Reference-free Quality Estimation of Entity Recognition and Linking over OCRed Historical Texts

Anonymous ACL submission

Abstract

Named Entity Recognition (NER) and Named Entity Linking (NEL) are core tasks in entity extraction, yet their robustness is limited when applied to noisy documents, such as those generated by Optical Character Recognition (OCR) over historical documents. Although large language models (LLMs) have shown strong zero-shot and few-shot performance on NER and NEL tasks, prior work has largely focused on using LLMs as direct predictors. In this study, we investigate the feasibility of using LLMs as evaluators to estimate the quality of NER/NEL outputs in the absence of human-annotated ground truth. Focusing on OCRed texts where gold labels are scarce, we design and analyze supervised approaches to improve LLMs' quality estimation. We design supervised based methods to improve quality judgments from LLMs and systematically compare their alignment with gold labels. Experiments on the HIPE-2020 benchmark across English, French, and German languages demonstrate that fine-tuned LLMs provide reliable estimates of output quality. Our findings suggest that LLM-based evaluation can support quality control and enable evaluation in noisy settings. Our source code is publicly available at XXX[1]

1 Introduction

The digitization of historical documents has significantly advanced research in the humanities, social sciences, and archival studies by converting vast collections of handwritten and printed records into machine-readable formats. This transformation relies heavily on Optical Character Recognition (OCR) technologies, which enable automated text extraction from scanned images and facilitate large-scale search, and analysis. However, historical documents present substantial challenges for OCR due to diverse layouts, physical degradation, and low-resource languages, resulting in noisy and error-prone outputs (Nguyen et al., 2019).

With the increasing digitization of large-scale document collections across domains such as historical archives, government records, and scientific literature, there is a growing need to assess the quality of these digital texts—especially when they are used as input to downstream tasks like NER and NEL. However, in many real-world scenarios, the ground truth annotations for these tasks are missing, making direct evaluation of extraction quality difficult. Traditional text-level metrics such as Word Error Rate (WER) and Character Error Rate (CER), while commonly used to evaluate transcription or OCR quality, are not well-suited for assessing the impact on NER or NEL performance, as shown in work (Hamdi et al., 2023). These metrics fail to capture task-specific errors that affect entity identification and linking. This highlights the need for alternative, task-aware and reference-free evaluation methods that can better estimate the utility and reliability of digitized documents in the context of entity-centric NLP applications [2].

Despite these challenges, digitized historical corpora offer valuable opportunities for large-scale entity extraction (EE). NER and NEL can reveal patterns and relationships within unstructured historical texts despite facing difficulties due to orthographic variation, shifting grammar, and evolving entity references, which lead to degraded performance compared to modern datasets (Hamdi et al., 2023). Recent LLMs, including GPT-3.5, GPT-4 (Achiam et al., 2023), and LLaMA (Grattafiori et al., 2024), have been found to be successful in entity extraction (Tudor et al., 2025), yet their per-

[1] To be provided after paper publication.

[2] For example, NLB system (Goh, 2017) mistakes "MoST" (intended to reference the "Museum of Shanghai Toys") with the word "most" in general text, falsely implying that a museum is referenced in the text. In large-scale applications such as historical research or public information portals, such errors could distort timelines, misrepresent affiliations, or incorrectly suggest connections that never existed.

formance varies in low-resource or historical document settings (González-Gallardo et al., 2023b).

The main problem is that the scarcity of gold-standard annotations in historical domains limits supervised training and evaluation for information extraction. This scarcity is due to several challenges specific to historical texts, including OCR errors, archaic language, and non-standardized spelling, which complicate reliable annotation. Moreover, the lack of clear annotation guidelines, sparse existing labels, and the need for domain expertise make the creation of gold-standard datasets both difficult and resource-intensive.

In response to this limitation, we advocate a novel approach that reframes the problem: rather than relying on annotated data for training or evaluation, we fine-tune LLMs to act as quality estimators that assess the plausibility and correctness of NER and NEL outputs produced by external systems applied to OCRed historical documents. Instead of comparing outputs to gold annotations, our approach allows fine-tuned LLMs to internally assess extraction quality using linguistic and contextual signals learned during training. Our work explores whether LLMs can effectively serve as a proxy for the reliability assessment of information extraction from imperfect OCR data.

The contributions of our paper are three-fold:

- We are the first to formulate the task of using LLMs as quality estimators for NER and NEL outputs in OCRed historical documents, particularly in the absence of gold-standard annotations.
- We investigate the feasibility of estimating the quality of NER and NEL results without relying on human-annotated ground truth, employing a fine-tuning strategy with LLMs and a transformer-based language model (encoder-based model).
- We perform a comparative analysis of LLM-based quality estimators against conventional confidence measures, demonstrating that fine-tuned LLMs can more accurately capture contextual and historical uncertainties in EE.

Our results suggest that LLMs, when carefully adapted, can serve not only as extractors but also as effective evaluators of historical text processing quality, even across multiple languages. This capability paves the way for scalable, annotation-free methods in digital humanities research, enabling more inclusive and multilingual exploration of historical corpora where gold-standard annotations are scarce or nonexistent.

The remainder of this paper is organized as follows: Section 2 reviews related work on NER, NEL, and LLM-based output estimation. Section 3 introduces our problem formulation and presents the proposed modeling approach. Section 4 describes the experimental setup, including data construction, synthetic supervision, and evaluation protocols. Finally, Section 6 offers a discussion of the findings and concludes the paper.

2 Related work

Since the main focus of our paper is evaluating the performance of EE tasks, specifically NER and NEL, we first discuss these tasks and then review related work on estimation using LLMs.

NER tasks Recent work has explored the application of LLMs to NER, moving beyond traditional token- or span-level classification approaches (Nadeau and Sekine, 2007; Hanh et al., 2021; Liu et al., 2021; Sun et al., 2024; Moncla and Zeghidi, 2025). LLM-based methods require distinct strategies due to their generative nature and contextual reasoning abilities. Zhang et al. (2024) propose a hybrid framework that integrates a fine-tuned local NER model with an LLM via an uncertainty-aware linking mechanism: the local model handles low-uncertainty predictions, while high-uncertainty cases are delegated to the LLM for classification. Wang et al. (2023) reformulate NER as a text-to-text generation task, leveraging in-context learning and instruction prompting (Tran et al., 2024a) to extract entity mentions. To reduce hallucinated outputs, a self-verification step is introduced for post-hoc validation. In the context of historical documents, where OCR noise and linguistic variation are prevalent, recent studies have employed transformer-based models (Boroş et al., 2020; González-Gallardo et al., 2023a), while more recent efforts have begun to explore the applicability of LLMs to such settings (González-Gallardo et al., 2024). These studies highlight the need for robust adaptation strategies for noisy, low-resource historical corpora.

NEL tasks State-of-the-art (SOTA) NEL approaches are predominantly transformer-based (Wu et al., 2019; De Cao et al., 2022; Shavarani and

Sarkar, 2023; Yamada et al., 2022). De Cao et al. (2022) model NEL as a sequence-to-sequence generation task, where entities are produced token by token using an auto-regressive decoder. To ensure valid entity identifiers, they incorporate a constrained beam search guided by a prefix tree constructed from a knowledge base and introduce language marginalization techniques to enhance both training and inference. In contrast, Shavarani and Sarkar (2023) frame NEL as a token classification task, assigning entity links at the token level and aggregating predictions for efficient mention level linking. The use of LLMs for NEL is still emerging and primarily supports context enrichment or disambiguation in noisy settings (Vollmers et al., 2025).

Although LLMs show promising performance, their effectiveness diminishes when applied to historical OCRed documents (González-Gallardo et al., 2023b, 2024).

LLMs as quality estimators Beyond task performance, LLMs have been used as estimators and evaluators for various NLP tasks, including simulating human-like judgment (Li et al., 2024), machine-generated text prediction (Tran et al., 2024b), output quality estimation (Lee and Lee, 2023), and confidence or uncertainty modeling (Liu et al., 2024). For instance, Kocmi and Federmann (2023) show that LLMs can be prompted to assess machine translation quality without reference translations, achieving SOTA performance at the system level. This has been widely cited as a breakthrough in reference-free quality estimation. Similar uses of LLMs for scoring and critiquing output have been demonstrated in tasks such as question answering (Lee et al., 2024) and dialogue systems (Krumdick et al., 2025). These trends suggest that LLMs can serve not only as generators for NER/NEL outputs but also as meta-models that assess the correctness and reliability of other system predictions. However, such approaches remain underexplored for tasks like NER and NEL, particularly when applied to noisy or OCR-degraded inputs.

To address this gap, we investigate the use of LLMs as quality estimators for downstream NER and NEL systems operating on noisy OCR input, without relying on ground truth annotations. Our approach aligns with broader efforts to build NLP models that are robust, interpretable, and effective in low-resource, high-noise environments.

3 Problem Formulation

Let $x \in \mathcal{X}$ denote an OCRed input sentence, and let $e \in \mathcal{E}$ be the corresponding output of an EE system (e.g., predicted entity tags or entity links). The true performance metric (e.g., F1 score) for this input-output pair is denoted by $y \in [0, 1]$, and our objective is to learn a function $p_\theta : \mathcal{X} \times \mathcal{E} \rightarrow [0, 1]$, parameterized by θ, such that:

$$\hat{y} = p_\theta(x, e) \approx F_1(x, e) \quad (1)$$

This formulation casts the performance estimation problem as a regression task, where the model predicts the evaluation score directly from the input-output pair.

3.1 Analysis Model

We assume a regression-based approach for testing the performance of EE systems, with a focus on NER and NEL. Our goal is to approximate the evaluation metric (e.g., F1 score) of a model's output without requiring ground truth labels at inference time.

Our analysis model consists of three primary components: (1) joint input encoding, (2) feature projection, and (3) regression output. An overview is illustrated in Figure 1. The OCRed texts are first processed by the external NER/NEL model to generate entity recognition and linking results. These outputs, along with the original OCRed texts, are then integrated in the `Join` module to form a unified representation for feature extraction. The final output is the predicted F1 score of the task. The following sections provide a detailed breakdown of each step in the pipeline.

Input Representation The input to the model is constructed by combining the OCRed text sentence x with the EE system output e. We represent this combination as a serialized textual form:

$$\tilde{x} = \texttt{Join}(x, e)$$

where `Join` denotes a deterministic function for merging x and e. `Join` is used as simple text concatenation, while e includes EE predictions and EE confidences (probability). The dataset is enriched with synthetic data to ensure a broader range of sample variations. Further details can be found in Section 4.1.

Feature Encoding The combined input $\tilde{x}$ is passed to a pretrained language encoder $\texttt{Encoder}_\phi$,

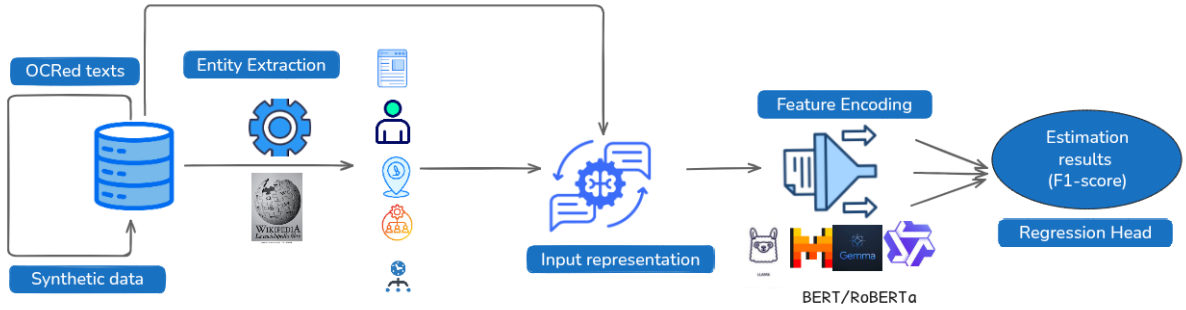


Figure 1: Overview of regression-based EE performance estimation.

which maps it to a fixed-dimensional latent representation:

$$h = \text{Encoder}_\phi(\tilde{x}) \in \mathbb{R}^d \tag{2}$$

where ϕ are the encoder parameters (e.g., from BERT, RoBERTa, or LLMs), and h can be extracted from a designated token (e.g., [CLS]) or by using mean/max pooling over token embeddings.

Regression Head A feed-forward linear projection layer transforms the encoded representation h into a scalar logit:

$$z = \mathbf{w}^\top h + b, \quad \mathbf{w} \in \mathbb{R}^d, \quad b \in \mathbb{R} \tag{3}$$

Output Activation To constrain the prediction $\hat{y}$ to lie in the interval $[0, 1]$, we apply the sigmoid function:

$$\hat{y} = \sigma(z) = \frac{1}{1 + \exp(-z)} \tag{4}$$

Training Objective The model is trained on a dataset of EE input-output pairs $\{(x_i, e_i, y_i)\}_{i=1}^N$, where each y_i is the gold evaluation score (e.g., F1) computed with reference annotations. We minimize a standard regression loss:

$$\mathcal{L}(\theta) = \frac{1}{N} \sum_{i=1}^{N} \ell(\hat{y}_i, y_i) \tag{5}$$

where $\hat{y}_i = p_\theta(x_i, e_i)$, and ℓ is a pointwise loss function, such as Mean Squared Error (MSE) or Mean Absolute Error (MAE):

$$\ell(\hat{y}, y) = (\hat{y} - y)^2 \quad \text{or} \quad |\hat{y} - y|$$

This approach enables label-free inference by generating performance estimates at test time without requiring ground truth labels. It supports task-agnostic representation, allowing the method to generalize across a wide range of EE tasks by jointly encoding both the input text and the system's output. Additionally, it facilitates model-agnostic evaluation, as it treats the output as an opaque signal, making the method compatible with any underlying EE system.

4 Experimental Setup

We conduct the analysis on the HIPE-2020 dataset, which was developed as part of a shared task on NER and NEL in historical documents. The dataset consists of three language-specific subsets: French (fr), German (de), and English (en), comprising newspaper articles from Switzerland, Luxembourg, and the United States, spanning the 19^{th} to 20^{th} centuries. Due to the limited number of annotated documents available for training dataset each subset, we generate synthetic data to improve model robustness. In the cross-lingual setting, we focus on HIPE-2020-en, which lacks a dedicated training set. Overall, this dataset contains 17,553 linked entity mentions annotated with a fine-grained label schema, including nested entities, mention components, and metonymic senses.

4.1 Dataset Construction

Due to the digitization process, the OCRed text of historical documents is often affected by various types of noise. To simulate such degradation and improve the model's robustness to real-world OCR errors, we adapt the approach proposed by Hamdi et al. (2023) to simulate common errors. Ground truth texts are first rendered as clean images and subsequently corrupted with noise. OCR engines (Tesseract and Google Cloud) are used to analyze the most common errors. Further details of the process, including the tools used, are provided in Appendix E. These common errors are used subse-

quently in the following perturbations:

Replacement: Random characters or words are substituted with visually or semantically similar alternatives, mimicking mis-recognitions.

Deletion: Characters or entire words are randomly removed, simulating cases in which parts of the text are lost or unreadable due to poor scan quality or document damage.

Insertion: Extraneous characters or words are inserted to reflect noise artifacts, such as smudging, overlapping lines, or layout issues that may cause OCR engines to hallucinate content.

Each perturbation is applied under three different conditions: (i) to entity tokens only, (ii) to the surrounding context of entities, and (iii) to all tokens in the text. These noise injection strategies allow us to systematically evaluate the model's robustness to varying levels and scopes of OCR-induced distortion, particularly in the context of named entity recognition and linking in historical texts. The distribution of training, validation, and test samples after pre-processing is provided in Table 1.

Split	fr	de	en
Original Train Set	5,532	3,310	N/A
Synthetic Train Set	71,916	43,030	N/A
Validation Set	1,227	1,165	N/A
Test Set	1,420	1,186	528

Table 1: Data distribution across splits (original, synthetic, validation, and test) for NER and NEL estimation on the HIPE-2020 dataset.

4.2 EE Model

NER We adopt the XLM-RoBERTa[3] model (XLM-R)[4]. This model is fine-tuned separately on the training split of each dataset. It serves as the external model for obtaining NER results, as it achieves the best results for the NER task across each dataset. Since the HIPE-2020 dataset is annotated at the document level and often exceeds the model's maximum token length, we segment documents into smaller units for training. Notably, we observe from the annotation files that entity labels can span sentence boundaries, with some annotations relying on context from preceding sentences. To preserve such dependencies, we split documents at the subgraph level, each subgraph consists of a few sentences. Specifically, a split occurs at sentence boundaries where the following line does not begin with an entity tag (i.e., not prefixed with `I-*`).

NEL For the NEL task, we adapt the multilingual mGENRE model De Cao et al. (2022) fine-tuned on five historical datasets (AJMC, HIPE-2020, TopRes19th, NewsEye, and SoNaR) available at the footnote link[5] as external model for obtaining NEL results. To ensure consistency, we apply the same document segmentation strategy used in NER to prepare data for NEL.

4.3 Regression Model

For the feature encoding model, we perform the analysis using two different approaches: LLMs-based and encoder-transformer-based. We use [CLS] token as output. For LLMs-based models, we use LoRA (Hu et al., 2022) with $r = 64$, $\alpha = 16$, and dropout 0.1. The confidence score is taken from the last softmax layer of the external NER and NEL model. For other BERT-based models, we perform full finetuning with a $lr = 1e - 5$.

NER

OCR: Gesandte der Vereinigten Staaten Amerikas verlangt Aufschluß über die schimpfliche Wegweisung zweier Prediger aus der Gemeinde Horgen . | NER prediction: O, O, B-loc, I-loc, I-loc, O, O, O, O, O, O, O, O, O, B-loc, I-loc, O | Confidence: 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00

NEL

OCR: Gesandte der [START] Vereinigten Staaten Amerikas [END] verlangt Aufschluß über die schimpfliche Wegweisung zweier Prediger aus der [START] Gemeinde Horgen [END] . | NEL mapping: _, _, Q30, Q30, Q30, _, _, _, _, _, _, _, _, _, _, Q68286, Q68286, _ | Confidence: 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00',

Figure 2: `Join` function output sample for NER and NEL tasks.

For the `Join` function in Section 3, we use the following format:
"OCR: ..." +"| Task results:..." +"| Confidence: ...".
The training samples for NER and NEL tasks can be found in Fig. 2.

[3] xlm-roberta-large-finetuned-conll03-english

[4] We use XLM-RoBERTa as it is a multilingual version of RoBERTa, pretrained on 100 languages, including those used in our experiments. Unlike RoBERTa, which is English-only, XLM-R has shown superior performance on non-English and zero-shot retrieval tasks (Tran et al., 2022; Tran, 2024).

[5] impresso-project/nel-mgenre-multilingual

Model	HIPE2020-de		HIPE2020-fr		HIPE2020-en	
	MSE (%)	MAE (%)	MSE (%)	MAE (%)	MSE (%)	MAE (%)
HIPE2020-de as the training set						
BERT (Devlin et al., 2019)	6.96	10.35	4.80	8.34	7.89	13.00
XLM-R (Conneau et al., 2019)	6.28	**9.57**	5.04	8.34	8.40	13.15
RobBERT (Delobelle et al., 2020)	6.88	10.17	4.72	**8.18**	7.84	12.55
LLaMA3.2 1B (Grattafiori et al., 2024)	6.06	9.85	4.67	8.68	6.51	**10.16**
LLaMA3.2 3B (Grattafiori et al., 2024)	6.33	10.11	**4.56**	9.07	6.12	11.71
Mistral 7B (Jiang et al., 2023)	7.23	10.28	5.50	8.70	8.74	13.43
Qwen2 7B (Yang et al., 2024)	5.92	9.99	5.02	9.65	7.16	13.14
Gemma 7B (Team et al., 2024)	6.55	10.11	5.05	8.83	7.26	12.42
LLaMA 8B (AI@Meta, 2024)	**5.62**	9.74	4.66	8.35	**5.88**	11.93
HIPE2020-fr as the training set						
BERT (Devlin et al., 2019)	6.38	10.21	4.60	8.31	6.51	11.43
XLM-R (Conneau et al., 2019)	6.77	10.24	4.62	7.99	6.90	11.24
CamemBERT (Delobelle et al., 2020)	**5.83**	10.56	4.44	8.41	6.45	11.51
LLaMA3.2 1B (Grattafiori et al., 2024)	6.60	10.00	4.62	7.95	6.98	11.88
LLaMA3.2 3B (Grattafiori et al., 2024)	6.29	**9.72**	4.66	**7.80**	6.39	10.99
Mistral 7B (Jiang et al., 2023)	6.24	10.04	4.56	8.18	6.34	11.05
Qwen2 7B (Yang et al., 2024)	6.09	10.08	**4.42**	8.13	**5.86**	10.81
Gemma 7B (Team et al., 2024)	6.94	10.13	5.10	8.39	6.46	**10.45**
LLaMA 8B (AI@Meta, 2024)	6.62	10.00	4.58	8.05	6.00	10.56

Table 2: Performance evaluation on NER tasks given HIPE2020-de and HIPE2020-fr as the training datasets, respectively. The highest score is highlighted in bold, and the second-highest is underlined.

5 Results

5.1 Comparative Analysis of Models

In this section, we evaluate the performance of various model types, including BERT-based models and LLMs. For the HIPE2020-fr dataset, we use CamemBERT (Martin et al., 2019), a variant of BERT pretrained specifically for French. For HIPE2020-de, we adopt RobBERT (Delobelle et al., 2020), a BERT-based model tailored for German. As a result of the ablation study, in this experiment, all models are fine-tuned using the synthetic version of each data set, EE (results and probability) and optimized with MSE loss. The results are summarized in Table 2 and Table 3.

NER Tasks As shown in Table 2, for the model trained on the HIPE2020-de dataset, LLaMA 8B consistently achieves the best performance across nearly all settings, particularly in the German dataset. The less parameter LLaMA 1B also shows strong in both monolingual and cross-lingual generalization, closely LLaMA 8B. Nonetheless, the performance differences between models are relatively minor, typically within 1%. This suggests that, under the current data constraints, overall model performance is effectively equivalent. One likely explanation for this limitation is the small size of the test set, which reduces the visibility of performance disparities.

While performance drops slightly in cross-lingual settings, the decrease is modest, indicating a certain degree of language dependency in the task. Notably, LLMs display greater robustness across languages, in contrast to BERT-based models, which suffer substantial performance degradation in out-of-language scenarios. Moreover, when evaluating on HIPE2020-en, models trained on HIPE2020-fr outperform those trained on HIPE2020-de an observation consistent with prior findings (Tran, 2024). Furthermore, BERT-based models exhibit higher error rates on English across both training settings. For example, XLM-R trained in German yields an MAE of 13.15%, compared to 10.16% for LLaMA 1B.

NEL Tasks BERT-based models generally achieve strong performance in monolingual or closely related cross-lingual settings, particularly when the target language is well-represented in the pretraining corpus. For instance, RobBERT trained on the HIPE2020-de dataset yields the lowest error on German (MSE: 2.61%, MAE: 5.17%). In contrast, CamemBERT performs significantly worse on the French dataset, suggesting that model–language alignment alone is insufficient for robust performance in all settings.

Interestingly, similar to observations in NER, we find that models trained on French transfer more effectively to English test sets than those trained on German. This suggests that the French training data may provide richer contextual signals that facilitate better generalization across languages.

Despite the enhanced cross-lingual capabilities of LLMs, even the best-performing models continue to exhibit relatively high error rates. This highlights the intrinsic challenges of historical entity recognition/linking in multilingual contexts and underscores the need for more robust architectures and richer, more diverse annotated datasets.

Model	HIPE2020-de		HIPE2020-fr		HIPE2020-en	
	MSE (%)	MAE (%)	MSE (%)	MAE (%)	MSE (%)	MAE (%)
HIPE2020-de as the training set						
BERT (Devlin et al., 2019)	2.96	5.77	8.94	13.49	4.15	8.47
XLM-R (Conneau et al., 2019)	2.85	5.28	9.75	13.65	**3.47**	**6.49**
RobBERT (Delobelle et al., 2020)	**2.61**	**5.17**	8.67	13.45	5.95	10.24
LLaMA3.2 1B (Grattafiori et al., 2024)	3.37	7.04	8.25	13.95	4.89	9.87
LLaMA3.2 3B (Grattafiori et al., 2024)	4.00	7.11	8.58	13.76	6.23	10.01
Mistral 7B (Jiang et al., 2023)	3.16	5.76	8.96	**12.89**	4.75	**8.10**
Qwen2 7B (Yang et al., 2024)	3.86	7.65	**7.89**	13.45	6.24	11.28
Gemma 7B (Team et al., 2024)	3.41	6.66	8.31	13.84	4.59	9.02
LLaMA 8B (AI@Meta, 2024)	3.03	6.80	**7.58**	13.23	4.81	9.74
HIPE2020-fr as the training set						
BERT (Devlin et al., 2019)	**2.84**	6.05	8.67	12.83	2.68	5.99
XLM-R (Conneau et al., 2019)	2.85	5.81	9.05	13.42	**2.66**	**5.76**
CamemBERT (Delobelle et al., 2020)	4.25	7.29	8.33	12.27	6.43	10.04
LLaMA3.2 1B (Grattafiori et al., 2024)	3.56	7.38	7.56	12.23	2.85	7.03
LLaMA3.2 3B (Grattafiori et al., 2024)	2.97	6.53	8.74	13.45	2.87	7.51
Mistral 7B (Jiang et al., 2023)	3.19	**5.71**	7.90	**11.18**	4.51	7.63
Qwen2 7B (Yang et al., 2024)	3.01	6.05	8.29	12.89	3.35	7.02
Gemma 7B (Team et al., 2024)	2.91	6.32	**7.27**	12.02	3.75	8.00
LLaMA 8B (AI@Meta, 2024)	3.19	6.25	7.80	12.31	3.27	6.79

Table 3: Performance evaluation on NEL tasks given HIPE2020-de and HIPE2020-fr as the training dataset, respectively. The highest score is highlighted in bold, and the second-highest is underlined.

5.2 Ablation study

In this part, we conduct experiments with several setup components using probability from the EE task, different loss functions (MAE, MSE), and synthetic data. In this setup, we use the same model, LLaMA 3.2 1B (Grattafiori et al., 2024)[6], utilizing synthetic data and the MAE objective function for a fair comparison. The overview results can be seen in Table 4.

Task	Setting	HIPE2020-de		HIPE2020-fr	
		MSE (%)	MAE (%)	MSE (%)	MAE (%)
NER	[1]	8.25	10.34	6.24	8.43
	[2]	7.02	9.85	5.83	8.49
	[3]	**6.06**	**9.85**	**4.62**	**7.95**
	[4]	9.19	11.2	7.33	9.57
NEL	[1]	5.27	7.19	11.19	14.21
	[2]	3.69	**6**	9.51	12.83
	[3]	**3.37**	7.04	**7.56**	**12.23**
	[4]	5.63	7.03	14.05	16.36

Table 4: Ablation study for NER and NEL tasks where [1] is *OCRed + EE results*; [2] *OCRed + EE (results + prob* ; [3] *OCRed + EE (results + prob) + MSE loss* ; [4] *OCRed + EE (results + prob) + MSE loss + W/o synthetic*.

For prediction NER task, the addition of probabilistic information (EE prob) leads to an approximately 1-2% improvement in both MSE and MAE scores, indicating a moderate but consistent benefit. This can be attributed to the fact that high-confidence predictions usually correspond to common words or clear entities, while low-confidence ones often indicate ambiguity or errors. As such, confidence serves as a valuable signal for models to account for uncertainty. Further replacing the standard configuration with MSE loss yields additional improvements over MAE loss, reinforcing the suitability of MSE as the optimization objective in this setting. Incorporating synthetic training data boosts performance further, achieving up to a 2% improvement compared to models using only original data.

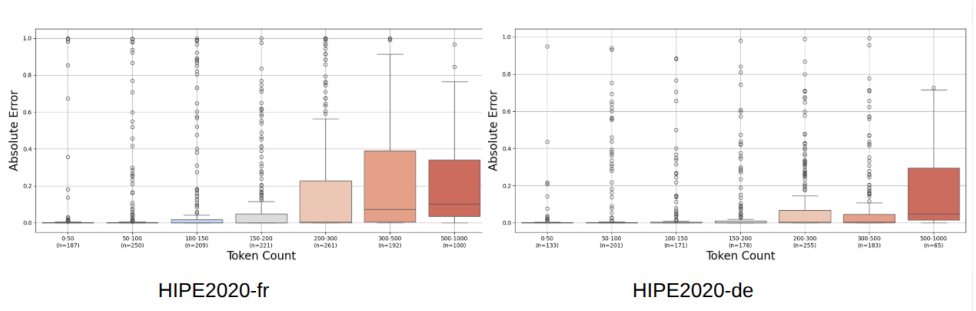
For the prediction of the NEL task, similar improvements can be observed when using synthetic data and optimizing with MAE loss; however, the resulting error patterns differ. The task exhibits lower error rates for HIPE2020-de, around 3% in MSE and 7% in MAE, but significantly higher error rates for HIPE2020-fr. This inconsistency can be attributed to the EE results of the external model, which is analyzed in the following section.

5.3 Analysis

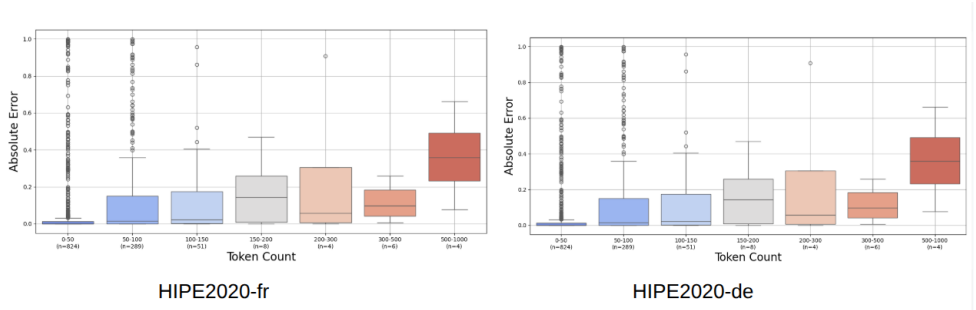
Effect of NER/NEL model Table 5 illustrates the relationship between the performance of NER/NEL models and their corresponding prediction F1 scores. It is evident that the relatively low performance of the NEL component in the HIPE2020-fr dataset leads to a lower overall prediction F1 score. Conversely, as the performance of the NEL model improves, particularly through fine-tuning the F1 score demonstrates a near-linear improvement. This trend highlights the dependency of the overall prediction accuracy on the effectiveness of the underlying NEL module.

Effect of input length Figure 3 presents the distribution of prediction errors in varying input lengths, measured by the number of tokens. The

[6] meta-LLaMA/LLaMA-3.2-1B



(a) NER



(b) NEL

Figure 3: Prediction Error Across Token Count Bins for predictions of NER (a) and NEL (b) with LLaMA 1B.

	HIPE2020-fr		HIPE2020-de	
	Strict	fuzzy	Strict	fuzzy
NER				
L3i (winner)	0.808	0.907	0.794	0.876
XML-R *	**0.828**	**0.917**	**0.798**	**0.885**
F1 Errors	~8%		~10%	
NEL				
L3i (winner)	0.602	0.620	0.506	0.525
mGENRE**	**0.661**	**0.661**	**0.863**	**0.863**
F1 Errors	~12%		~6%	

Table 5: NER and results (F1 score as in Ehrmann et al. (2022)) for HIPE2020-fr and HIPE2020-de.
* Fine-tuned on seperate HIPE2020-fr/HIPE2020-de training data.
** Fine-tuned on multiple historical dataset, evaluated on both HIPE2020-fr and HIPE2020-de.

lower error rates observed in short sentences can probably be attributed to the absence of named entities or to the overall simplicity of these inputs, which makes them easier for the model to handle. In contrast, we observe a notable increase in errors for inputs containing approximately 300 to 500 tokens. This may be due to the increased complexity and information density in longer sequences, which can overwhelm the model and lead to confusion. For inputs of medium length, error rates tend to be relatively low, with some exceptions that appear to result from a small number of outlier samples.

6 Conclusion

In this work, we investigate the use of LLMs as estimators for EE tasks, specifically NER and NEL, applied to historical OCRed texts. Rather than using LLMs purely as task solvers, our approach reframes the estimation process as a regression problem, leveraging the models to provide assessments of EE output quality in the absence of explicit ground truth. Such a role is especially valuable for uncertainty modeling and performance estimation in low-resource or cross-lingual contexts.

Our findings indicate that LLM-based estimation holds significant promise for assessing the quality of downstream EE tasks. Interestingly, results suggest that these tasks are relatively language independent, with LLMs demonstrating stable performance across different source and target languages. However, despite their generalization ability, even LLMs-based models still produce relatively high error rates, especially on noisy OCRed text, highlighting the persistent challenge of robust entity extraction in historical, multilingual settings.

Future work may explore leveraging agentic LLMs capable of self-assessing prediction confidence through function calling, prediction confidence in an end-to-end manner, facilitating improved uncertainty calibration.

7 Limitations

One limitation of the current prediction approach lies in the lack of interpretability inherent to LLM-based estimations. Since the models act as black boxes, it is difficult to understand or trace why a particular quality judgment is produced. This raises concerns about the transparency and reliability of the estimation process, especially in sensitive or decision-critical settings. One promising direction to address this is the use of reasoning models in the context of agentic scenarios, capable of generating not only outcome scores but also explanatory rationales. Such models could be further trained or aligned to produce consistent, high-quality estimations that might eventually serve as a proxy ground truth for benchmarking or guiding downstream tasks. Incorporating these models as judgment agents, rather than opaque predictors, could significantly enhance both the accountability and utility of LLM-based evaluation frameworks.

8 Ethics Statement

This work does not pose any ethical issues. All the data and tools used in this paper are publicly available under the CC BY-NC-SA 4.0 license. No private data or non-public information is used in this work.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

AI@Meta. 2024. Llama 3 model card.

Emanuela Boroş, 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.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Nicola De Cao, Ledell Wu, Kashyap Popat, Mikel Artetxe, Naman Goyal, Mikhail Plekhanov, Luke Zettlemoyer, Nicola Cancedda, Sebastian Riedel, and Fabio Petroni. 2022. Multilingual autoregressive entity linking. *Transactions of the Association for Computational Linguistics*, 10:274–290.

Pieter Delobelle, Thomas Winters, and Bettina Berendt. 2020. RobBERT: a Dutch RoBERTa-based Language Model. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3255–3265, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

Maud Ehrmann, Matteo Romanello, Sven Najem-Meyer, Antoine Doucet, Simon Clematide, Gulielmo Faggioli, Nicola Ferro, Alan Hanbury, and Martin Potthast. 2022. Extended overview of hipe-2022: Named entity recognition and linking in multilingual historical documents. In *CEUR Workshop Proceedings*, 3180, pages 1038–1063. CEUR-WS.

Rachael Goh. 2017. Using named entity recognition for automatic indexing.

Carlos-Emiliano González-Gallardo, Emanuela Boros, Edward Giamphy, Ahmed Hamdi, José G Moreno, and Antoine Doucet. 2023a. Injecting temporal-aware knowledge in historical named entity recognition. In *European Conference on Information Retrieval*, pages 377–393. Springer.

Carlos-Emiliano González-Gallardo, Emanuela Boros, Nancy Girdhar, Ahmed Hamdi, Jose G Moreno, and Antoine Doucet. 2023b. Yes but.. can chatgpt identify entities in historical documents? In *2023 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pages 184–189. IEEE.

Carlos-Emiliano González-Gallardo, Hanh Thi Hong Tran, Ahmed Hamdi, and Antoine Doucet. 2024. Leveraging open large language models for historical named entity recognition. In *International Conference on Theory and Practice of Digital Libraries*, pages 379–395. Springer.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Alexander Groleau, Kok Wei Chee, Stefan Larson, Samay Maini, and Jonathan Boarman. 2023. Augraphy: A data augmentation library for document images. In *Proceedings of the 17th International Conference on Document Analysis and Recognition (ICDAR)*.

Ahmed Hamdi, Elvys Linhares Pontes, Nicolas Sidere, Mickaël Coustaty, and Antoine Doucet. 2023. In-depth analysis of the impact of ocr errors on named entity recognition and linking. *Natural Language Engineering*, 29(2):425–448.

Tran Thi Hong Hanh, Antoine Doucet, Nicolas Sidere, Jose G Moreno, and Senja Pollak. 2021. Named entity recognition architecture combining contextual and global features. In *International Conference on Asian Digital Libraries*, pages 264–276. Springer.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. *arXiv preprint arXiv:2302.14520*.

Michael Krumdick, Charles Lovering, Varshini Reddy, Seth Ebner, and Chris Tanner. 2025. No free labels: Limitations of llm-as-a-judge without human grounding. *arXiv preprint arXiv:2503.05061*.

Bruce W Lee and Jason Hyung-Jong Lee. 2023. Prompt-based learning for text readability assessment. *arXiv preprint arXiv:2302.13139*.

Dongryeol Lee, Yerin Hwang, Yongil Kim, Joonsuk Park, and Kyomin Jung. 2024. Are llm-judges robust to expressions of uncertainty? investigating the effect of epistemic markers on llm-based evaluation. *arXiv preprint arXiv:2410.20774*.

Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. 2024. Llms-as-judges: a comprehensive survey on llm-based evaluation methods. *arXiv preprint arXiv:2412.05579*.

Zongxi Li, Xianming Li, Yuzhang Liu, Haoran Xie, Jing Li, Fu-lee Wang, Qing Li, and Xiaoqin Zhong. 2023. Label supervised llama finetuning. *arXiv preprint arXiv:2310.01208*.

Linyu Liu, Yu Pan, Xiaocheng Li, and Guanting Chen. 2024. Uncertainty estimation and quantification for llms: A simple supervised approach. *arXiv preprint arXiv:2404.15993*.

Zihan Liu, Yan Xu, Tiezheng Yu, Wenliang Dai, Ziwei Ji, Samuel Cahyawijaya, Andrea Madotto, and Pascale Fung. 2021. Crossner: Evaluating cross-domain named entity recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 13452–13460.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte de La Clergerie, Djamé Seddah, and Benoît Sagot. 2019. Camembert: a tasty french language model. *arXiv preprint arXiv:1911.03894*.

Ludovic Moncla and Hédi Zeghidi. 2025. Token and span classification for entity recognition in french historical encyclopedias. *arXiv preprint arXiv:2506.02872*.

David Nadeau and Satoshi Sekine. 2007. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26.

Thi-Tuyet-Hai Nguyen, Adam Jatowt, Mickaël Coustaty, Nhu-Van Nguyen, and Antoine Doucet. 2019. Deep statistical analysis of ocr errors for effective post-ocr processing. In *2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pages 29–38.

Hassan S Shavarani and Anoop Sarkar. 2023. Spel: Structured prediction for entity linking. *arXiv preprint arXiv:2310.14684*.

Wenjun Sun, Hanh Thi Hong Tran, Carlos-Emiliano González-Gallardo, Mickaël Coustaty, and Antoine Doucet. 2024. Lit: Label-informed transformers on token-based classification. In *International Conference on Theory and Practice of Digital Libraries*, pages 144–158. Springer.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, and 1 others. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Hanh Thi Hong Tran, Nishan Chatterjee, Senja Pollak, and Antoine Doucet. 2024a. Deberta beats behemoths: A comparative analysis of fine-tuning, prompting, and peft approaches on legallensner. In *Proceedings of the Natural Legal Language Processing Workshop 2024*, pages 371–380.

Hanh Thi Hong Tran, Matej Martinc, Antoine Doucet, and Senja Pollak. 2022. Can cross-domain term extraction benefit from cross-lingual transfer? In *International Conference on Discovery Science*, pages 363–378. Springer.

Hanh Thi Hong Tran, Tien Nam Nguyen, Antoine Doucet, and Senja Pollak. 2024b. L3i++ at semeval-2024 task 8: Can fine-tuned large language model detect multigenerator, multidomain, and multilingual black-box machine-generated text? In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 13–21.

Thi Hong Hanh Tran. 2024. *Neural approaches to automatic terminology extraction*. Ph.D. thesis, Université de La Rochelle; Institutu Jožef Stefan (Ljubljana).

Crina Tudor, Beata Megyesi, and Robert Östling. 2025. Prompting the past: Exploring zero-shot learning for named entity recognition in historical texts using prompt-answering LLMs. In *Proceedings of the 9th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2025)*, pages

216–226, Albuquerque, New Mexico. Association for Computational Linguistics.

Daniel Vollmers, Hamada Zahera, Diego Moussallem, and Axel-Cyrille Ngonga Ngomo. 2025. Contextual augmentation for entity linking using large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8535–8545.

Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, and Guoyin Wang. 2023. Gpt-ner: Named entity recognition via large language models. *arXiv preprint arXiv:2304.10428*.

Ledell Wu, Fabio Petroni, Martin Josifoski, Sebastian Riedel, and Luke Zettlemoyer. 2019. Scalable zero-shot entity linking with dense entity retrieval. *arXiv preprint arXiv:1911.03814*.

Ikuya Yamada, Koki Washio, Hiroyuki Shindo, and Yuji Matsumoto. 2022. Global entity disambiguation with bert. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3264–3271.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and 43 others. 2024. Qwen2 technical report. *Preprint*, arXiv:2407.10671.

Zhen Zhang, Yuhua Zhao, Hang Gao, and Mengting Hu. 2024. Linkner: linking local named entity recognition models to large language models using uncertainty. In *Proceedings of the ACM Web Conference 2024*, pages 4047–4058.

A NER and NEL results on HIPE2020-fr, HIPE2020-de dataset

In this section, we present the results of both NER and NEL tasks on the HIPE2020-fr and HIPE2020-de datasets.

A.1 NER

For NER, all chosen models are trained in a supervised fine-tuning manner, as listed in Table 6.

For the LLaMA model, we experiment with two different configurations: [1] *unmasked* - where we remove all attention masks from the transformer blocks (i.e., converting it into a bidirectional encoder-like structure), and [2] *causal* - where we retain the original causal masking. Our approach for supervised fine-tuning of LLaMA follows the method described in Li et al. (2023).

	Precision	Recall	F1 score
HIPE2020-fr			
L3i (winner)	78.6	83.1	80.8
XML-R	78.17	82.71	80.38
XML-R-large*	**81.68**	**85.21**	**83.40**
bert-large-cased	59.23	68.94	63.72
Unmask LLaMA2 7b	73.03	78.22	75.44
Unmask LLaMA3 3b	72.55	78.34	75.33
Causal LLaMA2 7b	51.24	62	56.11
Causal LLaMA3 3b	50.70	59.60	54.78
HIPE2020-de			
L3i (winner)	78.4	80.5	79.4
XML-R	67.05	73.97	70.33
XML-R-large*	**78.55**	**80.79**	**79.66**
bert-large-cased	54.05	56.19	55.09
Unmask LLaMA2 7b	56.61	61.73	59.06
Unmask LLaMA3 3b	67.71	74.16	70.19
Causal LLaMA2 7b	51.50	50.94	51.22
Causal LLaMA3 3b	42.8	49.03	45.48

Table 6: Performance comparison of different models for NER tasks. * represents *-finetuned-conll03-english*.

A.2 NEL

For NEL, we report the results of the pre-trained model for each data set in Table 7.

	Precision	Recall	F1 score
HIPE2020-fr			
NIL-BSL	20.9	20.9	20.9
SBB	70.7	51.5	59.6
L3i (winner)	60.2	60.2	60.2
Finetuning mGENRE	**66.1**	**66.1**	**66.1**
HIPE2020-de			
NIL-BSL	48.1	31.4	38.0
SBB	60.3	40.5	50.6
L3i (winner)	48.1	48.1	48.1
Finetuning mGENRE	**86.3**	**86.3**	**86.3**

Table 7: Performance comparison of different models for NEL task on HIPE2020-fr

B Synthetic sample results

In Fig. 4, we show synthetic samples with different strategies: [1] for entity tokens only, [2] for the surrounding context of entities, and [3] for all tokens in the text (random).

OCRed

— Le conseil supérieur des travaux publics , à Rome , a approuvé la construction d ' une voie ferrée Ortafrontière suisse (Brissago) par la rive droite du lac Majeur , soit par Pallanza , Intra et Canobbio .

Only entity

— Le conseil supérieur des travaux publics , à Kone , a approuvé la construction d ' une voie ferrée Gciulcprifièco suisse (Brissago) par la rive droite du lac Majeur , soit par Pallanza , Intra et Canobbio .

Surrounding entity

— Le conseil supérieur des travaux publics f l Bunc , a approuvé la construction d ' une u0lo lcnnyc Ortafrontière cntaao v Brissago) osc la rive alc6tio lv lac wuicnc k c6t1 osn Fd!sres w Jri1nu c1 iuo6htlu .

Random

— Le enhcot! supérieur des travaux publics , à Kqrno , a soonpnyj !u construction d g noc u0to loccic Gniutc0riftlcc suisse f Brissago) par !s nyc alcplic du !ur csiovc v soit par Pallanza , Intra et tsoethjq k

Figure 4: Illustration of synthetic data generated using different strategies. The red words indicate changes from the ground truth text.

C English Translation of the German Figure

Figure 5 provides an English translation of Figure 2 from the main text, originally presented in German.

NER

OCR: Envoy of the United States America demands clarification about the shameful expulsion of two preachers from the Horgen congregation. | NER prediction: O, O, B-loc, I-loc, I-loc, O, O, O, O, O, O, O, O, O, B-loc, I-loc, O | Confidence: 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00

NEL

OCR: Envoy of the [START] United States America [END] demands clarification about the shameful expulsion of two preachers from the [START] Horgen congregation [END]. | NEL mapping: _, _, Q30, Q30, Q30, _, _, _, _, _, _, _, _, _, _, Q68286, Q68286, _ | Confidence: 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00',

Figure 5: `Join` function output sample for NER and NEL tasks. (English translation)

D Visualization prediction

The worst-case predictions for the NER and NEL tasks are demonstrated from Fig. 6 to 9.

E Synthetic common OCR errors

To simulate realistic noise conditions, we first converted clean text corpora into synthetic document images (referred to as clean images). These images were then degraded using AuGraphy (Groleau et al., 2023)[7] to introduce document-level distortions. The following parameters were applied to simulate degradations: brightness texturization, dirty rollers (applied twice to enhance streaking artifacts), subtle noise, lighting gradient, and low-ink periodic lines. These effects mimic real-world conditions such as uneven lighting, faded ink, paper aging, and scanner noise. The degraded images were subsequently processed by multiple OCR engines to extract noisy text. This approach enabled us to capture a wide range of OCR error types and analyze their frequency and distribution. An illustration of the text-to-image degradation process is shown in Fig. 10.

[7] https://github.com/sparkfish/augraphy

```
================================================================================
Sample:
 OCR: In der Wiener Ausgabe des „ Völkischen Beobachters " wird bekanntgegeben , daß alle kirchlichen Schulen einschließlich der römisch -
katholischen Parochialschulen nach Beendigung der Sommerferien nicht wieder eröffnet werden . | NER prediction: 0, 0, 0, 0, 0, 0, B-prod, I
-prod, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0 | Confidence: 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.
00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 0.99
True F1 score: 0.0
Prediction F1 score: 0.9994738698005676

================================================================================
Sample:
 OCR: Herr Pescatore erhielt seitens des kath . Wahlcomites über 500 Stimmen , unb ging mit absoluter Stimmenmehrheit ( 939 St . ) aus dem
Wahlkampfe hervor . | NER prediction: B-pers, I-pers, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0 | Confidenc
e: 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.0
0, 1.00, 1.00, 1.00, 1.00
True F1 score: 0.0
Prediction F1 score: 0.998557984828949

================================================================================
Sample:
 OCR: Der „ Völkische Beobachter " erklärt zu dieser Maßnahme , der nationalsozialistische Staat stehe grundsätzlich auf dem Standpunkte ,
daß die Erziehung der Jugend eine Angelegenheit des Staates sei und diesem völlig vorbehalten bleiben müsse . | NER prediction: 0, 0, B-pro
d, I-prod, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0 | Confidence: 1.00, 1.00, 1.00,
1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 0.82, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00,
1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00
True F1 score: 0.0
Prediction F1 score: 0.9984606504440308

================================================================================
Sample:
 OCR: Nochenschau . | NER prediction: B-loc, 0 | Confidence: 1.00, 1.00
True F1 score: 0.0
Prediction F1 score: 0.9979233741760254

================================================================================
Sample:
 OCR: Die Alabamafrage , die mehr und mehr in den Vor ¬ dergrund tritt , besteht aus zwei Forderungen . | NER prediction: 0, B-loc, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0 | Confidence: 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.0
0, 1.00, 1.00, 1.00, 1.00, 1.00
True F1 score: 0.0
Prediction F1 score: 0.9974647760391235
```

Figure 6: Visualization errors on prediction NER on HIPE2020-de

```
================================================================================
Sample:
 OCR: DE LA FEUILLE OFFICIELLE EXTRAIT du jeudi 6 mai 1858 . | NER prediction: 0, 0, 0, 0, 0, B-time, I-time, I-time, I-time, I-time, 0 | C
onfidence: 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00
True F1 score: 0.0
Prediction F1 score: 0.9995765089988708

================================================================================
Sample:
 OCR: demandé chez Gallimard s ' il était possible de lui rendre visite . | NER prediction: 0, 0, B-pers, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0 | Co
nfidence: 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00
True F1 score: 0.0
Prediction F1 score: 0.9990845918655396

================================================================================
Sample:
 OCR: MONTES . | NER prediction: B-loc, 0 | Confidence: 1.00, 1.00
True F1 score: 0.0
Prediction F1 score: 0.9990519881248474

================================================================================
Sample:
 OCR: Faire sortir le catholicisme de son isolement , tel serait le but du concil œcuménique . | NER prediction: 0, 0, 0, B-prod, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0 | Confidence: 1.00, 1.00, 1.00, 0.99, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00
True F1 score: 0.0
Prediction F1 score: 0.9987344145774841

================================================================================
Sample:
 OCR: Un accès de désespoir lui a fait chercher la mort dans les Ilots du Rhin . | NER prediction: 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, B-lo
c, I-loc, I-loc, 0 | Confidence: 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00
True F1 score: 0.0
Prediction F1 score: 0.9979498982429504
```

Figure 7: Visualization errors on prediction NER on HIPE2020-fr

```
================================================================================
Sample:
 OCR: [START] ESPAGNE [END] . | NEL mapping: Q29, _ | Confidence: 1.00, 1.00
True F1 score: 0.0
Prediction F1 score: 0.9997603297233582

================================================================================
Sample:
 OCR: [START] Berne [END] , i " octobre . | NEL mapping: Q70, _, _, _, _, _ | Confidence: 1.00, 1.00, 1.00, 1.00, 1.00, 1.00
True F1 score: 0.0
Prediction F1 score: 0.9996993541717529

================================================================================
Sample:
 OCR: [START] BERNE [END] . | NEL mapping: Q70, _ | Confidence: 1.00, 1.00
True F1 score: 0.0
Prediction F1 score: 0.9996663331985474

================================================================================
Sample:
 OCR: [START] BERNE [END] . | NEL mapping: Q70, _ | Confidence: 1.00, 1.00
True F1 score: 0.0
Prediction F1 score: 0.9996663331985474

================================================================================
Sample:
 OCR: Grâce aux efforts de la [START] Confédération [END] et des cantons , personne n ' a souffert de la faim pendant la guerre . | NEL map
ping: _, _, _, _, _, Q39, _, _, _, _, _, _, _, _, _, _, _, _, _, _, _, _, _, _, _, _, _ | Confidence: 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.0
0, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00
True F1 score: 0.0
Prediction F1 score: 0.9995506405830383
```

Figure 8: Visualization errors on prediction NEL on HIPE2020-de

```
================================================================================
Sample:
 OCR: [START] ESPAGNE [END] . | NEL mapping: Q29, _ | Confidence: 1.00, 1.00
True F1 score: 0.0
Prediction F1 score: 0.9997603297233582

================================================================================
Sample:
 OCR: [START] Berne [END] , i " octobre . | NEL mapping: Q70, _, _, _, _, _ | Confidence: 1.00, 1.00, 1.00, 1.00, 1.00, 1.00
True F1 score: 0.0
Prediction F1 score: 0.9996993541717529

================================================================================
Sample:
 OCR: [START] BERNE [END] . | NEL mapping: Q70, _ | Confidence: 1.00, 1.00
True F1 score: 0.0
Prediction F1 score: 0.9996663331985474

================================================================================
Sample:
 OCR: [START] BERNE [END] . | NEL mapping: Q70, _ | Confidence: 1.00, 1.00
True F1 score: 0.0
Prediction F1 score: 0.9996663331985474

================================================================================
Sample:
 OCR: Grâce aux efforts de la [START] Confédération [END] et des cantons , personne n ' a souffert de la faim pendant la guerre . | NEL map
ping: _, _, _, _, _, Q39, _, _, _, _, _, _, _, _, _, _, _, _, _, _, _, _, _, _, _, _, _ | Confidence: 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.0
0, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00
True F1 score: 0.0
Prediction F1 score: 0.9995506405830383
```

Figure 9: Visualization errors on prediction NEL on HIPE2020-fr

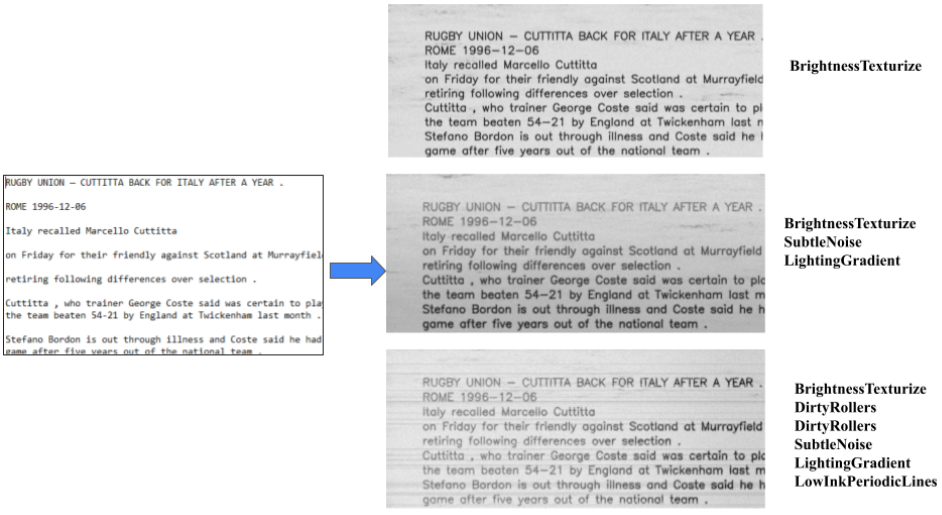


Figure 10: Illustration of the synthetic degradation pipeline: clean text is rendered as an image and corrupted with various degradations.