Reference-free Evaluation at Inference for NER/NEL over OCRed Historical Texts

Tien-Nam Nguyen[1], **Adam Jatowt**[2], **Ahmed Hamdi**[1], **Mickaël Coustaty**[1]
Thi-Hong-Hanh Tran[3], **Antoine Doucet**[1,4]
[1]L3i, La Rochelle University, La Rochelle, France
[2]University of Innsbruck, Innsbruck, Austria
[3]Arkhn, Paris, France
[4]Faculty of Computer and Information Science, University of Ljubljana, Slovenia
{tnguye28, ahmed.hamdi, mickael.coustaty, antoine.doucet}@univ-lr.fr,
adam.jatowt@uibk.ac.at, hanh.tran@arkhn.com

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is limited when applied to noisy documents, such as those generated by Optical Character Recognition (OCR) over historical documents.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 rather than evaluating extraction performance. In this study, we explore the feasibility of using LLMs as learned evaluators to estimate the quality of NER/NEL outputs, especially in settings where human-annotated references are unavailable at inference time. We propose supervised approaches that fine-tune LLMs to predict quality scores based on training data with gold annotations, enabling reference-free quality estimation once trained. 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.

Keywords: Large Language Models, Named Entity Recognition, Named Entity Linking, Evaluation

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The digitization of historical documents has significantly advanced research in the humanities, social sciences, and archival studies by converting vast collections of handwritten and printed records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 OCR technologies, which enable automated text extraction from scanned images, facilitate large-scale search and analysis. However, historical documents present substantial challenges for OCR due to diverse layouts, spelling variations, physical degradation, and low-resource languages, resulting in noisy and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 the 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 (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.

Named entities constitute a key mechanism for organizing and accessing information (Guo et al., 2009; Gefen, 2014; Chiron et al., 2017) through entity-based indexing and knowledge graph linking. Thus, accurate estimation of NER and NEL performance becomes essential. Such estimators provide insights into how effectively extracted entities support document indexing and retrieval, even when gold-standard annotations are unavailable.

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, but their performance remains degraded due to orthographic variation, shifting grammar, and evolving entity references (Hamdi et al., 2023). Recent LLMs, such as GPT-3.5, GPT-4 (Achiam et al., 2023), and LLaMA (Grattafiori et al., 2024), demonstrate an initial capability for entity extraction, particularly in zero- or few-shot settings (Tudor et al., 2025). However, their performance typically remains below that of fully supervised approaches, especially in low-resource or historical document contexts (González-Gallardo et al., 2023b).

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 893–904
11–16 May 2026.

The central challenge is the scarcity of gold-standard annotations in historical domains, which constrains both supervised training and evaluation for information extraction. This scarcity arises from OCR errors, archaic language, and non-standardized spelling, as well as the need for domain expertise and costly annotation processes. These factors make the creation of reliable evaluation benchmarks difficult, especially in low-resource settings. In this context, there is an increasing gap between the rapidly growing volumes of digitized text and the limited availability of annotated resources (Ehrmann et al., 2023). At the same time, NER and NEL play a crucial role in indexing, searching, and retrieval, meaning that unreliable entity extraction directly undermines the usability of large-scale digital archives (Chiron et al., 2017). This makes it essential to develop scalable methods that can approximate task-specific performance without costly annotation campaigns. Building an estimator that can predict extraction quality fills this gap, providing a practical tool for monitoring, filtering, and prioritizing data in real-world archival and information retrieval pipelines.

In response to this limitation, we propose a novel evaluation paradigm: rather than relying on annotated data, we fine-tune LLMs to act as quality estimators that assess the plausibility and correctness of NER and NEL outputs from OCRed historical documents. Instead of comparing outputs to gold annotations, our approach enables fine-tuned LLMs to internally judge extraction quality using linguistic and contextual signals learned during training. This reframing offers a task-aware, reference-free evaluation method that is directly applicable to annotation-scarce settings. Beyond historical corpora, our approach has broader implications for entity-centric NLP in other low-resource domains where labeled data is expensive or unavailable, aligning with current surveys (Ehrmann et al., 2023) highlighting the urgent need for scalable evaluation methods in data-scarce environments.

In addition to advancing evaluation methods, our work provides practical resources for the research community. The code and datasets used in our experiments are publicly available at our GitLab repository, enabling full reproducibility of the reported results and facilitating application of our methods to historical and low-resource corpora [1].

The contributions of our paper are three-fold:

- We are the first to propose finetuning LLMs as quality estimators for NER and NEL outputs on OCRed historical documents, enabling reference-free evaluation at inference without requiring gold-standard annotations.
- We investigate estimating the quality of NER and NEL outputs without relying on human-annotated ground truth, using fine-tuned LLMs and encoder-based transformer models to enable reliable reference-free evaluation.
- We perform a comparative analysis of LLMs-based quality estimators against conventional confidence measures, showing that fine-tuned LLMs capture contextual and historical uncertainties in EE more accurately.

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, inference-free methods in digital humanities research, enabling more inclusive and multilingual exploration of historical corpora where gold-standard annotations are scarce or nonexistent.

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

[1] gitlab/reference-free

(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 explore using LLMs as quality estimators for downstream NER and NEL systems on noisy OCR inputs, enabling evaluation without ground-truth annotations. Our approach contributes to broader efforts to develop NLP models that are robust, interpretable, and effective in low-resource, high-noise settings.

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 $p_\theta : \mathcal{X} \times \mathcal{E} \to [0, 1]$, parameterized by θ, such that:

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

This formulation casts the performance estimation problem as a regression task, where the model predicts the evaluation score directly from the input–output pair. By predicting document-level or sentence-level quality scores, our approach can identify low-quality OCRed documents that may require re-digitization or correction, enabling targeted improvements to downstream tasks such as NER and NEL. While entity-level quality estimation could in principle provide more granular feedback, we frame the task as a regression over aggregated scores to obtain more stable training signals and to better capture the overall usability of extracted entities. Moreover, our goal is to estimate holistic system performance in the absence of reference labels at inference time. The regression formulation thus aligns with our objective of producing consistent, interpretable quality estimates that support large-scale quality control.

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

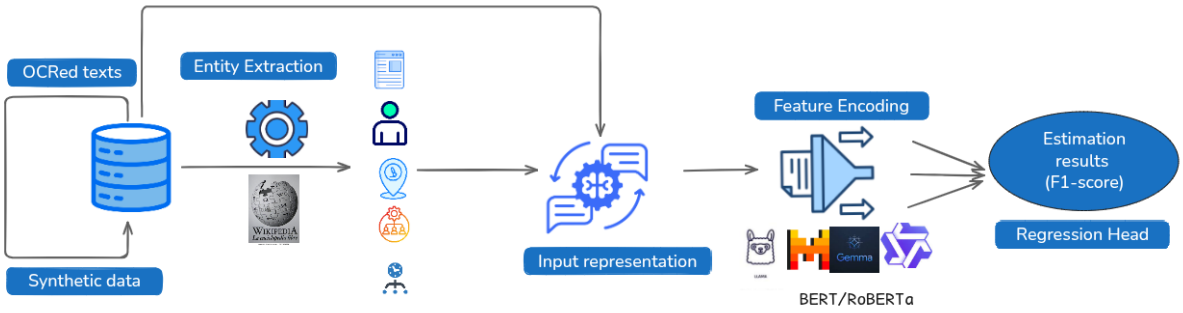

Figure 1: Overview of regression-based EE performance estimation.

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 $\texttt{Encoder}_\phi$, which maps it to a fixed-dimensional latent representation:

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

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

While the proposed method is general and not specific to OCR or historical documents, we focus our evaluation on the HIPE-2020 dataset, developed as part of a shared task on NER and NEL in historical documents (Ehrmann et al., 2022a). 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. Ideally, we would process entire document, but the effective input length for our model is unknown in practice. To handle this, we adopt a simple token-based chunking strategy: each document is split into sub-sequences that do not exceed a predefined maximum number of tokens, which may vary depending on the tokenizer and model used, regardless of sentence boundaries. This ensures model input constraints are respected while still capturing relevant context. Furthermore, due to the limited number of samples, we generate synthetic data to improve the robustness of the model. In the cross-lingual setting, we test with the English set which does not have a dedicated training set.

4.1. Dataset Construction

Our experiments are conducted on the HIPE-2020, a benchmark dataset for NER and NEL on historical

newspapers. Importantly, HIPE already contains naturally occurring OCR noise such as at different place: at named entity or surrounding named entity. The more details of noise perception of each dataset can read in (Ehrmann et al., 2022b). Our synthetic noise generation therefore complements existing OCR distortions and enables controlled robustness evaluation under progressively increased degradation levels.

We build upon the analysis of common OCR error patterns presented in Hamdi et al. (2023) and Jaud et al. (2025), and use these patterns to guide our synthetic noise generation process. Further details of the process, including the noise parameters and noise level used, are provided in Appendix 10.1.

These errors are used subsequently in the following three perturbations: **replacement**: Random characters or words are substituted with visually or semantically similar alternatives, mimicking misrecognitions); **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.

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 NER and NEL in historical texts. The distribution of training, validation, and test documents (original and synthetic) is provided in Table 1.

To test the generalizability of our performance estimation model, we conduct one more test samples that has not been fine-tuned on the HIPE-2020 dataset. The predictions from this out-of-domain model are used as inputs to our evaluator, simulating a realistic scenario where annotated data is limited or unavailable. In this setup, the test set naturally contains two types of predictions: high-confidence predictions from entities similar to those seen in training, and low-confidence predictions resulting from the out-of-domain model encountering unseen or difficult entities. This allows us to assess how well our evaluator can estimate the quality of model outputs under varying levels of confidence, without relying on additional annotated data for the evaluator itself.

4.2. EE Model

NER We adopt the XLM-RoBERTa[2] model (XLM-R).This model is fine-tuned separately on the training split of each dataset before being used as the external model for obtaining NER results, as it achieves the best results for the NER task across each dataset. To simulate handling new languages or examples not present in the training data, we use BERT-base-cased[3] without any task-specific fine-tuning as the second set input as mentioned in 4.1. This setup illustrates how the system can provide predictions for cases outside the scope of the original training datasets.

Split	fr	de	en
Original Train Set	158	103	N/A
Synthetic Train Set	2,212	1,442	N/A
Validation Set	43	33	N/A
Test Set	43	49	46

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

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)[4] as external model for obtaining NEL results. For second model, we use the same model architecture without pre-trained weights, meaning the model starts from scratch with no prior knowledge of the data.

4.3. Regression Model

For feature encoding, we investigate two modeling approaches: (i) LLM-based models and (ii) encoder transformer-based models.

For encoder-based models such as BERT and RoBERTa, we use the representation of the [CLS] token as the sentence-level feature. These models are fully fine-tuned with a learning rate of 1×10^{-5}, a batch size of 16, and trained for up to 5 epochs.

For LLM-based models, we apply parameter-efficient fine-tuning using LoRA (Hu et al., 2022) with rank $r = 16$, scaling factor $\alpha = 8$, and dropout rate 0.1. Due to computational constraints, LLM-based models are trained with a batch size of 4 for a maximum of 5 epochs.

The confidence score used in our analysis is obtained from the final softmax layer of the external NER and NEL models.

For the `Join` function in Section 3, we use the following format:
"OCR: ..." +"| Task results:..." +"| Confidence: ...".
Since the effective input length varies by model and tokenizer, we test three maximum-token thresholds (128, 256, 512) to examine the impact of chunk size

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

[3] google-bert/bert-base-cased

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

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, 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: Envoy of the United States America demands clarification about the shameful expulsion of two preachers from the Horgen congregation. | **NEL prediction:** _, _, 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 (translated into English for clarity).

on performance. The training samples for NER and NEL tasks can be found in Figure 2.

5. Results

5.1. Comparative Analysis of Models

In this section, we evaluate the performance of different model types, including BERT-based models (BERT, XLM-R) and LLMs (Llama 1B, Llama 3B, Gemma 7B). For the HIPE2020-fr dataset, we use CamemBERT (Martin et al., 2019), a BERT variant pretrained for French. For HIPE2020-de, we adopt GBERT (Chan et al., 2020), a BERT-based model tailored for German.

Following our ablation study, all models are fine-tuned on the synthetic versions of each dataset, using EE labels with associated probabilities, and optimized with mean squared error (MSE) loss.

To assess alternative approaches, we explore two strategies: a heuristic-based method and direct LLM-based prediction. In the heuristic-based approach, we consider a model's prediction trustworthy if its probability exceeds a given threshold; otherwise, we assign a new label. The predictions are then compared to the gold standard to compute evaluation scores. We can consider the heuristic-based method as baseline as it used as the direct information from prediction of NER/NEL model. In the LLM-based, we use in-context learning approach, we prompt the LLM with several labeled examples in the same format as the regression model training data to obtain predicted scores. In our experiments, we use Gemma 27B[5]. The results are summarized in Table 2 and Table 3.

[5] google/gemma-3-27b-it

Model	HIPE2020-de			HIPE2020-fr		
	512	256	128	512	256	128
LM/LLMs regression based						
BERT	3.83	4.65	5.39	7.22	**3.06**	3.47
XLM-R	**2.81**	3.34	**3.43**	**3.45**	3.17	**3.15**
GBERT[a]	3.25	**3.31**	3.57	7.01	4.12	3.47
LLaMA3.2 1B	4.67	6.31	4.08	14.20	12.11	4.71
LLaMA3.2 3B	4.19	5.34	4.94	9.09	11.51	7.95
Gemma 7B	7.09	7.40	6.50	7.79	4.87	6.41
Heuristics						
Probability-based	28.71	28.85	29.17	39.34	39.47	39.77
In-Context Learning						
Gemma 27B	48.37	52.22	57.37	37.77	42.52	47.08

[a] CamemBERT in case of HIPE2020-fr

Table 2: MAE (%) performance on NER tasks for different max token lengths.

Model	HIPE2020-de			HIPE2020-fr		
	512	256	128	512	256	128
LM/LLMs regression based						
BERT	1.87	1.96	2.18	2.42	3.31	**2.77**
XLM-R	**1.68**	**1.89**	**1.99**	**2.17**	**2.37**	2.98
GBERT[a]	2.24	2.25	2.54	2.48	2.56	2.95
LLaMA3.2 1B	1.92	2.15	2.37	2.42	2.96	2.87
LLaMA3.2 3B	1.91	2.09	2.38	2.64	2.84	2.87
Gemma 7B	1.97	3.33	1.90	2.20	2.24	2.82
Heuristics						
Probability-based	2.09	2.10	2.10	2.93	2.95	3.01
In-Context Learning						
Gemma 27B	80.64	83.45	88.69	24.19	14.47	10.55

[a] CamemBERT in case of HIPE2020-fr

Table 3: MAE (%) performance on NEL tasks for different max token lengths.

NER Tasks Table 2 reports MAE performance across the HIPE2020-fr and HIPE2020-de datasets for varying sequence lengths. Overall, regression-based models outperform heuristic and in-context learning baselines by a large margin, as the probability-based and Gemma 27B approaches yield substantially higher errors (28–57%). Among encoder-based models, BERT and CamemBERT show competitive in-language results, while XLM-R delivers the most consistent performance, maintaining MAE below 3.5% in all settings. In contrast, LLM-based models such as LLaMA 1B, 3B, Gemma 7B exhibit less stable results, particularly at longer input lengths, suggesting a tendency to overfit when training data is limited. These findings indicate that LLMs require substantially more supervised data or stronger regularization to adapt effectively to fine-grained regression tasks, whereas smaller encoder-based models generalize more robustly under data-constrained conditions.

NEL Tasks As shown in Table 3, the MAE results for NEL tasks reveal consistent advantages for regression-based models over both heuristic and in-context learning approaches, with the lat-

ter (Gemma 27B) exhibiting particularly high error rates. Among the regressors, encoder-based models achieve the lowest MAE, with XLM-R demonstrating the most balanced cross-lingual performance. Notably, LLaMA models (1B and 3B) perform competitively with these encoders, maintaining MAE values within a narrow range (1.9–2.9%) across languages and input lengths. This indicates that, for structured tasks like NEL, small LLMs can match the performance of fine-tuned encoder models when trained appropriately. MAE generally decreases with longer sequences, confirming that additional context supports more accurate entity disambiguation.

Finally, comparing NER and NEL, several patterns emerge. Encoder-based models remain robust across tasks, benefiting from bidirectional attention and full-context representations. LLMs with their autoregressive architecture and potential overfitting on small same-language datasets limit same-language NER performance, particularly at longer sequences. Sequence length effects are task-dependent: for NER, longer sequences primarily aid cross-lingual evaluation, whereas for NEL, longer sequences improve both same-language and cross-lingual performance due to the need for semantic context in linking (the same suggestion from (Xin et al., 2024)).

Model	HIPE2020-de			HIPE2020-en		
	512	256	128	512	256	128
HIPE2020-fr as the training set						
BERT	3.83	4.65	5.39	5.06	3.84	4.56
XLM-R	2.81	**3.34**	**3.43**	**3.50**	3.56	**3.20**
CamemBERT	**2.61**	5.78	7.58	4.71	**3.49**	4.23
LLaMA3.2 1B	4.67	6.31	4.08	9.28	9.88	5.29
LLaMA3.2 3B	4.19	5.34	4.94	7.25	8.80	7.43
Gemma 7B	3.36	9.94	10.06	6.21	9.13	8.51
Model	**HIPE2020-fr**			**HIPE2020-en**		
	512	256	128	512	256	128
HIPE2020-de as the training set						
BERT	**3.78**	**3.01**	**3.49**	7.01	7.39	5.58
XLM-R	4.04	3.92	3.55	5.81	**4.38**	**4.38**
GBERT	4.06	3.60	4.95	**5.10**	4.61	5.94
LLaMA3.2 1B	9.21	9.84	5.48	8.12	6.76	6.46
LLaMA3.2 3B	9.49	6.52	5.52	8.35	7.74	7.01
Gemma 7B	10.41	6.38	5.45	9.61	6.97	6.84

Table 4: MAE (%) performance on NER tasks for different in the cross-lingual settings

5.2. Cross-lingual analysis

Table 4 and 5 report MAE performance across the HIPE2020-fr, HIPE2020-de, and HIPE2020-en datasets for varying sequence lengths. Regression-based encoder models consistently achieve the lowest MAE across both NER and NEL tasks in cross-lingual transfer, with XLM-R showing particularly balanced performance across target languages. LLMs models (1B and 3B) perform competitively,

Model	HIPE2020-de			HIPE2020-en		
	512	256	128	512	256	128
HIPE2020-fr as the training set						
BERT	2.27	2.62	2.44	2.90	2.76	3.27
XLM-R	**1.88**	2.05	2.39	2.21	2.70	2.76
CamemBERT	2.58	2.28	2.87	2.57	2.87	3.13
LLaMA3.2 1B	1.96	2.25	2.29	2.19	2.57	**2.50**
LLaMA3.2 3B	2.03	2.13	2.29	2.13	2.47	**2.50**
Gemma 7B	**1.88**	**1.82**	**2.04**	**2.02**	**2.33**	**2.50**
Model	**HIPE2020-fr**			**HIPE2020-en**		
	512	256	128	512	256	128
HIPE2020-de as the training set						
BERT	2.76	3.06	3.59	2.39	3.20	3.18
XLM-R	2.54	**2.71**	3.08	2.37	**2.41**	2.89
GBERT	3.02	3.19	3.68	2.95	3.53	3.24
LLaMA3.2 1B	2.65	3.26	3.37	2.51	3.45	2.92
LLaMA3.2 3B	2.53	3.00	3.37	2.39	2.59	2.92
Gemma 7B	**2.52**	4.64	**2.63**	**2.34**	2.94	**2.20**

Table 5: MAE (%) performance on NEL tasks in cross-lingual settings

often approaching encoder-level accuracy, demonstrating that relatively small LLMs can generalize well when fine-tuned. NEL errors are generally lower than NER, likely due to fewer entities and a more constrained output space. Sequence length has a minor effect: longer contexts slightly reduce MAE for both tasks, suggesting additional input information supports more accurate for NER and NEL. Overall, these results confirm that regression-based fine-tuning enables robust cross-lingual quality estimation for both NER and NEL.

5.3. Ablation study

In this part, we conduct experiments with several setup components using probability from the EE task, MSE loss functions, and synthetic data. In this setup, we use the same base model, Bert-base-cased (Devlin et al., 2019)[6], utilizing synthetic data, the MSE objective function for a fair comparison.

Task	Setting	HIPE2020-de			HIPE2020-fr		
		512	256	128	512	256	128
NER	[1]	35.06	35.58	35.85	48.37	48.6	49.02
	[2]	33.24	33.63	34.07	44.92	45.2	45.55
	[3]	4.42	2.99	7.93	15.65	10.54	4.25
	[4]	**3.51**	**3.19**	**3.84**	**7.22**	**3.06**	**3.47**
NEL	[1]	3.01	3.27	3.17	3.69	4.11	4.49
	[2]	3.25	3.06	3.08	4.07	3.80	4.24
	[3]	2.00	2.52	2.39	3.05	**2.67**	3.62
	[4]	**1.71**	**1.78**	**2.18**	**2.42**	3.31	**2.77**

Table 6: Ablation study for NER and NEL tasks [1] : *OCRed + EE results*; [2] *OCRed + EE (results + prob)*; [3] *OCRed + EE (results) + synthetic*; [4] *OCRed + EE (results + prob) + synthetic*.

Table 6 illustrates the impact of progressively adding probability information and synthetic data on NER and NEL across sequence lengths. For

[6]Bert-base-cased

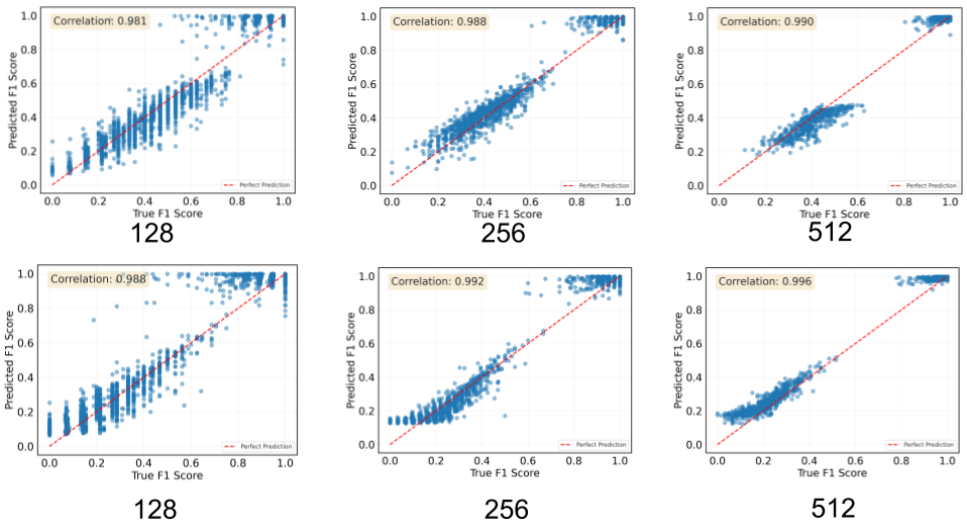

Figure 3: Correlation Between Predicted and True F1 Scores Across Token Lengths for HIPE2020-de (top) and HIPE2020-fr (bottom) for the NER task.

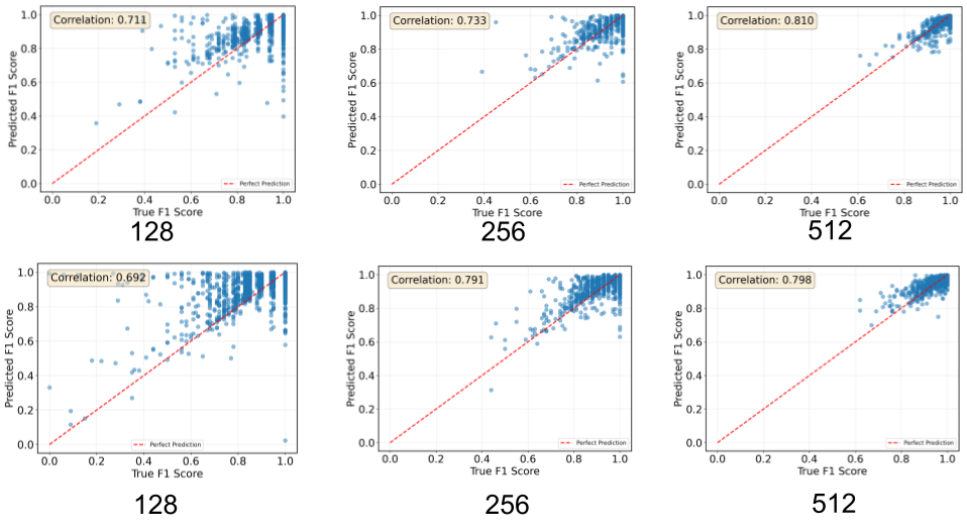

Figure 4: Correlation Between Predicted and True F1 Scores Across Token Lengths for HIPE2020-de (top) and HIPE2020-fr (bottom) for the NEL task.

NER, combining both enhancements ([4]) produces the largest MAE reductions, reaching 3.06–7.22% on French and 3.19–3.84% on German at 512 tokens, outperforming intermediate settings ([2] and [3]) and the baseline ([1]). Probability cues refine entity predictions, while synthetic data improves robustness, and once both are applied, sequence length has minimal additional effect.

For NEL, a similar trend is observed, with [4] achieving the lowest MAE across languages (e.g., 2.42 - 3.31% on French, 1.87–2.18% on German). Compared to NER, the MAE values for NEL are smaller overall, reflecting task simplicity: NEL focuses on a smaller set of entities and primarily resolves semantic linking, whereas NER requires precise boundary detection and label assignment, making it more challenging. Sequence length effects remain limited once enriched context is provided, emphasizing that local context combined with probabilistic and synthetic signals is sufficient for accurate entity linking.

5.4. Prediction Analysis

Figure 3 illustrates a strong linear relationship between predicted and true F1 scores for regression models on the HIPE2020-de and HIPE2020-fr datasets. Across different input lengths, correlations remain consistently high, indicating that NER performance estimation is stable and robust even with shorter contexts. This consistency un-

derscores the reliability of the regression approach for NER quality prediction.

In contrast to the near-perfect predictive fidelity observed for NER, the NEL task reveals a more pronounced sensitivity to input scale, with R^2 scores starting modestly at 0.711 and 0.692 for 128 tokens, respectively, before climbing to 0.810 and 0.798 at 512 tokens. This evident progression underscores the task's greater complexity, as NEL demands disambiguating entities across sparse, context-dependent signals that benefit substantially from expanded token windows, yet the persistently lower R^2 values relative to NER highlight inherent challenges like referential ambiguity and cross-document knowledge gaps, rendering accurate performance estimation more elusive despite the clear gains from richer token representations (See figure 4).

6. Conclusion

In this work, we investigate using LLMs as quality estimators for EE tasks, specifically NER and NEL, on historical OCRed texts. Rather than treating LLMs solely as task solvers, we reframe evaluation as a regression problem, enabling reference-free assessment of EE outputs without relying on ground-truth annotations. This approach positions LLMs as practical meta-models for estimating output quality in noisy, low-resource, and annotation-scarce environments. Such a role supports uncertainty modeling, performance monitoring, and cross-lingual generalization, highlighting the potential of LLM-based evaluators to complement traditional EE systems in challenging real-world.

Our findings show that LLM-based estimation holds significant promise for assessing downstream EE tasks. Results suggest these assessments are relatively language independent, though high error rates on noisy OCRed text highlight the persistent challenge of robust extraction in historical, multilingual settings.

Beyond these results, our study contributes three broader implications: it offers a scalable alternative to annotation-heavy evaluation, enables in domains with scarce gold labels such as historical archives, and highlights LLMs' potential as general-purpose evaluators for entity-centric NLP. Future work could explore agentic LLMs for self-assessment, extend evaluation to additional EE tasks, and test across multiple extraction systems to improve robustness and 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.

Although our evaluation focuses on 19^{th}–20^{th} newspapers, models are trained on large and diverse corpora that include various forms of noisy OCR text. This may facilitate a degree of transferability to related languages, scripts, or historical genres. However, adaptation to typologically distant languages or low-resource settings may require additional validation, as differences in linguistic structure, orthography, or data availability could affect performance.

8. Acknowledgments

This work has been co-funded by the European Union HORIZON-WIDERA-2023-TALENTS-01-01 grant 101186647 — AI4DH. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

This work has also been supported by the TERMITRAD (2020-2019-8510010) project funded by the Nouvelle-Aquitaine Region (France), and it has benefited from the computing resources of the L3i laboratory, operated and hosted by the University of La Rochelle, and funded by the French government and the Nouvelle-Aquitaine Region.

9. Bibliographical References

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

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.

Branden Chan, Stefan Schweter, and Timo Möller. 2020. German's next language model. In *Proceedings of the 28th international conference on computational linguistics*, pages 6788–6796.

Guillaume Chiron, Antoine Doucet, Mickaël Coustaty, Muriel Visani, and Jean-Philippe Moreux. 2017. Impact of ocr errors on the use of digital libraries: towards a better access to information. In *2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pages 1–4. IEEE.

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.

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

Maud Ehrmann, Ahmed Hamdi, Elvys Linhares Pontes, Matteo Romanello, and Antoine Doucet. 2023. Named entity recognition and classification in historical documents: A survey. *ACM Computing Surveys*, 56(2):1–47.

Maud Ehrmann, Matteo Romanello, Simon Clematide, and Antoine Doucet. 2022a. Introducing the HIPE 2022 Shared Task:Named Entity Recognition and Linking in Multilingual Historical Documents. In *Proceedings of the 44^d European Conference on IR Research (ECIR 2022)*, Stavanger, Norway. Lecture Notes in Computer Science, Springer.

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

Alexandre Gefen. 2014. Les enjeux épistémologiques des humanités numériques. *Socio-La nouvelle revue des sciences sociales*, pages 61–74.

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, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Jiafeng Guo, Gu Xu, Xueqi Cheng, and Hang Li. 2009. Named entity recognition in query. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, page 267–274, New York, NY, USA. Association for Computing Machinery.

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, et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Alexandre Jaud, Ahmed Hamdi, Antoine Doucet, Adam Jatowt, and Mickaël Coustaty. 2025. Beyond cer and wer: How does ocr really impact information retrieval? In *2025 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pages 30–39.

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*.

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.

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, 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.

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*.

Amy Xin, Yunjia Qi, Zijun Yao, Fangwei Zhu, Kaisheng Zeng, Xu Bin, Lei Hou, and Juanzi Li. 2024. Llmael: Large language models are good context augmenters for entity linking. *arXiv preprint arXiv:2407.04020*.

Ikuya Yamada, Koki Washio, Hiroyuki Shindo, and Yuji Matsumoto. 2022. Global entity disambigua-

tion 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.

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.

10. Appendices

10.1. Synthetic Noise Parameters

We adapt the analytical noise model proposed by Jaud et al. (2025) and integrate it into our synthetic noise generation framework. The corruption process is controlled by two parameters: the *threshold level* and the *noise level*.

The **threshold level** determines the proportion of tokens to which noise is applied. It ranges from 0 to 1 with a step size of 0.5.

The **noise level** specifies the number of characters replaced within each selected token, ranging from 1 to 4 characters.

The parameter combinations results in 20 initial noise configurations. For clarity and comparability, we retain only four representative noise levels per perturbation setting (entity tokens, surrounding context, and all tokens; see Section 4.1). This selection is applied independently to each HIPE2020 language subset. Details are reported in Table 7.

Language	Noise	Entity		Context		All
		WER	CER	WER	CER	WER/CER
FR	L1	0.014	0.009	0.064	0.035	0.15 / 0.10
	L2	0.024	0.014	0.099	0.064	0.25 / 0.17
	L3	0.034	0.024	0.129	0.100	0.35 / 0.29
	L4	0.044	0.038	0.161	0.125	0.45 / 0.37
DE	L1	0.013	0.011	0.055	0.027	0.15 / 0.11
	L2	0.022	0.020	0.089	0.042	0.25 / 0.15
	L3	0.034	0.003	0.113	0.078	0.35 / 0.21
	L4	0.043	0.044	0.147	0.120	0.45 / 0.27

Table 7: WER and CER across noise levels for FR and DE.

The corrupted files are used consistently for both NER and NEL to ensure fair and comparable evaluation across settings. Predictions on the corrupted data are obtained using fine-tuned XLM-RoBERTa models for NER and NEL (Section 4.1). The final joint predictions are publicly available[7].

10.2. Prompting

In this section, we present the prompts used for ICL with Gemma 27B, corresponding to the results reported in Section 5.

[7] gitlab/reference-free

NER

```
You are an expert Named Entity
↪  Recognition evaluation assistant.

Instructions:
1. You are given a predicted entity (or
↪  entities) for some OCR text.
2. Evaluate the predictions against the
↪  correct entity categories (B-pers,
↪  I-pers, B-loc, I-loc, B-org, I-org,
↪  B-org, I-org, B-prod, I-prod, 0) per
↪  token.
3. Compute the overall F1 score for the
↪  prediction.
Return ONLY valid JSON in the format:
{{ "f1": float between 0 and 1 }}

Examples (for illustration):
**Input**: "Sentence": "le département
↪  des finances est autorisé :"
"Predicted_entities (per token, in
↪  order)": ["O", "O", "O", "O", "O",
↪  "O", "O"]
**Output**: {{"f1": 0.0}}

Now process the following:
"Sentence": {sentence}
"Predicted_entities (per token, in
↪  order)": {pred_entities}
```

NEL

```
You are an expert Named Entity Linking
↪  evaluation assistant.

Instructions:
1. You are given a predicted entity (or
↪  entities) for some OCR text.
2. Compare the predicted Wikidata ID(s)
↪  with the correct Wikidata ID(s)
↪  based on the entity name and its
↪  context.
3. Compute the overall F1 score for the
↪  prediction.
Return ONLY valid JSON in the format:
{{ "f1": float between 0 and 1 }}

Examples (for illustration):

**Input**: "Sentence": "NOUVELLES
↪  SUISSES - En 1887 , la Société
↪  suisse du Grutli s ' est accrue"
"Predicted_entities (per token, in
↪  order)": ["_", "_", "_", "Q7826",
↪  "Q7826", "_", "_", "Q683672",
↪  "Q683672", "Q683672", "Q683672", "_",
↪  "_", "_", "_"]
**Output**: {{"f1": 0.5}}

Now process the following:
"Sentence": {sentence}
"Predicted_entities (per token, in
↪  order)": {pred_entities}
```