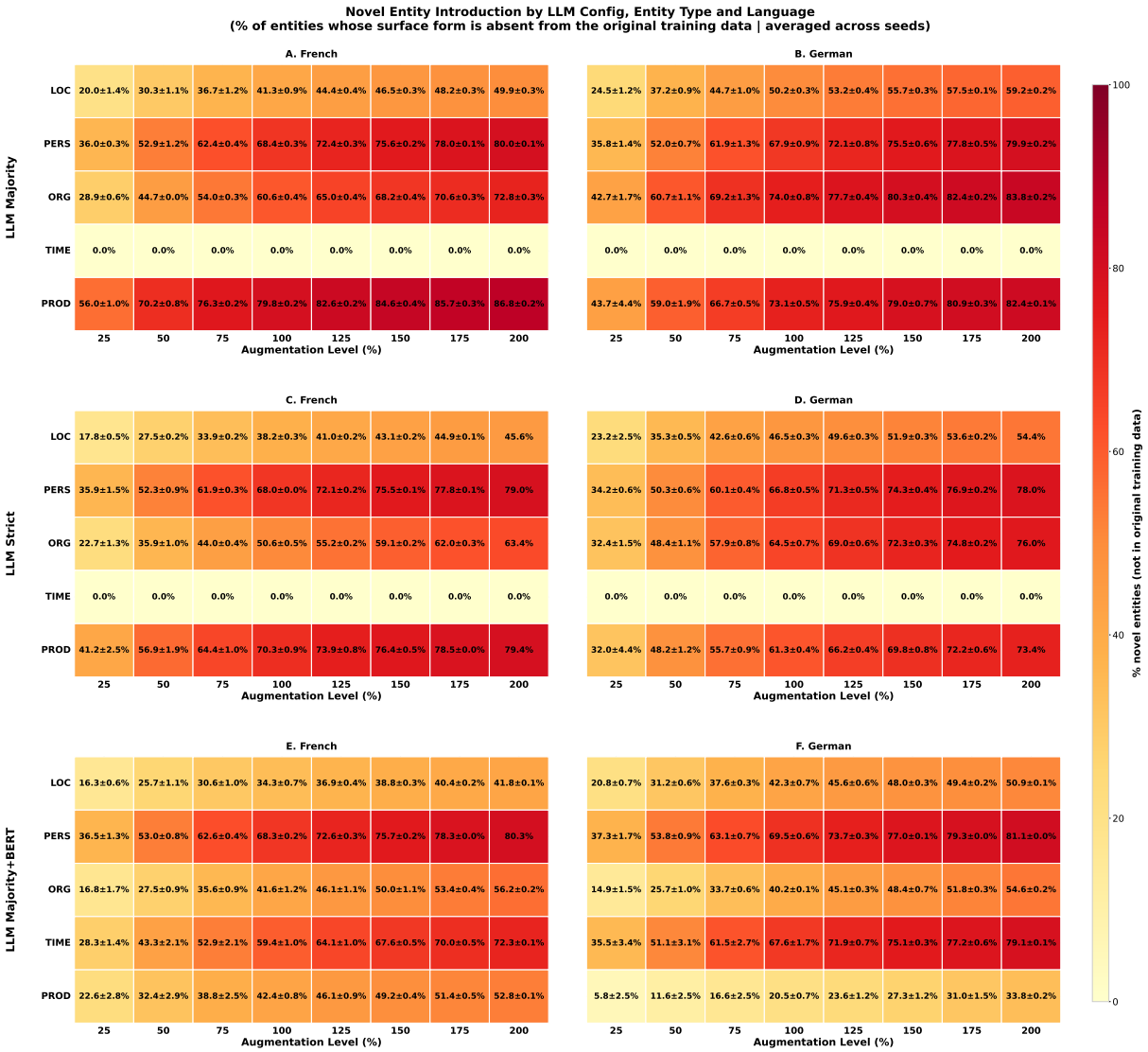

Figure 5: Proportion of novel entities (% of entity surface forms absent from the original training data) introduced by each LCA configuration (LLM-Majority, LLM-Strict, LLM-Majority+BERT), per entity type and augmentation level, for French (A, C, E) and German (B, D, F).