INSA Lyon at HIPE 2026: A Modular Ensemble Pipeline for Person–Place Relation Classification in Historical Newspapers

Sang Nguyen[1,*], Ruixing Zheng[2], Diana Nurbakova[3,*] and Killian Barrere[3]

[1]INSA Lyon, 69100 Villeurbanne, France
[2]Sorbonne Université, 75006 Paris, France
[3]INSA Lyon, CNRS, Universite Claude Bernard Lyon 1, LIRIS, UMR5205, 69100 Villeurbanne, France

Abstract

This paper presents our participation in the HIPE-2026 shared task at CLEF 2026, which addresses person–place relation extraction in multilingual historical newspapers. The task requires predicting, for each person–location pair, whether the person was ever present at the given location before the document's publication, and whether they were present there around the time of publication. We propose a modular heterogeneous ensemble pipeline combining multilingual Transformer encoders (mDeBERTa-v3 and XLM-RoBERTa-large), zero-shot large language models (Gemma 4 31B and Llama 3.3 70B), and handcrafted features. Predictions are combined through ensemble strategies including plurality voting and lookup-table stacking. We submitted three runs representing different performance–efficiency trade-offs: a full heterogeneous ensemble (Run 1), a compact single-encoder baseline (Run 2), and a frugal local ensemble without Large Language Models inference (Run 3). On the official impresso test set, Run 1 achieved a macro-recall of 0.639, ranking 11th out of 45 submissions. Post-hoc analysis against the released gold labels reveals zero recall on the minority PROBABLE class across all runs, with errors concentrating in partially linked pairs and under cross-domain transfer to unseen French literary text. The code and resources associated with this work are available at github.com/sangnguyen2803/insalyon-at-hipe2026.

Keywords

Relation Extraction, Person-place Relations, Large Language Models, Transformers, Historical Newspapers

1. Introduction

Semantic relations between persons and places in historical documents are a fundamental building block for historical knowledge graphs, cultural heritage preservation, and large-scale archive retrieval. Historical newspapers, in particular, capture rich traces of human activity across time and space, making person–place relation extraction a valuable yet underexplored task in the digital humanities.

However, applying relation extraction systems to historical text raises significant challenges. Optical Character Recognition (OCR) noise, spelling variants, and archaic vocabulary substantially degrade the performance of Natural Language Processing (NLP) systems trained on modern corpora. Beyond surface-level noise, person–place relations in historical archives carry a strong temporal dimension: whether a person *was ever* at a location is a fundamentally different assertion from whether they *were present near the time of publication*, requiring systems to perform both cross-sentence reasoning and temporal anchoring. To advance on this challenge, HIPE-2026 proposes multilingual historical newspaper data (German, English, and French) [1, 2] and defines two person–place relation labels targeting different temporal scopes. Here, the classification goes beyond a simple true/false distinction.

Prior work in historical document processing has established strong baselines for named entity recognition and linking [3]. Systems participating in HIPE-2020 [4] and HIPE-2022 [5] demonstrated that Transformer architectures [6, 7] combined with historically pre-trained language models such as hmBERT [8] and CRF layers consistently outperform symbolic and BiLSTM-based approaches on

CLEF 2026, September 21-24, 2026, Jena, Germany
*Corresponding author.
sang.nguyen@insa-lyon.fr (S. Nguyen); zhengruixingfr@outlook.com (R. Zheng); diana.nurbakova@insa-lyon.fr (D. Nurbakova); killian.barrere@insa-lyon.fr (K. Barrere)
0009-0006-6323-1342 (S. Nguyen); 0000-0002-6620-7771 (D. Nurbakova); 0009-0004-6955-5191 (K. Barrere)


NERC tasks [9, 4, 5]. More recently, Large Language Models (LLMs) have shown strong generalization in few-shot and low-resource settings [10, 11]. However, extending these methods to person–place relation classification, particularly under the temporal constraints introduced by HIPE-2026, remains an open problem.

In this paper, we present our system for the HIPE-2026 shared task and report results across three submitted runs corresponding to different performance–efficiency operating points. The main contributions of this paper are as follows:

- We propose a modular heterogeneous ensemble system for person–place relation extraction in historical newspapers, effectively combining handcrafted features, fine-tuned Transformer encoders, and zero-shot LLM predictions;
- We systematically compare three configurations of varying scale and complexity, quantifying the contribution of LLM components and ensemble strategies to overall performance;
- Through post-hoc error analysis on the released gold labels, we identify zero recall on the minority PROBABLE class, the impact of entity linking coverage on prediction quality, and the main failure modes under cross-domain generalization as the primary residual challenges.

The remainder of this paper is organized as follows: Section 2 reviews related work, while Section 3 describes the task and data. In Section 4, we present our system architecture. Section 5 then details the experimental setup, before reporting obtained results and providing analyses in Section 6. Finally, limitations and future works are discussed in Section 7.

2. Related Work

Relation Extraction (RE) is a core task in NLP and Information Extraction (IE), aiming to automatically identify semantic relationships between entities in unstructured text, typically represented as triplets of the form (*entity$_1$, relation, entity$_2$*) [12, 13].

Early RE approaches relied primarily on rule-based heuristic pattern matching and traditional machine learning algorithms such as Support Vector Machines, Conditional Random Fields, and Hidden Markov Models. This gradually gave rise to the Open Information Extraction (Open IE) paradigm that requires no predefined relation types [12, 13]. With the widespread adoption of Deep Learning, neural architectures such as Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) were extensively applied to RE tasks, enabling automatic learning of long-range dependencies and deep sentence-level features while substantially reducing reliance on manual feature engineering [13]. The emergence of pre-trained Language Models further advanced the field: through a systematic analysis of 137 papers published at ACL conferences from 2020 to 2023, Diaz-Garcia and Lopez [10] found that BERT-based Transformer methods [7] consistently dominate RE benchmarks and achieve state-of-the-art performance, while large language models (LLMs) such as T5 [14] and GPT [15] demonstrate strong generalization in few-shot and low-resource settings, particularly in identifying previously unseen relation types.

Traditional RE has focused primarily on sentence-level extraction, identifying relations between entity pairs appearing within a single sentence. However, in real-world scenarios, relational information is often distributed across different parts of a document, motivating a shift toward Document-level Relation Extraction (DocRE). Compared to sentence-level RE, DocRE offers a broader analytical context but also introduces greater challenges, requiring models to handle entity coreference, extract transient relations, and perform logical reasoning across multiple sentences [16]. To systematically assess this challenge, Yao et al. [17] constructed DocRED, a large-scale manually annotated dataset built from Wikipedia and Wikidata. Their analysis shows that at least 40.7% of relational facts can only be extracted through cross-sentence reasoning, including logical inference, coreference resolution, and common-sense reasoning, posing significant challenges to existing methods. Exemplified by ATLOP and DREEAM, Delaunay et al. [16] found that Transformer-based approaches achieve approximately 65% F1 on DocRED, confirming that document-level RE remains an open problem.

Despite notable progress on modern standard text, applying RE to historical documents and cultural heritage materials introduces fundamentally different challenges. Historical texts suffer from OCR noise, spelling variants, archaic vocabulary, and multilingual content, all of which substantially degrade the performance of NLP systems trained on modern corpora [4]. To address named entity processing in historical documents, the HIPE shared task series (HIPE-2020, HIPE-2022) conducted NER and entity linking evaluations on multilingual historical newspapers, providing important benchmarks for the field [4, 5]. Participating systems typically combined BERT and CamemBERT encoders with CRF layers and employed character-level feature extraction, multi-model ensembling, and historically pre-trained language models such as hmBERT to handle OCR noise and spelling variation.

It is also worth noting that entity relations in historical documents carry a strong temporal dimension. While conventional RE typically treats relations as static facts, in historical archives the interactions between persons and places (e.g. *born in*, *resided in*, or *worked at*) are highly dependent on specific time points or periods. Time-aware RE thus becomes a core requirement for processing historical corpora [18].

HIPE-2026 addresses this challenge [1, 2] by defining two person–place relation types with different temporal scopes: `at` (whether a person was ever at a given place) and `isAt` (whether a person was present at a place around the time of publication). These systems are required to simultaneously handle noise tolerance, cross-sentence reasoning, and temporal anchoring [1, 2]. This dual emphasis on temporal awareness and noise robustness represents the current frontier of relation extraction research in the digital humanities.

3. Data and Task Description

3.1. Task Description

HIPE-2026 addresses person–place relation extraction in multilingual historical documents. For each sampled person–location pair in a document [1, 2], systems are required to predict two related labels, `at` and `isAt`, based on the evidence provided by the text. Unlike approaches based solely on entity co-occurrence, the task requires contextual understanding and temporal reasoning to infer the underlying relations.

The `at` label captures whether the document supports the inference that the person was at the specified location at any time prior to publication. This is formulated as a three-way classification task: TRUE when the text clearly supports the relation, PROBABLE when the evidence is suggestive but not conclusive, and FALSE when the relation is unsupported or contradicted by the context.

The `isAt` label is a stricter temporal variant of `at`. It asks whether the text indicates that the person was at the location within approximately one month before the publication date, and is therefore treated as a binary classification task (TRUE/FALSE). By definition, `isAt` presupposes `at`: a pair cannot receive `isAt=TRUE` if `at=FALSE`.

The dataset used in this task consists of a training set (Train), a primary test set (Test A), and a generalization test set (Test B). The data are sourced from historical newspaper archives and literary texts, and are designed to evaluate model robustness across different languages, historical periods, and text genres.

3.2. Train Data

The training set contains 104 documents and 1,251 annotated person–location pairs (see [1, 2] for full statistics). Two aspects of this data are especially important for system development. First, the label distribution is strongly imbalanced: FALSE accounts for 55.2% of instances, whereas PROBABLE accounts for only 9.6%, making the minority middle class difficult to learn. Second, Wikidata coverage is uneven across entity types. Person entities are linked in only 43.1% of cases, compared with 75.9% for locations, which limits the extent to which the system can rely on external entity matches during training.

3.3. Test Data

The test set consists of two parts: Test A (`impresso`), the primary in-domain evaluation set, and Test B (`surprise-test-fr`), an out-of-domain generalization test. Together they contain 87 documents and 1,118 annotated person–location pairs (see [1, 2] for full statistics). Relative to the training data, two changes are especially important for evaluation. First, French documents are more strongly represented in the test set (49 out of 87), introducing a moderate shift in language distribution. Second, Wikidata coverage for person entities drops further to 31.7%, which reduces the availability of external entity information at test time. Together, these differences make generalization more difficult.

4. System Description

4.1. System Architecture Overview

We present our system for HIPE 2026 as a modular pipeline for multilingual person–place relation classification in historical newspapers [1, 2], instantiated in three submitted runs. Given a document and a pre-identified person–location pair, the system predicts two relation labels: `at`, a three-class label (TRUE, PROBABLE, FALSE) representing a general association between the person and the location, and `isAt`, a binary label (TRUE, FALSE) indicating whether the person is physically present at the location within the temporal context of the document.

Figure 1 presents an overview of the proposed architecture.

At a high level, the system consists of five blocks: preprocessing, representations, models, a decision layer, and evaluation. Preprocessing builds pair-level instances and extracts local and temporal context (Section 4.2). Representations prepare the inputs used by the downstream predictors. The model block comprises handcrafted models (Section 4.3), multilingual encoders (Section 4.4), and prompted LLMs (Section 4.5). Their outputs are combined in the decision layer (Section 4.6) to produce the final `at` and `isAt` labels used in the three submission runs (Section 4.7).

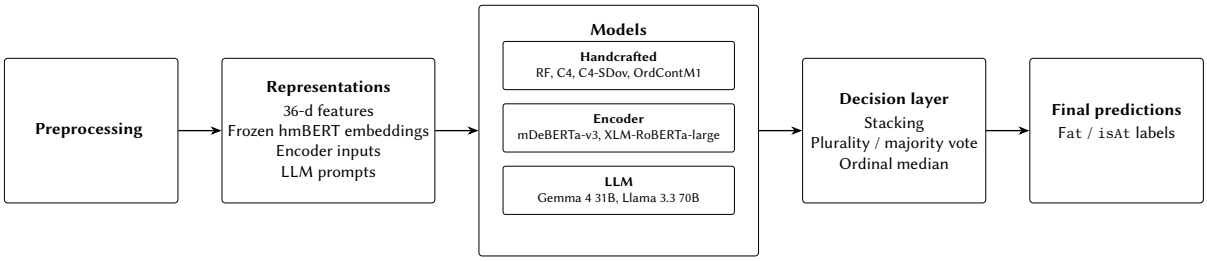

Figure 1: Overview of our relation extraction pipeline.

4.2. Data Preprocessing

We first transform the original document-level data into pair-level classification instances. Each article may contain multiple annotated person-location pairs, and each pair is treated as a separate example while preserving the document identifier, language, and publication date.

For each pair, we extract a local textual context around the target mentions using a sentence-based heuristic. If the person and location mentions occur close to each other, we keep the sentence span covering both mentions; otherwise, we retain a reduced context around the two mention regions. We also apply light text normalization by removing line breaks and collapsing repeated whitespace.

We then attach the temporal information used by the downstream models. This includes automatically detected temporal expressions, lexical temporal signals, coarse tense/aspect and negation cues from the main predicate. When temporal expressions can be aligned with document time, we compute coarse distance features relative to the publication date. These outputs are reused across the handcrafted, encoder-based, and LLM-based model families.

4.3. Handcrafted Models

The handcrafted component of the pipeline combines explicit feature engineering with representations derived from the frozen hmBERT [8] model. One component in this family is a Random Forest (RF)[19] classifier, which operates only on a 36-dimensional handcrafted feature vector. These features encode temporal cues, sentence position, document metadata, language indicators, entity-linking information, person-status categories, mention counts, and coarse context statistics. In parallel, the other models in this branch (C4, C4-SDov and OrdContM1) use frozen relation representations extracted by applying hmBERT to the marked local context through a MASK template and taking the resulting MASK and entity-span hidden states as dense features. We use hmBERT because it is pretrained on multilingual historical newspaper text and is therefore better suited than a generic modern-language BERT to OCR-noisy historical documents and small-data transfer. The full feature list is reported in Appendix A.

4.4. Transformer-based Models (mDeBERTa and XLM-RoBERTa)

The encoder-based part of the pipeline uses multilingual transformers as supervised relation classifiers. We work with two encoders, mDeBERTa-v3-base [20] and XLM-RoBERTa-large [21], to cover a compact option and a larger option within the same model family. The system does not assume that both are used at the same time. Instead, the encoder block can use either model, depending on which configuration is submitted

Both encoders take the same pair-level input built from the target entities and their local context. In each case, the model is trained to predict the `at` and `isAt` labels. Training details, including the loss function and regularization choices, are reported in Section 5.

4.5. LLM-based Models

The LLM-based part of the pipeline relies on prompted zero-shot inference rather than task-specific fine-tuning. We use Gemma 4 31B[22] as the main LLM predictor because it is the only LLM applied to both `at` and `isAt`. Among the LLMs considered in this work, it provides the most suitable balance between reasoning performance and deployment cost. Llama 3.3 70B[23] is used only for `isAt`, where it contributes an additional voting signal alongside the other predictions.

The prompt is constructed from a pair-level representation, consisting of the local article context and an ordered target entity pair. In the official LLM runs, we use a zero-shot pairwise prompt variant, denoted as `P-AB`, where `A` and `B` refer to the head and tail entities, respectively. No retrieved examples or auxiliary Wikidata or temporal evidence blocks are included. The full prompt template is provided in Appendix B.

In the proposed architecture, the LLM-based predictors are not treated as an isolated standalone system. Instead, they provide an additional reasoning signal that is combined with encoder-based and handcrafted predictors in the final decision layer.

4.6. Ensemble and Decision Strategy

The decision layer takes as input one label prediction per model for a given person–location pair and returns the final `at` and `isAt` labels. Depending on the run, these inputs come from feature-based models, encoder-based models, and LLMs, and they are combined either by direct voting or by a two-stage stacking procedure.

For `at`, the main stacked component combines four complementary predictors: a Random Forest over handcrafted features, a logistic-regression[24] model over frozen hmBERT-based relation features with a post-hoc rule-based overlay, a compact contrastive MLP over the same frozen hmBERT features, and Gemma as the LLM predictor. For each training instance, the resulting four-model vote tuple is paired with the gold `at` label and we record the empirical frequency of each observed tuple.

At inference time, the observed vote tuple is looked up in this table and mapped to the corresponding output label. To handle sparse tuples (those observed only once or twice in the training pool) we use

a simple, conservative backoff policy: if the tuple frequency is >= 3 we return the empirical majority label for that tuple; if the tuple occurs exactly twice and both occurrences agree on the same label we accept that label; otherwise (frequency 1, or frequency 2 with disagreement) we back off to a robust ensemble fallback: plurality voting across the available model outputs (including encoder predictions when present), with ordinal-median tie-breaking on the ordered set FALSE < PROBABLE < TRUE. This policy avoids relying on noisy single observations while still exploiting tuple-specific signals when supported by enough training examples. The stacked prediction can be used on its own or combined with encoder outputs in a subsequent voting step.

For `isAt`, the final combinations are simpler: the system uses either a single predictor or a majority vote across several predictors. No ordinal tie-breaking is needed because `isAt` is a binary task. For `at`, direct voting is done by plurality over the available model outputs. When plurality voting produces a tie, we resolve it with the ordinal median over the ordered label set

$$\text{FALSE} < \text{PROBABLE} < \text{TRUE}.$$

4.7. Submission Runs

Submission to HIPE-2026 is limited to three related configurations derived from the same overall system. Run 1 is the reference configuration, while Run 2 and Run 3 are simpler variants obtained by removing specific parts of Run 1 to study their impact on performance and deployment cost. For reference, Table 1 recalls the component composition of each run.

Table 1
Component composition of the three submitted runs. Included components are marked with a check mark; absent components are marked with a dash.

	mDeBERTa	XLM-R	Stacker	Gemma	Llama	Parameters	Size
Run 1	✓	✓	✓	✓	✓	102B	195.7 GB
Run 2	✓	—	—	—	—	278M	1.1 GB
Run 3	✓	✓	—	—	—	839M	3.2 GB

Run 1. Run 1 is the full heterogeneous ensemble. For the `at` task, it combines XLM-RoBERTa-large, the 4-model lookup-table stacker, and Gemma 4 31B through plurality voting with ordinal-median tie-breaking. For the `isAt` task, it combines XLM-RoBERTa-large, Gemma 4 31B, and Llama 3.3 70B through majority voting. This is the most complete submission and uses the richest combination of handcrafted, encoder-based, and LLM-based models.

Run 2. Run 2 is the compact joint-encoder baseline. It uses only mDeBERTa-v3 base, with one shared encoder and two task-specific heads for `at` and `isAt`. Compared with Run 1, it removes LLM-based reasoning, disagreement stacking, and multi-model voting. This makes it the clearest single-model reference and the smallest submitted system in terms of model size and parameter count.

Run 3. Run 3 is a frugal local ensemble that removes hosted LLM (Gemma and Llama) inference at test time. For the `at` task, it combines XLM-RoBERTa-large, mDeBERTa-v3 base, and the Random Forest classifier through plurality voting with ordinal-median tie-breaking. For the `isAt` task, it uses XLM-RoBERTa-large alone. Relative to Run 1, it keeps predictor diversity but restricts the system to locally deployable models, making it a useful compromise between the full ensemble and the single-model baseline.

Taken together, the three runs correspond to three interpretable operating points of the same system: a full heterogeneous ensemble, a compact single-encoder baseline, and a frugal local ensemble without hosted LLM inference.

5. Experimental Setup

This section will detail the experimental configurations of our submitted systems, including the fine-tuning setup for the encoder-based models, the training objectives and regularization choices, implementation details and model sizes, and the evaluation metrics used in this shared task.

5.1. Hyperparameters and Fine-tuning

The two supervised encoder models in our system, mDeBERTa-v3 base and XLM-RoBERTa-large, use the same general fine-tuning setup so that performance differences can be interpreted mainly in terms of model capacity rather than unrelated optimization choices. Specifically, both mDeBERTa-v3-base and XLM-RoBERTa-large are fine-tuned for 40 epochs with a learning rate of 2e-5, an effective batch size of 16, label smoothing of 0.05, and inverse squared root class weighting. The train/validation split is 1063/188.

For Run 2, mDeBERTa-v3 base is fine-tuned as a joint model with two prediction heads, one per task. XLM-RoBERTa-large is fine-tuned as the stronger encoder expert used in the ensemble settings. In contrast, the LLM components are not fine-tuned: Gemma 4 31B and Llama 3.3 70B are used in zero-shot mode with the prompt. On the feature-based side, the Random Forest classifier is trained on the 36-dimensional hand-crafted vector, while the C4 model uses a `StandardScaler` followed by logistic regression, and OrdContM1 is implemented as a compact contrastive MLP over frozen hmBERT-derived representations. Across all supervised components, the train/validation split is 1063/188, reserving $\approx 15\%$ of the labeled pool for evaluation using a stratified split per task.

5.2. Loss Function

For the encoder-based models, we use weighted cross-entropy as the main training objective, motivated by the label imbalance of both tasks: the low frequency of the PROBABLE class in `at` and the skew toward FALSE in `isAt`. The weighting scheme increases the contribution of minority labels during optimization, reducing the tendency of the model to over-predict the dominant classes.

We combine this objective with label smoothing. The aim is to avoid overly sharp posterior distributions on ambiguous historical cases, where the distinction between explicit evidence, weak evidence, and absence of evidence is often subtle. In practice, weighted cross-entropy and label smoothing work together as a simple way to improve robustness without altering the underlying encoder architecture.

5.3. Implementation Details

The three submitted systems combine components with different deployment conventions, so we report both parameter counts and on-disk model sizes using the HIPE-2026 evaluation guideline. Fine-tuned encoder checkpoints and the frozen hmBERT encoder are counted at FP32, whereas hosted LLMs are counted at BF16. For sklearn and compact neural artifacts, we report the deployed on-disk size of the corresponding `joblib` or `.pt` file. Accordingly, Table 2 summarizes the parameter counts and on-disk sizes of all reusable components, providing the basis for the efficiency comparisons reported in Section 6.

The official submission files are written by merging pair-level predictions back into the original nested article-level JSONL structure, producing one file per language and test collection. All three official runs follow the same output format.

5.4. Metrics

The official evaluation follows the HIPE-2026 metric definition[1, 2]. For each task, performance is measured using macro-recall (MR), which gives equal importance to each class regardless of class frequency. We report MR(`at`) for the three-way association task and MR(`isAt`) for the binary document-time presence task. The official headline score is the average of these two values.

Table 2
Parameter counts and deployed model sizes.

Component	Parameters	Size (MB)	Precision / format
Random Forest (handcrafted)	0.84M	20	`joblib`
C4 (StandardScaler + LR)	11.6K	1	`joblib`
OrdContM1	2.16M	8	`.pt`
hmBERT encoder	110.6M	424	FP32
mDeBERTa-v3 base	278M	1,061	FP32
XLM-RoBERTa-large	560M	2,136	FP32
Gemma 4 31B	30.7B	58,556	BF16
Llama 3.3 70B Instruct	70.6B	134,571	BF16

For internal validation and model development, we also inspect validation macro-F1 reported by the encoder training runs, since it provides a complementary view of class-balanced classification quality during fine-tuning. When discussing official performance, however, we prioritize macro-recall because it is the shared-task target metric.

6. Results

6.1. Results before Submission

We validated all three runs on the reserved evaluation split described in Section 5 using macro-recall (MR). For `isAt`, feature-based and ensemble combinations were also evaluated on the official v1.0 `isAt-test` set. Table 3 summarizes the main pre-submission results.

Table 3
Pre-submission comparison of the three submitted runs. MR(`at`) is reported on the reserved validation split, whereas MR(`isAt`) is reported on v1.0 `isAt-test`. Bootstrap intervals are 95% CIs.

Run	MR(`at`)	95% CI	MR(`isAt`)	95% CI	Size (GB)
Run 1	0.7765	[0.6953, 0.8590]	0.8411	[0.7747, 0.9041]	195.716
Run 2	0.6189	[0.5375, 0.7039]	0.7782	[0.6999, 0.8521]	1.061
Run 3	0.7198	[0.6275, 0.8107]	0.7854	[0.7072, 0.8609]	3.217

Accuracy comparison. Run 1 is the strongest submitted configuration on both tasks. It reaches 0.7765 MR on `at` and 0.8411 MR on `isAt`, while Run 3 is consistently second and Run 2 is the weakest of the three on validation performance. The gap between Run 1 and Run 2 is especially large on `at`, where removing LLM-based reasoning, stacking, and voting costs about 0.16 MR.

Contribution of model combination. The main benefit of Run 1 comes from combining complementary predictors rather than relying on any single component alone. On `at`, Run 1 improves over the 4-model lookup-table stacker (0.7269) by 0.0496 MR, suggesting that the additional XLM-RoBERTa signal helps the final plurality decision. On `isAt`, Run 1 also slightly improves over Gemma alone (0.8273 to 0.8411), which is consistent with a small but useful complementarity between XLM-RoBERTa-large, Gemma 4 31B, and Llama 3.3 70B. By contrast, Run 3 stays close to the LLM-free reference point: on `at` it remains near the stacker-level range, and on `isAt` it stays close to the corresponding local alternative.

Efficiency comparison. The three runs define a clear efficiency hierarchy. Run 2 is the lightest deployable system at 1.061 GB and serves as the simplest single-model baseline. Run 3 remains fully local at 3.217 GB while recovering most of Run 1's `at` performance. Run 1 is by far the heaviest configuration, but it is also the only one that consistently gives the strongest validation performance on both tasks.

6.2. Headline Numbers

In the official HIPE-2026 evaluation across all three impresso languages (German, English, French), we (INSA Lyon) submitted three runs that trade off performance, generalization, and computational footprint. We first report the leaderboard results of the submitted systems, then provide a post-hoc analysis against the released gold labels (Section 6.3), which also makes component-level comparison possible.

Accuracy Profile (test-set performance on impresso v1.0): Among the submitted systems, Run 1 achieved the strongest impresso result, with overall MR 0.639 and rank 11th out of 45 submissions. Run 3 scored 0.4731 (rank 33) and Run 2 scored 0.4708 (rank 34). The large gap between Run 1 and the two lighter systems shows the value of the full heterogeneous ensemble, whereas the near tie between Run 2 and Run 3 suggests that the local ensemble brings little advantage over the compact single-model baseline on this test set. As shown below, this run-level ranking does not fully match the component-level ranking under the released official gold labels.

Generalization Profile (surprise-test-fr French unseen data): On the unseen French surprise test, Run 1 again remained the strongest INSA Lyon submission, with MR 0.4705 and rank 26th out of 46 runs. Run 2 reached 0.4231 (rank 29) and Run 3 reached 0.3986 (rank 30). Relative to its impresso score, Run 1 drops by 0.168, which shows that its stronger in-domain performance transfers only partially to the surprise setting. Within our submitted set, however, it still generalizes better than the two lighter alternatives.

Accuracy-Efficiency Profile: The Accuracy-Efficiency ranking highlights the cost of stronger performance. Run 2 is the most efficient submission, ranking 10 with a 1.061 GB footprint and 278M parameters. Run 3 ranks 21 with 3.217 GB and 839M parameters, remaining fully local while recovering part of the ensemble benefit. Run 1 ranks 25 with 195.7 GB and 102B parameters, reflecting the cost of the LLM components.

6.3. Post-hoc Error Analysis

After release of the official gold labels, we performed a post-hoc pair-level analysis by matching each prediction to its gold pair using the tuple (document_id, pers_entity_id, loc_entity_id). We re-scored not only the three official submissions, but also the main ensemble components and development candidates preserved in the repository.

Table 4 summarizes the label distribution of the pooled impresso test set and of surprise-test-fr. The table also shows why transfer analysis on surprise-test-fr is most informative for MR(at), as the released isAt split contains no positive instances.

Table 4
Gold-label distribution of the released test sets used in the post-hoc analysis.

Split	Task	TRUE	PROBABLE	FALSE	Total
impresso	at	214	85	339	638
impresso	isAt	213	–	425	638
surprise-test-fr	at	130	60	290	480
surprise-test-fr	isAt	0	–	480	480

Table 5 compares the three submitted runs with two non-submitted reference systems: pure Gemma and the 4-model lookup stacker. The main finding is simple: under the released official gold labels, Pure Gemma performs better than all three submitted runs on pooled impresso and also transfers better to surprise-test-fr. By contrast, the 4-model lookup stacker does not outperform its strongest component.

Table 5
Task-level post-hoc breakdown on the released official test labels. `Impresso` scores are computed on pooled pair-level predictions across the three languages; `surprise-test-fr` is reported as MR(`at`) only because the released `isAt` split contains no positive support. The first two rows are non-submitted reference systems; the remaining rows are the three official submissions.

Run	Impresso MR(`at`)	Impresso MR(`isAt`)	Impresso pooled global	Surprise-fr MR(`at`)
Pure Gemma	0.5625	0.8109	0.6867	0.5352
4-model stacker	0.4659	0.6595	0.5627	0.4611
Run 1	0.4953	0.7991	0.6472	0.4705
Run 2	0.4208	0.5304	0.4756	0.4231
Run 3	0.3730	0.5864	0.4797	0.3986

Table 6 makes the component-level comparison explicit on pooled `impresso`. The ranking is notable in two ways. First, Gemma is the strongest standalone model on both tasks, not just on the global score. Second, the combinations that looked promising during development do not remain dominant after gold release: the lookup stacker is clearly worse than Gemma on `at`, and the majority-vote `isAt` ensemble used in Run 1 is slightly worse than Gemma alone on `isAt`. In other words, the official submission logic improved robustness within the submitted set, but not beyond the best zero-shot component.

Table 6
Component-level comparison on pooled `impresso` test pairs (Test A). Global score is the mean of MR(`at`) and MR(`isAt`).

Component	MR(`at`)	MR(`isAt`)	Global
Gemma 4 31B	0.5625	0.8109	0.6867
Llama 3.3 70B	0.5138	0.7708	0.6423
Run 1 voting scheme	0.4953	0.7991	0.6472
4-model stacker	0.4659	0.6595	0.5627
mDeBERTa-v3	0.4208	0.5304	0.4756
xlm-RoBERTa-large	0.3416	0.5864	0.4640

Table 7
Development-to-official shift on pooled `impresso` MR(`at`). For the trained components, the development reference is cross-validated MR; for Gemma, it is the zero-shot development estimate used in the repository analysis.

Component	Development reference	Official test	Shift
RF (handcrafted)	0.5844	0.3473	-0.2371
C4-SDov	0.5787	0.4187	-0.1600
OrdContM1	0.5545	0.3785	-0.1760
Gemma 4 31B	0.5506	0.5625	+0.0119
4-model stacker	0.6443	0.4659	-0.1784

Table 7 makes this reversal in ranking more explicit. On the `at` task, the trained components lose substantial macro-recall from development to official test: RF drops from 0.5844 to 0.3473, C4-SDov from 0.5787 to 0.4187, OrdContM1 from 0.5545 to 0.3785, and the final lookup stacker from 0.6443 to 0.4659. By contrast, pure Gemma is the only major component whose official-test score remains close to its development estimate, with a shift of +0.0119 MR. The most plausible interpretation is severe train-to-test mismatch: the models trained on the 1,251 labeled instances learn patterns that do not transfer reliably to the released official distribution, whereas the zero-shot LLM remains comparatively stable. While this finding raises the question of why Gemma was not submitted as a standalone run, the ensemble-first strategy reflected our validation-time assumptions: the heterogeneous ensemble outperformed individual components on the held-out validation set, making it the natural choice for submission before the official test distribution became known.This ranking, however, did not carry over to the official evaluation: Run 1's MR(at) dropped from 0.7765 on validation to 0.4953 on the official

labels, whereas zero-shot Gemma remained stable (0.5506 → 0.5625). This reversal is precisely the validation-to-official-test discrepancy discussed above.

At the same time, the problem is not simply lack of diversity. The four `at` stacker components agree on only 28.9% of the 1,118 official-test pairs, and an oracle that always selected the correct component would reach 87.9% plain accuracy. In other words, there is ample complementary evidence in principle, but the learned combination rule does not identify the reliable component under distribution shift. Post-hoc, this makes the ensemble result look less like a failure of diversity and more like a failure of selection.

Within the three official outputs, the most important class-level finding is shown in Table 8. All three submitted systems collapse on the minority PROBABLE label in the `at` task, with zero recall in the final official outputs. Thus, most of the observed `at` performance comes from separating TRUE and FALSE, while the uncertain middle class remains unresolved. Even the stronger standalone components recover this class only marginally on pooled `impresso`: Gemma reaches recall 0.0118 and Llama 0.0588, confirming that the difficulty is corpus-level rather than specific to a single submitted run.

Table 8
Per-class recall on pooled `impresso` test pairs. All three submitted runs show zero recall on the PROBABLE class in the final official outputs.

Run	at: TRUE	at: PROBABLE	at: FALSE	isAt: TRUE	isAt: FALSE
Run 1	0.6776	0.0000	0.8083	0.7981	0.8000
Run 2	0.4953	0.0000	0.7670	0.1268	0.9341
Run 3	0.2991	0.0000	0.8201	0.5352	0.6376

Figure 2 visualizes the same result. Among the submitted runs, the full ensemble (Run 1) is strongest not because it recovers PROBABLE, but because it keeps substantially higher recall on positive TRUE cases in both tasks while maintaining competitive recall on FALSE. The compact baseline (Run 2) is especially conservative on `isAt`, where it achieves high recall on FALSE but very low recall on TRUE. The frugal ensemble (Run 3) is less conservative on `isAt`, but falls further behind on `at`.

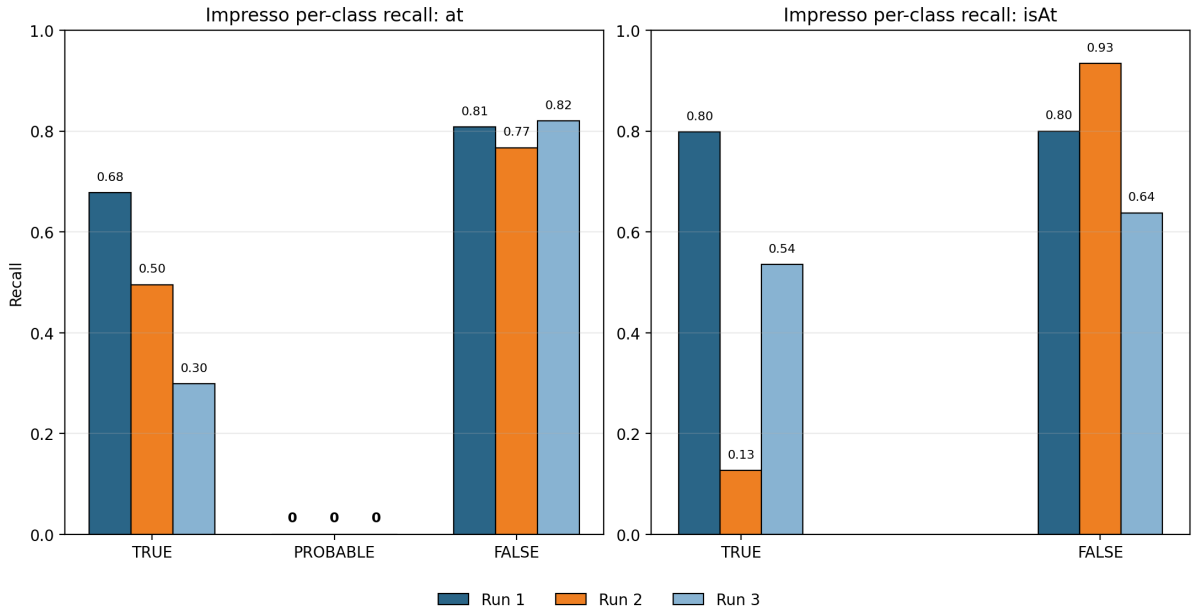

Figure 2: Per-class recall on pooled `impresso` test pairs for the three official runs. The plot makes the zero-recall collapse on PROBABLE visually explicit and highlights the stronger positive-class recall of Run 1 on both tasks.

The confusion patterns (see Figure 5) show where these errors come from. For the full ensemble, the dominant `at` errors on `impresso` are PROBABLE → TRUE (58 cases) and PROBABLE → FALSE (27 cases). In other words, the system does not preserve the middle category at decision time; it pushes

ambiguous cases toward one of the two extreme labels. The compact baseline shows the same failure more strongly (PROBABLE → TRUE: 50; PROBABLE → FALSE: 35), while the frugal ensemble is the most pessimistic system on `at`, with TRUE → FALSE as its largest non-diagonal confusion (144 cases).

Generalization to unseen French is also uneven across gold labels. Figure 3 shows the `at` error rate of the full ensemble by gold label on pooled `impresso` and on `surprise-test-fr`. The main transfer loss comes from gold TRUE cases, whose error rate rises under transfer. By contrast, gold PROBABLE already fails completely on `impresso` and therefore remains unchanged on the surprise split. This pattern is consistent with the broader component-level picture: transfer degrades most sharply in models whose decision boundaries were tuned on the small labeled pool, while the zero-shot LLM baseline remains more stable across test conditions.

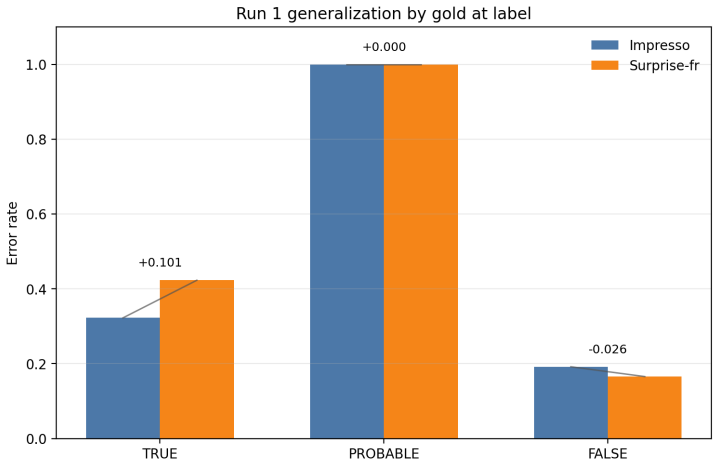

Figure 3: Run 1 error rate by gold `at` label on pooled `impresso` versus `surprise-test-fr`. The plot shows where generalization degrades: the error rate rises for gold TRUE, remains maximal for gold PROBABLE, and decreases slightly for gold FALSE.

To identify where the strongest submitted system still fails often, we inspect the full ensemble at finer granularity. Figure 4 reports error rates by language, entity-linking status, and mention multiplicity on pooled `impresso`. According to our results, our model performs worse on English text for the full ensemble on both tasks. Second, errors concentrate in less well anchored pairs: pairs with only one Wikidata-linked argument are harder than pairs with both arguments linked, and pairs with multiple

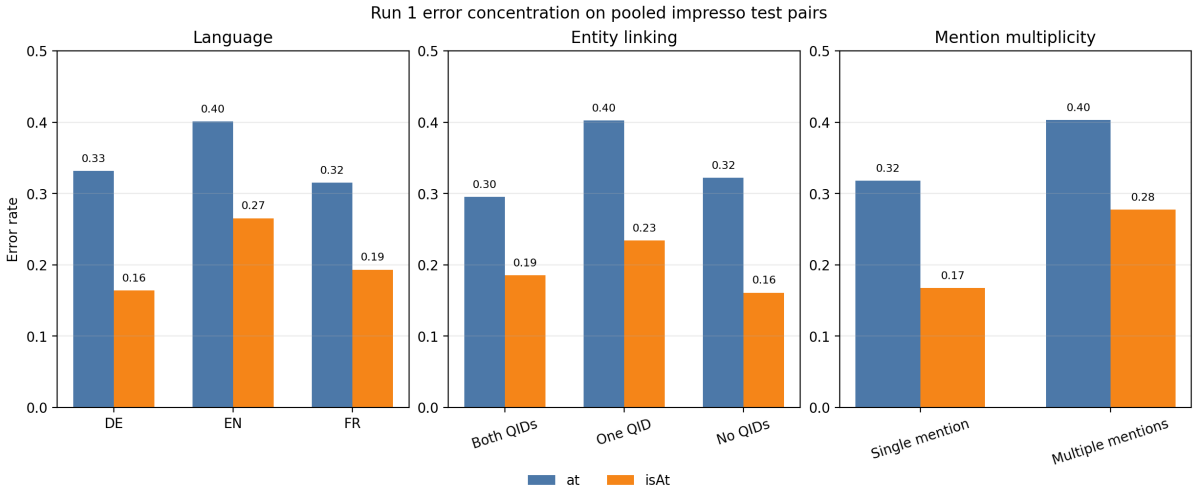

Figure 4: Error concentration for Run 1 on pooled `impresso` test pairs, broken down by language, entity-linking status, and mention multiplicity. Higher values indicate subsets where the strongest submitted system still fails more often.

mention strings are harder than single-mention pairs.

Table 9 gives the same bucket analysis numerically. The differences are substantial. For the `at` task, the error rate rises from 0.295 in the fully linked bucket to 0.402 in the partially linked bucket, and from 0.318 for single-mention pairs to 0.403 for multiple-mention pairs. For the `isAt` task, the same direction holds, although the gaps are smaller. The same tendency also appears on `surprise-test-fr`, where the highest `at` error rate is found in pairs with no linked arguments (0.442).

Table 9
Error-rate breakdown for Run 1 on pooled `impresso` test pairs. Error rate is the proportion of pair-level predictions that do not match the released gold label.

Bucket family	Subset	N pairs	`at` error rate	`isAt` error rate
Language	DE	238	0.332	0.164
Language	EN	162	0.401	0.265
Language	FR	238	0.315	0.193
Linking	Both arguments linked	264	0.295	0.186
Linking	One argument linked	256	0.402	0.234
Linking	No linked arguments	118	0.322	0.161
Mentions	Single-mention pairs	447	0.318	0.168
Mentions	Multiple-mention pairs	191	0.403	0.277

Finally, Figure 5 presents the standard row-normalized confusion matrices for the full ensemble on pooled `impresso`. The matrices make the two dominant residual error modes explicit. For `at`, the entire PROBABLE row is off-diagonal, confirming that the final decision layer does not preserve the middle class. For `isAt`, the remaining errors are concentrated on the binary boundary itself, with substantial mass in both FALSE → TRUE and TRUE → FALSE.

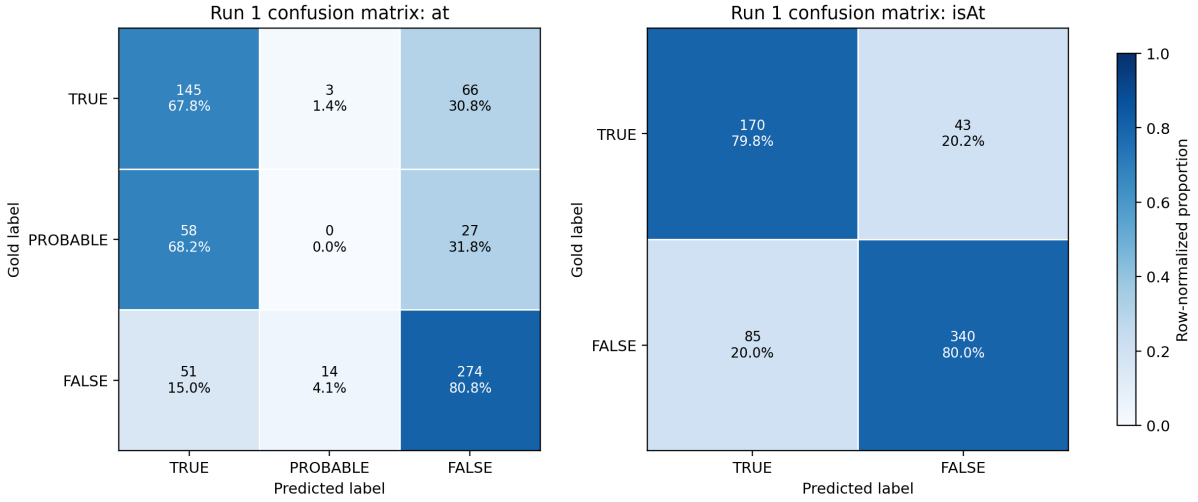

Figure 5: Standard confusion matrices for Run 1 on pooled `impresso` test pairs. Cells are annotated with raw counts and row-normalized percentages.

Taken together, the post-hoc analysis yields four main conclusions. First, Run 1 remains the strongest official submission, but pure Gemma would have been the strongest standalone system in our repository on the released official gold labels. Second, all final submitted outputs fail on the minority PROBABLE class, and even stronger components recover it only marginally. Third, the remaining errors concentrate at the positive-vs-negative decision boundary and in pairs with weaker entity grounding or more dispersed mention structure. Fourth, the largest system-level weakness is distribution shift: the trained components and their lookup-based combination degrade sharply from development estimates to official-test performance, whereas the zero-shot LLM baseline remains comparatively stable.

7. Limitations and Future Work

The current system has three main limitations: sensitivity to distribution shift, weak handling of uncertainty, and a large deployment gap between the strongest and most practical configurations.

Distribution shift. Although the heterogeneous ensemble was designed to combine complementary signals, the post-hoc comparison in Tables 5 and 6 shows that several trained components degraded sharply once evaluated on the released official test set. The central weakness is therefore not the absence of model diversity, but the instability of the learned combination under changing test conditions. In particular, the lookup-table stacker depends on patterns estimated from a small labelled pool and does not reliably preserve the strongest component when the distribution shifts.

Uncertainty modelling. Table 8 and Figure 2 show that all three submitted runs collapse on the PROBABLE class, indicating that the current pipeline behaves much better as a polar classifier than as a graded relation classifier. This is partly a consequence of the present decision design: plurality voting with ordinal tie-breaking is robust for aggregate decisions, but it does not protect a minority middle class when evidence is sparse or ambiguous.

Deployment cost. Run 1 is the strongest submission, but also the most expensive to deploy. Its total footprint is around 100B parameters and nearly 200 GB, making the local alternatives substantially more practical for reproduction and large-scale use.

The post-hoc analysis suggests a consistent explanation for these limitations. Tables 5 and 6 show that Pure Gemma is the strongest overall test-set performer after gold release, which indicates that broad zero-shot reasoning transferred more reliably than the smaller supervised components. At the same time, Figures 2 and 5 show that ambiguity is usually resolved by collapsing PROBABLE cases toward TRUE or FALSE, while Figure 3 and Figure 4 show that the remaining errors concentrate on gold TRUE cases and weakly anchored instances. These findings suggest three near-term directions for future work: a more robust decision layer under distribution shift, more explicit modelling of ordinal uncertainty, and stronger local models that retain some of the robustness of Gemma without relying on external LLM. Beyond component improvements, the setup should also be extended toward end-to-end evaluation, where OCR noise, mention multiplicity, entity linking, and candidate generation are treated as part of the modeling problem rather than fixed inputs.

8. Conclusion

Through our participation in the HIPE-2026 shared task, we proposed a modular heterogeneous ensemble pipeline combining multilingual Transformer encoders, zero-shot large language models, and handcrafted features, and submitted three runs representing different performance–efficiency trade-offs. Run 1 (full heterogeneous ensemble) achieved a macro-recall of 0.639 on the official impresso test set, ranking 11th out of 45 submissions, while Run 2 (compact single-encoder baseline) and Run 3 (frugal local ensemble without LLM inference) provided competitive efficiency alternatives with 278M and 839M parameters respectively. Despite these encouraging results, post-hoc analysis against the released gold labels revealed three systematic limitations.

First, all submitted systems show zero recall on the minority PROBABLE class in the `at` task, indicating that the current decision layer fails to preserve the uncertain middle category. Second, trained components degrade substantially from validation to official test performance, whereas the zero-shot Gemma baseline remains comparatively stable, pointing to train-to-test distribution shift as the central challenge. Third, errors concentrate in pairs with low entity linking coverage and under cross-domain generalization to unseen French literary text (Test B).

Addressing these limitations defines the main directions for future work: building a more robust decision layer under distribution shift, explicitly modelling ordinal uncertainty to improve recall on

the PROBABLE class, and developing stronger local models that retain the generalization robustness of Gemma without the cost of hosted inference.

Acknowledgments

Thanks to the developers of ACM consolidated LaTeX styles https://github.com/borisveytsman/acmart, the developers of Elsevier updated LaTeX templates https://www.ctan.org/tex-archive/macros/latex/contrib/els-cas-templates, and the PAGODA platform (https://jupyterhub.pagoda.liris.cnrs.fr/) of LIRIS, CNRS, France.

Declaration on Generative AI

During the preparation of this work, the authors used generative AI tools, including GPT-5.4 and Claude Sonnet 4.5, for translation, grammar correction, wording refinement, and limited drafting assistance based on author-provided ideas and analyses. These tools were not used to generate the scientific contributions, experimental results, or conclusions of the paper. All outputs were reviewed, revised, and validated by the authors, who take full responsibility for the final manuscript.

References

[1] J. Opitz, M. Ehrmann, C. Raclé, A. Michail, M. Romanello, S. Clematide, Overview of HIPE-2026: Person–Place Relation Extraction from Multilingual Historical Texts, in: M. Hagen, M. Potthast, B. Stein, P. Schaer, E. Zangerle, S. MacAvaney, J. M. Struß, E. Sánchez Salido, A. Barrón-Cedeño, A. García Seco de Herrera (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Seventeenth International Conference of the CLEF Association (CLEF 2026), Lecture Notes in Computer Science (LNCS), Springer, 2026. URL: https://doi.org/10.1007/978-3-032-21321-1_46. doi:10.1007/978-3-032-21321-1_46.

[2] J. Opitz, M. Ehrmann, C. Raclé, A. Michail, M. Romanello, E. Boros, S. Gabay, S. Clematide, Extended Overview of HIPE-2026: Evaluating Accurate and Efficient Person–Place Relation Extraction from Multilingual Historical Texts, in: E. Sánchez Salido, A. Barrón-Cedeño, A. García Seco de Herrera, S. MacAvaney, J. M. Struß (Eds.), CLEF 2026 Working Notes, CEUR Workshop Proceedings, CEUR-WS, 2026. doi:10.5281/zenodo.20344461.

[3] M. Ehrmann, A. Hamdi, E. Linhares Pontes, M. Romanello, A. Doucet, Named entity recognition and classification in historical documents: A survey, ACM Computing Surveys 56 (2023) 1–47. doi:10.1145/3604931.

[4] M. Ehrmann, M. Romanello, S. Bircher, S. Clematide, Introducing the CLEF 2020 HIPE shared task: Named entity recognition and linking on historical newspapers, in: Advances in Information Retrieval. ECIR 2020, volume 12036 of *Lecture Notes in Computer Science*, Springer, Cham, 2020, pp. 524–532. doi:10.1007/978-3-030-45442-5_68.

[5] M. Ehrmann, M. Romanello, A. Doucet, S. Clematide, Introducing the HIPE 2022 shared task: Named entity recognition and linking in multilingual historical documents, in: Advances in Information Retrieval. ECIR 2022, volume 13186 of *Lecture Notes in Computer Science*, Springer, Cham, 2022, pp. 347–354. doi:10.1007/978-3-030-99739-7_44.

[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems, volume 30, Curran Associates, Inc., 2017, pp. 5998–6008. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[7] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long

and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423/. doi:10.18653/v1/N19-1423.

[8] S. Schweter, L. März, K. Schmid, E. Çano, hmBERT: Historical multilingual language models for named entity recognition, in: CLEF 2022 Working Notes, CEUR-WS, 2022, pp. 1109–1129. URL: https://ceur-ws.org/Vol-3180/paper-87.pdf.

[9] E. Boros, C.-E. González-Gallardo, E. Giamphy, A. Hamdi, J. G. Moreno, A. Doucet, Knowledge-based contexts for historical named entity recognition & linking., in: CLEF (Working Notes), 2022, pp. 1064–1078. URL: https://ceur-ws.org/Vol-3180/paper-84.pdf.

[10] J. A. Diaz-Garcia, J. A. D. Lopez, A survey on cutting-edge relation extraction techniques based on language models, Artificial Intelligence Review 58 (2025) 287. doi:10.1007/s10462-025-11280-0.

[11] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, in: Advances in Neural Information Processing Systems, volume 33, Curran Associates, Inc., 2020, pp. 1877–1901. URL: https://arxiv.org/abs/2005.14165.

[12] S. C. de Abreu, T. L. Bonamigo, R. Vieira, A review on relation extraction with an eye on portuguese, Journal of the Brazilian Computer Society 19 (2013) 553–571. URL: https://doi.org/10.1007/s13173-013-0116-8. doi:10.1007/s13173-013-0116-8.

[13] K. Detroja, C. Bhensdadia, B. S. Bhatt, A survey on relation extraction, Intelligent Systems with Applications 19 (2023) 200244. URL: https://doi.org/10.1016/j.iswa.2023.200244. doi:10.1016/j.iswa.2023.200244.

[14] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, Journal of Machine Learning Research 21 (2020) 1–67. URL: http://jmlr.org/papers/v21/20-074.html.

[15] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, Improving Language Understanding by Generative Pre-Training, Technical Report, OpenAI, 2018. URL: https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf.

[16] J. Delaunay, H. T. H. Tran, C.-E. González-Gallardo, G. Bordea, N. Sidere, A. Doucet, A comprehensive survey of document-level relation extraction (2016–2023), 2023. URL: https://arxiv.org/abs/2309.16396. arXiv:2309.16396.

[17] Y. Yao, D. Ye, P. Li, X. Han, Y. Lin, Z. Liu, Z. Liu, L. Huang, J. Zhou, M. Sun, DocRED: A large-scale document-level relation extraction dataset, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 764–777. URL: https://doi.org/10.18653/v1/p19-1074. doi:10.18653/v1/p19-1074.

[18] C. Yuan, Q. Xie, S. Ananiadou, Zero-shot temporal relation extraction with ChatGPT, in: Proceedings of the 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks, Association for Computational Linguistics, Toronto, Canada, 2023, pp. 92–102. URL: https://aclanthology.org/2023.bionlp-1.7/. doi:10.18653/v1/2023.bionlp-1.7.

[19] L. Breiman, Random forests, Machine Learning 45 (2001) 5–32. doi:10.1023/A:1010933404324.

[20] P. He, J. Gao, W. Chen, DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing, in: International Conference on Learning Representations, 2023. URL: https://openreview.net/forum?id=sE7-XhLxHA, iCLR 2023.

[21] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 8440–8451. URL: https://aclanthology.org/2020.acl-main.747/. doi:10.18653/v1/2020.acl-main.747.

[22] Gemma Team, Gemma 4 model card, 2026. URL: https://ai.google.dev/gemma/docs/core/model_card_4.

[23] A. Grattafiori, et al., The Llama 3 herd of models, 2024. URL: https://arxiv.org/abs/2407.21783. arXiv:2407.21783.

[24] D. R. Cox, The regression analysis of binary sequences, Journal of the Royal Statistical Society: Series B 20 (1958) 215–232. doi:10.1111/j.2517-6161.1958.tb00292.x.

A. Handcrafted Features

#	Feature	Description
1	verb_is_present	Whether any extracted verb cluster is in present tense.
2	verb_is_past	Whether any extracted verb cluster is in past tense.
3	verb_is_negated	Whether any extracted verb cluster is negated.
4	verb_is_future	Whether any extracted verb cluster is in future tense.
5	sentence_position_norm	Normalised sentence position of the pair mention.
6	is_lead_paragraph	Whether the pair appears in the lead part of the article.
7	signal_negation	Whether the temporal-signal category is negation.
8	signal_present	Whether the temporal-signal category is present.
9	signal_relative	Whether the temporal-signal category is relative-only.
10	signal_none	Whether no temporal signal is detected.
11	has_timex_in_window	Whether a temporal expression falls within the isAt document-time window.
12	nearest_timex_distance_norm	Normalised distance to the nearest temporal expression.
13	person_is_deceased	Whether the person status is dead-historical.
14	person_status_known	Whether a non-empty person-status category is available.
15	ocr_quality	Precomputed OCR-quality score.
16	lang_en	English language indicator.
17	lang_fr	French language indicator.
18	lang_de	German language indicator.
19	lang_lb	Luxembourgish language indicator.
20	person_status_unknown	Person-status one-hot feature.
21	person_status_no_match	Person-status one-hot feature.
22	person_status_alive_past	Person-status one-hot feature.
23	person_status_alive_now	Person-status one-hot feature.
24	person_status_alive_no_longer	Person-status one-hot feature.
25	person_status_dead_historical	Person-status one-hot feature.
26	person_status_timex_in_lifespan	Person-status one-hot feature.
27	person_status_timex_after_death	Person-status one-hot feature.
28	has_pers_qid	Whether the person mention has a Wikidata QID.
29	has_loc_qid	Whether the location mention has a Wikidata QID.
30	n_pers_mentions	Number of person mentions in the instance.
31	n_loc_mentions	Number of location mentions in the instance.
32	n_temporal_expressions	Number of detected temporal expressions.
33	n_temporal_signals	Number of detected temporal signals.
34	n_tense_aspect_clusters	Number of extracted tense/aspect clusters.
35	text_len_chars_norm	Normalised full-text length.
36	context_len_chars_norm	Normalised local-context length.

Table 10
Hand-crafted features used in the handcrafted part of the pipeline.

B. Prompts

```
## System Prompt

You are an expert annotator for historical named-entity relation extraction
in the HIPE-2026 shared task.

CONTEXT:
- Documents are historical newspaper articles from 1798-2018.
- Languages: English, French, German (read as-is - do NOT translate).
- OCR may introduce minor noise; focus on clear verbal structures and entities.

YOUR TASK:
Classify TWO relations for a given person-location pair:

1. "at" - GENERAL ASSOCIATION
   Does this person have any connection to this location?
   (born there, lived there, worked there, visited, etc.)
   Values: TRUE / PROBABLE / FALSE

2. "isAt" - PHYSICAL PRESENCE AT PUBLICATION TIME
   Was this person physically present at this location around the time the
   article was published (within approximately one month)?
   Values: TRUE / FALSE

LOGICAL CONSTRAINT: If isAt=TRUE, then at must be TRUE or PROBABLE.
(If someone is currently at a place, they have an association with it.)

DECISION FRAMEWORK:

For "at" (general association):
  - Direct textual evidence of connection -> TRUE
  - Implicit or uncertain connection -> PROBABLE
  - No connection -> FALSE
  - Person deceased before publication -> FALSE

For "isAt" (temporal presence):
  - Resolve all temporal expressions relative to the publication date
  - Present tense in news context -> supports TRUE
  - "yesterday"/"hier"/"gestern" -> resolve to pub_date - 1 -> within window -> TRUE
  - Past tense with distancing language -> FALSE
  - Person deceased -> FALSE
  - No temporal anchor within ~1 month of publication -> FALSE

LANGUAGE-SPECIFIC NOTES:
  - French: passe compose is ambiguous; present historique may be stylistic, not
    temporal -- check surrounding tense. "ancien" (former) -> FALSE.
  - German: Perfekt -> recent (TRUE-leaning); Praeteritum -> narrative past (neutral).
  - English: present perfect / progressive -> TRUE-leaning; "former", "late" -> FALSE.

OUTPUT FORMAT:
Respond with exactly one line:
  at=LABEL isAt=LABEL | conf_at=X.XX conf_isAt=X.XX

Example: at=TRUE isAt=FALSE | conf_at=0.92 conf_isAt=0.85
Do NOT output explanations or any other text.

## User Message
```

```
Publication date: [PUBLICATION_DATE]
Language: [LANGUAGE]

[ENTITY PAIR]
Person: '[PERSON_SURFACE]' (QID: [PERSON_QID_OR_NONE])
Location: '[LOCATION_SURFACE]' (QID: [LOCATION_QID_OR_NONE])

[ARTICLE TEXT]
"[ARTICLE_TEXT_WITH_<PERSON>...</PERSON>_AND_<LOCATION>...</LOCATION>_MARKERS]"

Classify both the "at" and "isAt" relations for this person-location pair.
```