Where Were They? Person-Place Relation Extraction in Multilingual Historical Archives*

Ron Keinan[1,*], Reut Tsarfaty[1]

[1]Bar Ilan University, Ramat-Gan, Israel

Abstract

Relation extraction (RE) is the automatic identification of semantic relationships between named entities in text. Large language models (LLMs) have recently made RE accessible through zero- and few-shot prompting, with strong performance, without the need for task-specific fine-tuning. LLM-based RE is particularly valuable in multilingual settings, where annotated training data is scarce for most language pairs, yet unlike in English, performance often degrades in non-English and lower-resource settings, even for models with broad multilingual coverage. These difficulties are further compounded in historical document archives, where OCR digitisation introduces systematic character-level noise, 19th-century spelling and syntax diverge sharply from modern pre-training distributions, and domain-specific adaptation is rarely feasible. We investigate this challenging intersection across historical newspaper corpora in German, English, and French, qualifying annotated person–place pairs along two relations, `at` (was the person at the location?) and `isAt` (was the person at the location around the time of publication?), as part of the HIPE 2026 evaluation lab at CLEF 2026. Our approach centers on carefully engineered multilingual prompts: we design a unified cross-lingual prompting scheme, with explicit relation definitions, evidence-grounded explanations, and OCR-robustness rules, and evaluate it systematically across all three document languages with no task-specific fine-tuning. A central design question we address empirically is whether prompting in a high-resource language (English) or in the native source language (German, French) better serves performance across models and language pairs. All in all, our best submission ranks **9th out of 46** for French and **20th overall**; on the out-of-domain surprise set, our most generalisable model ranks **3rd out of 44** in binary evaluation, demonstrating that careful prompt engineering alone yields competitive, cross-lingual performance, with encouraging but uneven generalisation to an out-of-domain test set. We present a structured multilingual prompting approach for historical person–place RE that performs competitively as a zero-shot baseline, though performance varies substantially across models and does not always transfer from development to test — an outcome we analyse in detail.

Keywords

relation extraction, historical newspapers, large language models, multilingual NLP, zero-shot prompting, few-shot learning, person–place relations, HIPE 2026

1. Introduction

Relation extraction (RE) is the task of automatically identifying and classifying semantic relationships between named entities in unstructured text, underpinning knowledge graph construction, database population, and downstream applications such as question answering and fact verification. Classical RE systems required large task-specific annotated datasets, making them expensive to extend to new domains or languages. The emergence of large language models (LLMs) has transformed this picture: through in-context learning, modern LLMs can perform competitive RE without updating any model weights, enabling rapid, annotation-free deployment across new settings.

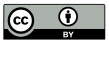
This zero-shot capability is especially valuable in multilingual settings: decades of NLP progress have been concentrated on English, leaving most other languages without the supervised benchmarks needed to train dedicated RE systems [1], and even broadly multilingual LLMs exhibit performance drops outside English, sensitive to linguistic distance and pre-training coverage.

These difficulties compound further in historical newspaper archives, where OCR digitisation introduces systematic noise and archaic spelling conventions diverge from modern pre-training distributions — challenges detailed in Section 2 — making zero-shot prompting especially attractive as an approach.

CLEF 2026 Working Notes, 21 – 24 September 2026, Jena, Germany

*Corresponding author.

ronke21@gmail.com (R. Keinan); reut.tsarfaty@biu.ac.il (R. Tsarfaty)



We explore this intersection in a systematic study targeting person–place relation qualification in historical newspaper corpora across German, English, and French. The task requires classifying each annotated pair along two semantically subtle, class-imbalanced relations, `at` (was the person at the location?) and `isAt` (was the person at the location around the time the article was published?), demanding multi-step inference rather than surface matching. This investigation is conducted as part of the **HIPE 2026 evaluation lab**[1] at CLEF 2026, a competitive shared task with 17 participating teams, who together submitted 45 runs (46–47 submissions per language-specific ranking category), evaluated by global macro-recall [2, 3].

At the core of our approach is a carefully engineered multilingual prompting framework operating across all three document languages with no task-specific fine-tuning, incorporating explicit relation definitions, evidence-grounded explanations, OCR-robustness rules, and structured JSON output. A key empirical question is the effect of *prompt language*: prompting in a high-resource language (English) versus the native document language (German, French). We systematically evaluate this alongside shot count (0, 3, 5, and 10 in-context examples) across all four models, selecting the best configuration per model and language on the development set.

Our submissions demonstrate that carefully crafted zero-shot LLM prompting is a strong strategy: our best run ranks **9th out of 46** for French and **20th overall**, and on the out-of-domain surprise test set our most generalisable model ranks **3rd out of 44** in binary evaluation, strong on French and out-of-domain generalisation, mid-table on aggregate. We further find that English-language prompts generalise well across all document languages, observe a pattern consistent with few-shot examples benefiting our sparse Mixture-of-Experts model more than the denser models tested (though with only one MoE architecture evaluated, this remains a hypothesis rather than a controlled finding), and find that development rankings do not always transfer to the test set.

Our primary contribution is a systematic empirical investigation of multilingual LLM-based RE under historical OCR conditions; the following enumerate the specific components and findings it yields.
Our contributions are:

1. A carefully engineered (not fine-tuned) cross-lingual prompting scheme for historical person–place RE incorporating explicit relation definitions, OCR-robustness rules, evidence-grounded explanations, and structured JSON — serving as the controlled experimental vehicle for contributions 2–4 and as a reproducible baseline for the community.
2. A systematic empirical study of prompt language (English vs. native) and shot count (0–10) across four open-weight models and three languages, yielding actionable insights on when native-language prompts help and when English generalises better.
3. A qualitative error analysis identifying four recurring error types: PROBABLE/TRUE over-prediction, origin-location confusion, OCR-induced parse failures, and cross-entity context bleed.
4. A fully reproducible baseline: all code, prompts, and result files are publicly released at https://github.com/Ronke21/HIPE-2026-Evaluating-Person-Place-Relation-Extraction, addressing the reproducibility gap in multilingual RE research [4].

Sections 2–3 describe the challenge and task; Section 4 covers data and experimental setup; Section 5 reports results; Section 6 discusses findings and error analysis; Section 7 covers related work and future directions; Section 8 concludes.

2. The Challenge

Multilingual LLMs underperform on RE outside English, with the gap widening for morphologically complex languages such as German and for domains underrepresented in pre-training [1, 5].

These challenges compound sharply in the historical newspaper domain. The HIPE 2026 corpus is digitised from microfilm, introducing OCR errors that corrupt the tokens a model needs to reason over. Common artefacts in the corpus include character substitutions arising from Gothic typefaces,

[1] https://hipe-eval.github.io/HIPE-2026/

the letter 'M' frequently rendered as 'rn', the historical long-s glyph misread as 'f', hyphenation at line boundaries that splits proper names across lines, and corrupted entity mentions. A representative sentence from the corpus might arrive as:

```
Herr v. Rnöller, Bürgermeifter zu Müncben, ift in Ber-
lin eingetroffen.
```

where `Rnöller` reflects Gothic M → rn confusion, `Bürgermeifter` the long-s → f substitution, `Müncben` a ch → cb substitution, and `Ber-lin` a hyphenation artefact, all compounding in a single entity-bearing sentence. Beyond OCR, 19th-century German uses pre-reform spelling conventions (*Thür* for *Tür*), and French sources of the same period employ period-specific typography and abbreviations absent from modern training corpora. Models pre-trained primarily on contemporary web text may support the languages in principle yet lack exposure to these historical registers. More broadly, linguistic phenomena that already challenge cross-lingual RE in contemporary settings, compounding, inflection, pro-drop, and variable word order, are further compounded when OCR noise corrupts the very morphological markers a model must resolve to identify entity-bearing tokens [6]. In this sense, historical newspaper archives present the characteristics of a low-resource domain even for languages that are otherwise well-resourced: archaic register, pre-reform spelling, and OCR-degraded surface forms place the text far outside modern pre-training distributions, effectively confronting the model with an out-of-distribution language variant [4, 7].

The `at` relation qualification task adds semantic difficulty on top of these surface challenges. Deciding whether a person was `at` a location requires inferring presence from indirect cues, a title implying a residence, a reported action entailing presence, a temporal phrase indicating a past visit, rather than simple entity co-occurrence. The `isAt` relation is subtler still: it requires evidence that the person was at the location around the time the article was published, not merely a past, one-off mention. Both relations are strongly class-imbalanced: `FALSE` dominates in most article–pair combinations, with minority positive classes (`at=TRUE`, `at=PROBABLE`, `isAt=TRUE`) carrying the discriminative signal. This skew is typical of distantly supervised multilingual RE [8, 9, 10]; the all-`FALSE` baseline and metric choice are formalised in Section 3.

Together, cross-lingual degradation, historical OCR noise, and class-imbalanced semantic subtlety define a setting where prompt engineering and in-context calibration matter as much as model choice.

3. The HIPE 2026 Task

3.1. Task Definition

At a high level, the task asks the following: given a historical newspaper article and a pre-identified (person, place) pair mentioned in it, determine two things. First, does the article provide evidence that the person was at the location at any point before publication, and if so how strong is that evidence? Second, was the person at that location around the time the article was published? The first question has three possible answers (`TRUE`, `PROBABLE`, or `FALSE`), depending on whether the evidence is direct, indirect, or absent. The second is binary (`TRUE` or `FALSE`) and strictly stronger: `isAt=TRUE` logically entails `at=TRUE` (being at a place around publication time requires having been there). A system must produce both labels for every (person, place) pair in every article. HIPE 2026 [2, 3] represents one of the few controlled evaluation environments for multilingual RE, providing consistent annotation protocols and a unified evaluation metric across three languages simultaneously, enabling direct cross-lingual comparison under a shared task infrastructure [11, 12].

We formalise this as follows.

Definition 1 (Person–Place Relation Qualification)**.** Let d be a historical newspaper article in language $\ell \in \{\mathrm{DE}, \mathrm{EN}, \mathrm{FR}\}$, and let $\mathcal{P}(d) = \{(e_p, e_l)\}$ be the set of pre-annotated *person–place pairs* in d, where e_p is a person mention and e_l is a location mention. The task is to learn a function

$$f : d, (e_p, e_l) \mapsto \left(\hat{y}_{\texttt{at}}, \hat{y}_{\texttt{isAt}}\right) \tag{1}$$

where $\hat{y}_{\texttt{at}} \in \mathcal{Y}_{\texttt{at}} = \{\text{TRUE}, \text{PROBABLE}, \text{FALSE}\}$ and $\hat{y}_{\texttt{isAt}} \in \mathcal{Y}_{\texttt{isAt}} = \{\text{TRUE}, \text{FALSE}\}$, subject to the *consistency constraint*:

$$\hat{y}_{\texttt{isAt}} = \text{TRUE} \;\Rightarrow\; \hat{y}_{\texttt{at}} = \text{TRUE} \tag{2}$$

No task-specific fine-tuning is applied; all model weights are fixed at inference time.

3.2. Evaluation Metric

A predicted label is evaluated by **exact string equality** against the gold annotation: the official evaluation script[2] compares the predicted label string (`"TRUE"`, `"PROBABLE"`, or `"FALSE"`) to the gold string character-by-character, with no fuzzy matching. This is important to note: OCR noise affects the entity mentions (e_p, e_l) that are given as input, but the evaluation operates only on the output label strings, which are controlled vocabulary. The gold labels were produced by human annotators who reviewed and corrected LLM-generated pre-annotations, and are treated as ground truth.

The primary metric is **global macro-recall** (GMR), computed as a two-level average: first, per-relation macro-recall (uniform average over that relation's label classes); then a uniform average across the two relations:

$$\text{GMR} = \tfrac{1}{2}\left[\frac{1}{|\mathcal{Y}_{\texttt{at}}|}\sum_{c\in\mathcal{Y}_{\texttt{at}}}\text{Recall}(c) \;+\; \frac{1}{|\mathcal{Y}_{\texttt{isAt}}|}\sum_{c\in\mathcal{Y}_{\texttt{isAt}}}\text{Recall}(c)\right] \tag{3}$$

Each of the three `at` classes and each of the two `isAt` classes contributes equally to the score. A system predicting all pairs as FALSE achieves $\text{GMR} = \frac{1}{2}\left(\frac{1}{3} + \frac{1}{2}\right) = \frac{5}{12} \approx \mathbf{0.4167}$ — the dominant-class baseline.

A secondary **binary evaluation** maps $\hat{y}_{\texttt{at}} = \text{PROBABLE} \rightarrow \text{TRUE}$, reducing $\mathcal{Y}_{\texttt{at}}$ to two classes (TRUE/FALSE). This pools all evidence of presence, whether the model was certain (TRUE) or uncertain (PROBABLE), into a single positive class, giving an upper-bound view of how well a system identifies any spatial connection at all.

Following the official HIPE 2026 terminology, submissions are evaluated along three **profiles**: the **Accuracy profile** (predictive performance on the impresso test sets, reported via the main and binary macro-recall above), the **Efficiency profile** (a composite rank over accuracy, model parameter count, and deployed model size), and the **Generalization profile** (predictive performance on the out-of-domain surprise-FR test set). We submit to the Accuracy and Generalization profiles; we do not submit an Efficiency profile entry.

4. Experimental Design

4.1. The Data

The data was provided by the Impresso project (https://impresso-project.ch), which digitized and semantically enriched historical European newspaper archives. The HIPE 2026 corpus extends the HIPE-2022 v2.1 named entity-annotated collection, covering newspapers in German, English, and French from roughly the 1780s to the 1950s. Named entity annotations include surface mention offsets, entity type (person, location), and Wikidata [13] QID links where available. Person–place pairs were first pre-annotated by an LLM ensemble, then manually reviewed and corrected by domain experts.

Table 1 shows the number of articles and annotated person–place pairs per language and split. The French portion of the training set is substantially larger than German and English combined, reflecting the composition of the underlying newspaper archives.

The **surprise test set** was drawn from a French literary works corpus not disclosed before submission, designed to probe out-of-domain generalisation.

[2] We use data release v1.0 and the official scorer, both available at https://github.com/hipe-eval/HIPE-2026-data/releases/tag/v1.0.

Table 1
Dataset statistics (articles / annotated pairs per split).

Split	German	English	French	Total
Train	88 / 1,224	56 / 496	317 / 4,450	461 / 6,170
Dev	32 / 432	17 / 151	107 / 1,498	156 / 2,081
Test (impresso)	19 / 238	19 / 162	19 / 238	57 / 638
Test (surprise-FR)	—	—	30 / 480	30 / 480

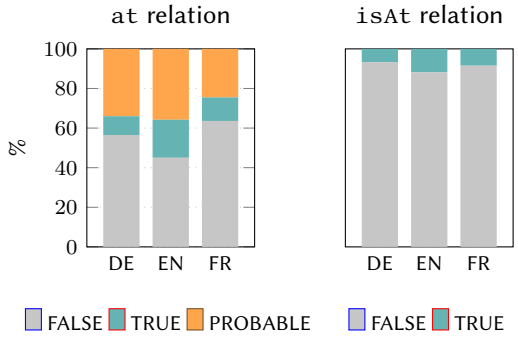


Figure 1: FALSE dominates both relations in all three languages; `isAt=TRUE` is the rarest label (<12%), motivating the use of macro-recall over accuracy.

4.1.1. Label Distribution and Class Imbalance

The label distribution is strongly skewed toward FALSE in both training and development data. `isAt=TRUE` is rarest of all, since it requires evidence that the person was at the location around the time of publication, rather than any past mention. This imbalance means a system can achieve GMR of 0.4167 without learning anything meaningful; improving substantially beyond it requires correctly recalling the minority classes. Figure 1 shows the label distributions in the development set.

4.1.2. Data Characteristics

The HIPE 2026 corpus is distinct from contemporary RE benchmarks in four ways: (i) OCR noise from microfilm digitisation; (ii) archaic 19th–20th century language in German and French; (iii) long documents with many entity pairs requiring per-pair scoping; and (iv) implicit or indirect evidence for positive relation instances, requiring multi-step inference rather than surface matching.

4.2. Methodology

4.2.1. Our Approach

We address the relation qualification problem of Definition 1 as a **per-pair text classification** task via zero/few-shot prompting of large language models. Each pair $(e_p, e_l) \in \mathcal{P}(d)$ is classified independently: the model receives the full article d and the two entity mention strings, and produces $(\hat{y}_{\texttt{at}}, \hat{y}_{\texttt{isAt}})$ subject to the consistency constraint (Eq. 2).

A fundamental design commitment is **multilingual generality**: rather than designing separate systems for each language, we develop a single unified prompting framework and test it across all three document languages in both English and native-language prompt variants. The motivation is threefold. First, the training set is small relative to the task complexity, making fine-tuning risky, especially for German and English (fewer than 100 articles each). Second, modern instruction-tuned LLMs have broad multilingual coverage and capture fine-grained semantic distinctions relevant to historical text. Third, zero/few-shot prompting allows rapid iteration with interpretable, evidence-grounded explanations.

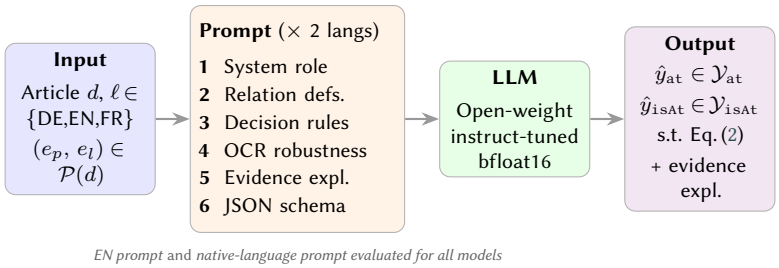


Figure 2: Per-pair inference pipeline. Each article–pair input is processed by a six-component structured prompt (1 system role, 2 relation definitions, 3 decision rules, 4 OCR-robustness instructions, 5 evidence-explanation format, 6 JSON schema), evaluated in both English and native-language variants. The LLM returns a JSON object containing evidence-grounded explanations and a label pair $(\hat{y}_{\text{at}}, \hat{y}_{\text{isAt}})$ satisfying the consistency constraint (Eq. 2).

Figure 2 gives an overview of the full inference pipeline; the six components are detailed in the prompt design subsection below.

4.2.2. Prompt Design

Systematic prompt design is the experimental vehicle for our central contribution: a controlled empirical study of what drives performance in cross-lingual, OCR-degraded historical RE. Each prompt is designed to directly counter the three challenges identified in Section 2: cross-lingual degradation, OCR noise, and semantic subtlety. A key design decision is to build a *single* prompt framework for all three document languages without any language-specific adaptation, making multilingualism a first-class design principle rather than an afterthought. We evaluate prompts in two language variants, English (applied to all document languages) and native (German or French, applied to their respective corpora), making the prompt language itself an explicit experimental variable and directly testing whether high-resource English instruction-following or source-language surface-form matching drives better performance.

Each prompt has six carefully designed components (the complete English template is given in Appendix B):

1. **System role**: The model is instructed to act as a historian working on person–location relations in multilingual historical European newspapers, grounding its reasoning strictly in the article text.
2. **Relation definitions**: Explicit definitions of `at` and `isAt`, their temporal scope, valid label sets, and the distinction between TRUE and PROBABLE for `at`. The prompt operationalises `isAt` as whether the person was at the location in a current, ongoing, or very recent frame relative to the publication date, consistent with the task's temporal-proximity criterion; Section 6.3 (E2) discusses cases where the model over-applies this criterion to single completed events rather than an ongoing presence.
3. **Decision guidance**: Labelling rules specifying when each label applies: PROBABLE only when evidence is indirect or hedged; `isAt` is never PROBABLE; the logical constraints between the two relations must be respected.
4. **Robustness rules**: Explicit instructions to handle OCR noise, line-break hyphenation artefacts, and historical spelling variation, without using external knowledge or resolving ambiguities beyond what the article provides.
5. **Evidence-grounded explanation format**: Before assigning each label, the model produces a brief evidence-only explanation (at most 100 words, citing only article text and explicitly excluding intermediate reasoning steps), grounding the decision in an interpretable, auditable justification.
6. **Structured JSON output**: The final answer is a JSON object with keys `at_explanation`, `at`, `isAt_explanation`, `isAt`, separating the reasoning trace from the label and enabling reliable

Table 2
Open-weight instruction-tuned models evaluated in this study.

Model key	HuggingFace ID	Params	Architecture
gemma4	google/gemma-4-26B-A4B-it	26B (4B active)	MoE
gemma3	google/gemma-3-27b-it	27B	Dense
mistral	mistralai/Mistral-Small-24B-Instruct-2501	24B	Dense
aya	CohereForAI/aya-expanse-32b	32B	Dense

parsing.

The OCR-robustness component in particular mirrors adversarial-training intuitions: rather than exposing the model to perturbed inputs during training to encourage noise-invariant representations [4], robustness is achieved through explicit instruction, directing the model to reason through surface-level corruption rather than treating it as meaningful signal.

Native-language prompt variants (German, French) translate all prose components while keeping the label strings (TRUE, PROBABLE, FALSE) in English, since the evaluation expects English labels. We hypothesised that native-language prompts would benefit models with weaker cross-lingual transfer, while English prompts would generalise more consistently for models whose instruction-following is predominantly English-trained.

4.2.3. Few-Shot Examples

We experimented with 0, 3, 5, and 10 in-context examples per prompt. Examples were drawn from the training set and selected to cover a balanced mix of TRUE, PROBABLE, and FALSE label combinations, avoiding majority-class bias. Each example includes the full article snippet, entity mentions, and the complete JSON output with explanations. Although 10-shot achieved the best development scores for Gemma-4 in post-hoc analysis, the submitted runs used 3-shot for German and French and 5-shot for English, which were the best configurations confirmed prior to the submission deadline.

4.2.4. Models

We evaluated four open-weight instruction-tuned models, all loaded via HuggingFace Transformers in bfloat16 precision on NVIDIA A100-SXM4-80GB GPUs (full hardware and hyperparameter details in Appendix C). Table 2 summarises the models.

Gemma-4 [14] uses a Mixture-of-Experts (MoE) architecture, activating only 4B parameters per forward pass despite 26B total parameters, making it computationally efficient for long-context processing. Gemma-3 is a 27B dense instruction-tuned model with strong multilingual coverage [15]. Mistral-Small-24B is included as a dense, non-MoE point of comparison to Gemma-4 at a similar total-parameter scale, and as a well-established open-weight instruction-tuned model with demonstrated multilingual and European-language competence, letting us probe, without full experimental control, whether MoE sparsity is associated with the few-shot scaling behaviour analysed in Section 6, though model family and scale remain partially confounded with architecture in this comparison. Aya-Expanse is purpose-built for multilingual instruction following, covering over 23 languages [16]; it was evaluated on the development set only (see Table 3).

4.3. Evaluation Setup

For each combination of model $\times$ prompt language $\times$ shot count, we ran inference on the full development set (2,081 pairs across three languages) and computed GMR using the official HIPE 2026 evaluation script. We selected the best configuration per model per language on development GMR. We submitted under team name **BIU_NLP**; the top three models were submitted as run1, run2, and run3, yielding 12 submission files (3 runs $\times$ 4 test files: impresso-DE, impresso-EN, impresso-FR, surprise-FR). For the

Table 3
Best development set results (global macro-recall); all-FALSE baseline ≈0.4167. Aya-Expanse was evaluated on the development set only; time constraints prevented test-set submission.

Model	DE	EN	FR	Best config
gemma-4-26B (MoE)	**0.7518**	**0.7803**	**0.7481**	native/10-shot (DE); en/10-shot (EN, FR)
aya-expanse-32b	0.5950	0.5967	0.6183	en / 0-shot
Mistral-Small-24B	0.5917	0.5935	0.5914	native / 0-shot
gemma-3-27b	0.5622	0.7209	0.6611	native/0-shot (DE); en/0-shot (EN, FR)

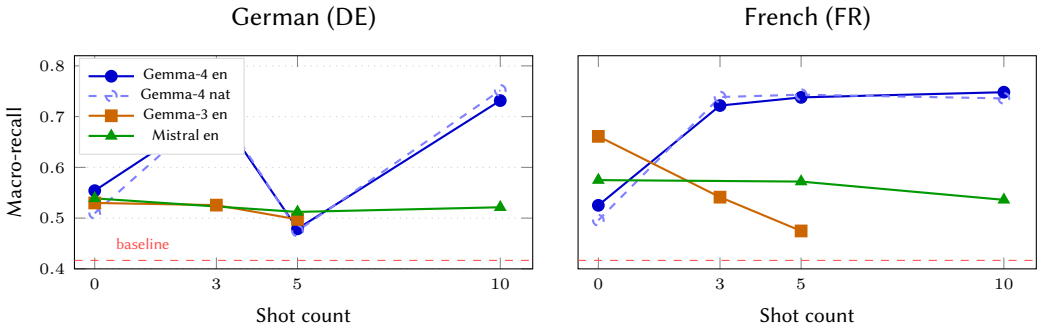


Figure 3: Development GMR vs. few-shot count for German and French; solid = English prompt, dashed = native; red dashed = all-FALSE baseline (0.4167).

surprise French test set, we used the same configuration as for impresso-FR, since the surprise corpus was not disclosed before submission.

5. Results

5.1. Development Set Results

Table 3 reports the best development set GMR per model across all configurations. Gemma-4 is the strongest model on the development set, reaching 0.7518, 0.7803, and 0.7481 for German, English, and French respectively, gains of over 0.33 above the all-FALSE baseline, with native-language 10-shot prompts performing best for German and English-language 10-shot prompts for English and French. All models substantially exceed the all-FALSE baseline (≈0.4167).

Increasing in-context examples consistently improved Gemma-4 up to 10 shots (e.g. +0.18 GMR for German from 0-shot to 10-shot with English prompts: 0.5541→0.7316). For Gemma-3 and Mistral, few-shot examples were inconsistent and sometimes degraded performance, possibly due to prompt length effects or reduced attention to the target article as context grows. Figure 3 plots GMR against shot count for German and French, illustrating these contrasting behaviours.

5.2. Official Test Results

Tables 4 and 5 report the official HIPE 2026 test results. Gemma-3 (run2) achieves rank 20/46 overall, rank 9/46 for French (main), and rank 19/47 for English. Mistral (run3) performs best on German (rank 20/46 main). Gemma-4 (run1) ranks near the bottom on the impresso test sets (Accuracy profile) but reverses on the surprise French test (Generalization profile): rank 6/46 (main) and rank 3/44 (binary) with a binary score of 0.8804. All three models exceed the all-FALSE baseline in every language and evaluation mode.

Figure 4 visualises the ranks across all languages and test sets. Our single best result is Gemma-4 on the surprise-FR binary evaluation (Generalization profile): **rank 3/44**, score **0.8804**, a +0.46 gain over the all-FALSE baseline, demonstrating strong out-of-domain generalisation to a corpus unseen at

Table 4
Main evaluation (global macro-recall): rank/total — score. DE/EN/FR/Overall rows report the impresso test sets (Accuracy profile); Surprise-FR reports the out-of-domain surprise-FR test set (Generalization profile).

Language / Set	run1 (gemma4)	run2 (gemma3)	run3 (mistral)
Overall (mean DE/EN/FR)	41/46 — 0.4429	**20/46 — 0.5781**	24/46 — 0.5390
German	37/46 — 0.4495	31/46 — 0.5050	**20/46 — 0.5451**
English	36/47 — 0.4583	**19/47 — 0.5762**	28/47 — 0.5101
French	42/46 — 0.4208	**9/46 — 0.6529**	23/46 — 0.5617
Surprise-FR	**6/46 — 0.6837**	22/46 — 0.5265	17/46 — 0.6000

Table 5
Binary evaluation (PROBABLE "at" → TRUE): rank/total — score. DE/EN/FR/Overall rows report the impresso test sets (Accuracy profile); Surprise-FR reports the out-of-domain surprise-FR test set (Generalization profile).

Language / Set	run1 (gemma4)	run2 (gemma3)	run3 (mistral)
Overall (mean DE/EN/FR)	41/46 — 0.5252	**20/46 — 0.6721**	24/46 — 0.6274
German	40/46 — 0.5329	24/46 — 0.6231	**23/46 — 0.6381**
English	36/47 — 0.5382	**21/47 — 0.6524**	28/47 — 0.5904
French	42/46 — 0.5044	**14/46 — 0.7409**	23/46 — 0.6538
Surprise-FR	**3/44 — 0.8804**	26/44 — 0.6473	20/44 — 0.7262

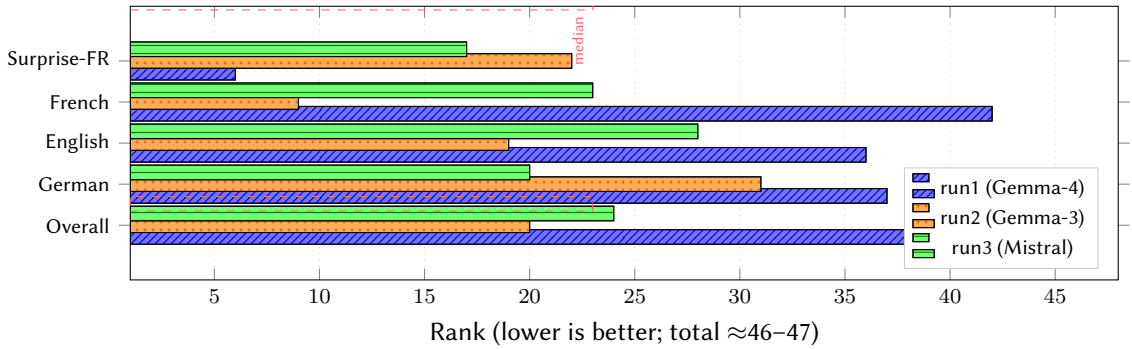


Figure 4: Official main-evaluation ranks per language/test set and run; lower is better; red dashed = rank 23 (approximate median).

development time. Complete rankings for all participating teams and runs are available in the official overview papers [2, 3].

6. Analysis

6.1. Gemma-3 as the Most Reliable System

Gemma-3 (run2) is our strongest overall submission on the impresso test sets, ranking 20/46 and achieving rank 9/46 for French (main, Table 4). Its consistent zero-shot English performance across all three document languages, without any native-language adaptation, makes it a robust baseline for multilingual historical RE. The particularly strong French result (GMR 0.6529) may partly reflect the larger French training set (317 articles, 4,450 pairs; Table 1), providing richer in-domain calibration data for development-time evaluation and example selection.

6.2. The Gemma-4 Anomaly

The most striking finding is the divergence between Gemma-4's development and impresso test performance. On the development set it was the clear winner (Table 3: 0.75–0.78 across all three languages),

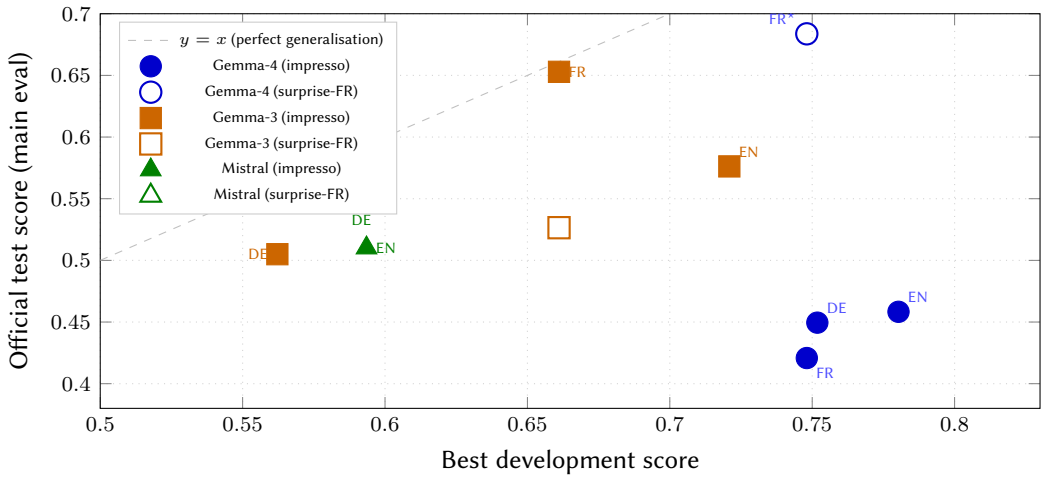


Figure 5: Development vs. official test score (main eval); filled = impresso, hollow = surprise-FR (⋆); dashed diagonal = perfect transfer.

Table 6
Error distribution on the development set (Gemma-3, 0-shot).

Type	Description	German (n = 205)	French (n = 806)
E1	`at` PROBABLE/TRUE over-pred.	—	385 (48%)
E2	`isAt` over-prediction	—†	187 (23%)
E3	Origin-location confusion	systematic (qual.)	
E4	OCR parse failures & cross-entity bleed	53 (26%)	—‡
E1+E2 combined		—	572 (71%)
Other / mixed		152*	234 (29%)

* Non-trivial German errors (JSON-parse failures excluded).

† E2 recurs in German but was not separately counted.

‡ Cross-entity bleed also observed qualitatively in French.

yet on the impresso test sets it ranked last among our runs (Table 4: 41/46 overall). Conversely, on the surprise French test it ranked 6/46 (main) and 3/44 (binary, score 0.8804), our best single result. Several hypotheses may explain this pattern: (i) **dev/test domain shift within impresso**, Gemma-4 may have overfit to dev-set idiosyncrasies during configuration selection; (ii) PROBABLE **over-prediction**, Gemma-4 predicts PROBABLE more aggressively, hurting the main evaluation when the test distribution has fewer ambiguous cases (Table 5 shows a smaller main-to-binary improvement for run1 vs. run2, consistent with this); (iii) **few-shot example bias**, the 3/5-shot examples may not generalise to the impresso test domain; (iv) **surprise-set affinity**, the literary works corpus may be stylistically closer to Gemma-4's pre-training distribution.

Figure 5 makes the dev/test divergence concrete.

6.3. Qualitative Error Analysis

To validate the hypotheses raised in the preceding subsections, we manually inspected development-set errors for the best-submitted model per language (Gemma-3, native prompt, 0-shot for German; Gemma-3, English prompt, 0-shot for French and English). We categorise the 152 non-trivial German errors (predictions with explanations, excluding 53 JSON-parse failures; 205 total) and 806 French errors into four recurring types, summarised in Table 6.

(E1) PROBABLE/TRUE over-prediction. The most frequent error in French (385 false-positive TRUE predictions) is the model assigning `at=TRUE` where the gold label is PROBABLE. For example, *"îsabél . . . est vue au Musée du Louvre, le guide à la main"* ("Isabel . . . is seen at the Louvre Museum, guidebook in hand") elicits a TRUE prediction, but the annotators coded this as PROBABLE because the evidence is from a biographical sketch rather than a direct contemporaneous report. The model treats clear textual evidence of presence as sufficient for TRUE regardless of whether that evidence is deictic or habitual.

(E2) isAt over-prediction. A related pattern, accounting for 187 isAt errors in the French dev set and recurring in German as well, occurs when the model assigns `isAt=TRUE` based on a present-tense or recent-arrival mention, even when the article gives no indication that the person remained at, or was otherwise settled in, the location around publication time. In one German example, *"Hr. James Fazy sei der Konsulate wegen nach Paris gekommen"* ("Mr. James Fazy has come to Paris because of the consulate") triggers `isAt=TRUE`, while the gold label is `at=PROBABLE` / `isAt=FALSE`, because the sentence reports a single completed arrival rather than an established presence at the location around publication time. This suggests the model conflates a recent or newsworthy arrival event with a genuinely current or ongoing state, treating any temporally proximate action as sufficient evidence for `isAt=TRUE`. This over-generalisation systematically inflates false-positive rates for this label.

(E3) Origin-location confusion. Constructions like *"Escher-Bodmer von Zürich"* ("of Zürich") or *"Herren Georg Dörtenbach . . . von Zürich"* are interpreted by the model as evidence of presence in Zürich, producing a PROBABLE prediction where gold is FALSE. The prepositional marker *von* in 19th-century German denotes origin or affiliation, not current location, and the archaic register makes this disambiguation error systematic across models.

(E4) OCR-induced parse failures and cross-entity bleed. A substantial fraction of German errors (53 of 205 total) are JSON-parse failures caused by OCR noise in long articles: hyphenated names (*"Ber-lin"*, *"Ober-hausen"*) and corrupted entity strings trigger malformed model outputs that fall back to FALSE. A separate pattern appears when articles describe multiple persons: the model attributes a location mentioned for person A (e.g. Isabel Hutchinson at the Louvre) to a co-mentioned person B (Marianne Beaugrand), producing a false-positive for the wrong entity.
These findings directly inform our future directions: E1/E2 motivate tighter calibration between PROBABLE and TRUE in the prompt decision guidance; E3 motivates adding explicit *von*-interpretation rules to the German prompt; E4 motivates constrained-decoding JSON validation and per-entity scoping instructions.

6.4. Prompt Language Effects

Native-language prompting consistently benefits models on German: Mistral improves by +0.053 (0.5917 vs. 0.5389, 0-shot), Gemma-3 by +0.032 (0.5622 vs. 0.5299, 0-shot), and Gemma-4 by +0.020 at 10 shots (0.7518 native vs. 0.7316 English). For French, native prompts provided no consistent benefit across models; English prompts matched or exceeded them in most configurations. The pattern suggests that native-language prompting matters most when the document and instruction language would otherwise be mismatched, particularly for historical German, where the gap between 19th-century text and English training distributions is largest. German's rich inflectional morphology — combined with Gothic-typeface OCR artefacts that preferentially corrupt morphological endings, the very markers that disambiguate case and grammatical function — creates a surface distribution that diverges sharply from both modern German and English text; a native-language prompt reduces this instruction–document mismatch by sharing at least vocabulary and morphological patterns with the target. French, typologically closer to English and lacking Gothic-typeface artefacts in the Impresso corpus, presents a smaller distributional gap, which helps explain why English prompts matched or exceeded native French prompts across most configurations. This is consistent with cross-lingual

transfer findings that surface-form alignment between instruction and document language benefits morphologically complex, OCR-affected settings, while a high-resource pivot such as English serves as a more neutral instruction medium when the language gap is smaller [4, 17]. Empirical studies confirm that multilingual LLMs often perform better when prompted in English regardless of document language, though the degree of this English advantage varies with linguistic distance and the model's multilingual training [17, 18].

6.5. Few-Shot Scaling and Progress Beyond Baseline

As shown in Table 7 (see Appendix A), few-shot examples improved Gemma-4 consistently up to 10 shots but were inconsistent for Gemma-3 and Mistral. This may reflect Gemma-4's MoE architecture: activating specialised parameter subsets per token may enable better integration of long in-context demonstrations, while denser models lose focus on the target pair as context grows [19, 20]. We emphasise that this explanation rests on a single MoE model compared against three dense models of different families and scales, and should be treated as a hypothesis rather than an isolated causal effect. All submitted runs exceed the all-FALSE baseline in every setting: run2 achieves GMR 0.5781 overall (main) and 0.6721 (binary), gains of $+0.16$ and $+0.26$; the French result (0.6529, rank 9/46) represents a gain of $+0.24$; and the surprise-FR binary result (0.8804, rank 3/44) a gain of $+0.46$. These gains demonstrate that structured LLM prompting is a strong baseline for this task without task-specific fine-tuning; our framework incorporated evidence-grounded explanations and explicit relation definitions, though no ablation isolates the specific contribution of either component.

7. Related Work

7.1. Relation Extraction

Early RE systems relied on pattern-based rules, dependency tree kernels, and feature-engineered classifiers operating on sentence-level entity pairs with a small closed set of relation types [21]; for a comprehensive overview of methods and benchmarks, see Zhao et al. [22]. Neural approaches [23] improved over these baselines on benchmarks such as ACE [24] and TACRED [25], and the introduction of pre-trained transformers (BERT [26] and its variants) enabled fine-tuned systems to reach near-human performance on standard English benchmarks. End-to-end generative approaches such as REBEL [27] further extended this paradigm by directly outputting relation triplets as text sequences, demonstrating strong adaptability across relation types; end-to-end information extraction architectures have similarly been applied in specialised domains such as biomedicine, where named entity recognition and relation classification are jointly modelled [28]. Document-level RE, which aggregates evidence across multiple sentences, has attracted increasing attention through datasets such as DocRED [29] and methods with adaptive context pooling [30]; this is the natural setting for historical newspaper text, where person–place evidence may span several paragraphs. Cross-lingual RE has been explored via multilingual pre-trained models (XLM-R [31]) and translation-based pipelines, but both require parallel data or target-language annotations that are rarely available for historical domains. Cross-lingual performance varies substantially by language family: fine-tuned models consistently outperform zero-shot cross-lingual transfer for well-represented languages such as German and French, while the gap widens for morphologically complex or typologically distant languages [6, 32].

7.2. LLM-Based Relation Extraction

Large language models have demonstrated strong in-context learning across NLP tasks including RE [33, 34, 35]. Zero-shot and few-shot prompting can match or exceed fine-tuned baselines on standard benchmarks when the model is given explicit relation definitions and label constraints; label verbalization and entailment-based approaches enable effective zero-shot RE without any training on the target relation classes [36], while document-level RE with LLMs has been explored through

autoregressive formulations [37]. Prompting-based approaches have emerged as a primary methodology for multilingual RE, achieving competitive cross-lingual performance without task-specific fine-tuning [4]; competitive zero-shot results have been demonstrated in multilingual relation classification using only structured output formats and English class labels [8]. Even small amounts of multilingual instruction-tuning substantially improve cross-lingual instruction-following, suggesting that instruction-following capabilities transfer efficiently across languages [38]. Chain-of-thought prompting, which elicits an extended intermediate reasoning trace before the final answer, has been shown to improve performance on complex relation semantics in prior work [34]. Our framework instead requires a concise, evidence-only explanation (at most 100 words, explicitly excluding reasoning steps) before each label, closer to a structured evidence justification than full chain-of-thought reasoning; we do not present an ablation isolating its contribution. Even adding a small number of in-context examples in the target language on top of English instruction can substantially improve over purely zero-shot cross-lingual baselines [6], motivating the few-shot prompting regime we explore. Structured output formats such as JSON, combined with intermediate reasoning or evidence-justification steps [39], improve both the reliability and interpretability of predictions, especially when the task requires multi-step inference over indirect evidence. The application of LLMs to RE has been studied primarily in English general-domain settings; work on multilingual and historical settings is limited [40, 41, 42, 43], and our study contributes directly to this gap. Multilingual text classification methods, including feature-based approaches applied to low-resource shared tasks such as cross-lingual sexism detection, provide complementary evidence that careful feature engineering and cross-lingual transfer strategies can generalise across tasks and language families [44, 45].

7.3. Historical Text NLP

Historical newspaper NLP has been driven by the HIPE shared tasks (2020–2026), which provide multilingual benchmarks for named entity recognition and linking on digitised newspaper archives [11, 12]. OCR noise and archaic language have been studied primarily in the NER setting; domain-adaptive language models trained on historical corpora improve NER performance [46] but require large amounts of in-domain text that is impractical to collect for every language and period. RE in historical settings is considerably less explored than NER, and zero-shot cross-lingual RE on historical text without any task-specific training data, the HIPE 2026 setting, has to our knowledge not been studied systematically before this work. More broadly, NLP and text mining pipelines have been applied to diverse digitised text sources, including online health communities and social networks during public health events [47, 48], highlighting the shared challenge of adapting general-purpose models to domain-specific language distributions and user-generated noise.

7.4. Future Directions

Fine-tuning a smaller model (7B–13B parameters) on the HIPE 2026 training data could match zero-shot 24–27B performance at significantly lower inference cost; continued pre-training on historical newspaper text may further improve recall of minority classes. Retrieval-augmented few-shot selection, choosing in-context examples by semantic similarity to the target article, could address the inconsistency of fixed examples observed in Gemma-3 and Mistral (Section 6). The logical constraints between `at` and `isAt` (Definition 1) could be enforced via constrained decoding (e.g. using the Outlines library) or a post-processing repair step rather than relying on prompt instructions alone; quantifying constraint-violation rates in our current outputs would establish the residual error from this source, and an ablation comparing prompt-only versus hard-constraint enforcement is a natural next step. Finally, the French training set is $6\times$ larger than German and $9\times$ larger than English; systematic cross-lingual transfer experiments from French could reduce the data imbalance and improve minority-class recall in German and English. Joint entity-relation extraction approaches [49], which reduce cascading errors from pipeline architectures, are especially relevant in noisy domains where entity boundary errors propagate into relation classification [4]; a prompting-based analogue that simultaneously scopes entity mentions

and classifies their relation could address this limitation in future work. Synthetic data generation and language-family-based transfer have shown consistent gains in low-resource multilingual RE [4]; applying these strategies to the German and English subsets of HIPE, where training data is smallest, is a natural direction for improving minority-class recall.

8. Conclusions

We have shown that a carefully engineered multilingual prompting scheme, evidence-grounded explanations, structured JSON output, explicit OCR-robustness rules, and a unified cross-lingual design, achieves strong performance on French historical text and strong out-of-domain generalisation on the surprise set, with more modest aggregate results, all without task-specific fine-tuning. Our best submissions rank 9th for French (main), 3rd on the out-of-domain surprise set (binary), and 20th overall out of 46–47 submissions (17 participating teams). We note, however, that the strongest development-set model (Gemma-4) generalised poorly to the main impresso test sets (Section 6), underscoring that these results reflect specific prompt/model/configuration combinations rather than a uniformly reliable methodology. The surprise-FR binary result (0.8804 macro-recall, Generalization profile) is a promising signal of out-of-domain generalisation, though given Gemma-4's inconsistent dev-to-test transfer on the impresso sets (Accuracy profile), we treat this as a single encouraging data point rather than evidence that the approach generalises broadly across domains. A key lesson is that development-set rankings do not reliably predict test performance, underscoring the importance of held-out out-of-domain evaluation. We release all code, prompts, and results as a reproducible baseline for the community at https://github.com/Ronke21/HIPE-2026-Evaluating-Person-Place-Relation-Extraction. Reproducibility remains a broader challenge in multilingual RE research, where many systems lack public code or complete implementation details [4]; our full prompt templates, result tables, and configurations are intended to enable direct comparison with future work.

Acknowledgments

This research was supported by a grant from the Israeli Innovation Authority (KAMIN), by a grant from the Israel Science Foundation (ISF), grant number 670/23, and by a generous VATAT grant to the BIU NLP team, which contributed computational resources to this project. We thank the HIPE 2026 organizers for their support and for providing the dataset and evaluation infrastructure. We also thank the anonymous reviewers for their constructive feedback.

Declaration on Generative AI

During the preparation of this work, the author(s) used Claude Code to assist with coding and implementation. Gemini was used for rephrasing and editorial assistance.

Table 7
Complete development set results (**Glb** = global macro-recall; `at` = `at` macro-recall; `isAt` = `isAt` macro-recall; best per model×language in **bold**; all-FALSE baseline Glb≈0.4167).

			German (DE)			English (EN)			French (FR)		
Model	**Prompt**	**Shots**	Glb	at	isAt	Glb	at	isAt	Glb	at	isAt
Gemma-3-27b	en	0	0.5299	0.3927	0.6671	**0.7209**	0.5890	0.8528	**0.6611**	0.5147	0.8074
	en	3	0.5256	0.3668	0.6843	0.6662	0.4428	0.8895	0.5413	0.3952	0.6875
	en	5	0.4975	0.3919	0.6032	0.6881	0.4956	0.8805	0.4747	0.3551	0.5944
	native	0	**0.5622**	0.4919	0.6326	**0.7209**	0.5890	0.8528	0.6528	0.5280	0.7777
	native	3	0.5393	0.4314	0.6472	0.6662	0.4428	0.8895	0.4972	0.3557	0.6386
	native	5	0.5163	0.4281	0.6046	0.6881	0.4956	0.8805	0.4520	0.3511	0.5528
Gemma-4-26B (MoE)	en	0	0.5541	0.4322	0.6760	0.7347	0.6558	0.8137	0.5252	0.4327	0.6178
	en	3	0.7119	0.6238	0.8000	0.7349	0.6411	0.8287	0.7218	0.6567	0.7869
	en	5	0.4789	0.3925	0.5652	0.7550	0.6534	0.8565	0.7379	0.6612	0.8145
	en	10	0.7316	0.6495	0.8136	**0.7803**	0.6854	0.8753	**0.7481**	0.6725	0.8236
	native	0	0.5096	0.4060	0.6132	—	—	—	0.4957	0.4044	0.5870
	native	3	0.7229	0.6578	0.7879	—	—	—	0.7386	0.6821	0.7950
	native	5	0.4748	0.3844	0.5652	—	—	—	0.7432	0.6757	0.8108
	native	10	**0.7518**	0.6529	0.8508	—	—	—	0.7359	0.6714	0.8004
Mistral-Small-24B	en	0	0.5389	0.5002	0.5775	0.5741	0.5168	0.6314	0.5749	0.5205	0.6293
	en	5	0.5122	0.4505	0.5738	0.5653	0.4715	0.6591	0.5720	0.4969	0.6472
	en	10	0.5214	0.4396	0.6033	0.5496	0.4604	0.6389	0.5360	0.4528	0.6193
	native	0	**0.5917**	0.5875	0.5959	**0.5935**	0.5279	0.6591	**0.5914**	0.5677	0.6151
	native	5	0.5595	0.5280	0.5909	0.5773	0.4715	0.6832	0.5776	0.5456	0.6097
	native	10	0.5720	0.5038	0.6402	0.5586	0.4506	0.6667	0.5356	0.4861	0.5851
Aya-Expanse -32b	en	0	**0.5950**	0.4862	0.7039	**0.5967**	0.4045	0.7888	**0.6183**	0.4591	0.7775
	en	3	0.5680	0.3867	0.7493	0.5823	0.3922	0.7723	0.4260	0.3407	0.5113
All-FALSE baseline			0.4167	0.3333	0.5000	0.4167	0.3333	0.5000	0.4167	0.3333	0.5000

Appendix A. Complete Experimental Results

This appendix reports the complete numerical results for all experiments. Table 7 covers every combination of model, prompt language, and shot count evaluated on the development set. Tables 8 and 9 give the full per-metric, per-language official test results.

Appendix A.1. Development Set — Full Evaluation Grid

Appendix A.2. Official Test Results

The official results were drawn from the HIPE 2026 evaluation repository (final de-anonymised release, 2026-05-20). Ranks are given as rank / total submissions in that language–evaluation category. The surprise French test set contains only `at` annotations; `isAt` columns are therefore not applicable (—) for those rows. Submitted configurations (prompt language / shot count) are listed in the **Config** column.

Appendix B. English Prompt Template

The following is the complete English-language prompt template used for all three document languages in the English-prompt condition. The placeholders `{publication_date}`, `{article_text}`, and `{pair_context}` are filled at inference time with the article metadata, full article text, and the target person–place pair respectively. German and French variants translate all prose components while retaining the label strings (TRUE, PROBABLE, FALSE) in English, since the evaluation expects English labels.

```
You are a historian working on person-location relations in multilingual historical
```

Table 8
Official test results — main evaluation (**Score** = global macro-recall; **at Rec** = `at` macro-recall; **at Acc** = `at` accuracy; **isAt Rec** = `isAt` macro-recall; **isAt Acc** = `isAt` accuracy). The **Test Set** column distinguishes the impresso test sets (DE/EN/FR, Accuracy profile) from the surprise-FR test set (Generalization profile).

Run	Model	Config	Test Set	Rank	Score	at Rec	at Acc	isAt Rec	isAt Acc
run1	Gemma-4-26B (MoE)	en / 3-shot	DE	37/46	0.4495	0.3765	0.6429	0.5224	0.7311
		en / 5-shot	EN	36/47	0.4583	0.3962	0.4444	0.5205	0.6111
		en / 3-shot	FR	42/46	0.4208	0.3387	0.5672	0.5030	0.6597
		en / 3-shot	Surprise-FR	**6/46**	**0.6837**	0.6837	0.8354	—	—
run2	Gemma-3-27b	native / 0-shot	DE	31/46	0.5050	0.4740	0.5882	0.5360	0.7311
		en / 0-shot	EN	**19/47**	**0.5762**	0.4807	0.5309	0.6718	0.7346
		en / 0-shot	FR	**9/46**	**0.6529**	0.5703	0.6387	0.7356	0.7773
		en / 0-shot	Surprise-FR	22/46	0.5265	0.5265	0.4729	—	—
run3	Mistral-Small-24B	native / 0-shot	DE	**20/46**	**0.5451**	0.4452	0.6261	0.6450	0.7899
		native / 0-shot	EN	28/47	0.5101	0.4586	0.5062	0.5615	0.6481
		native / 0-shot	FR	23/46	0.5617	0.5743	0.6891	0.5492	0.6891
		native / 0-shot	Surprise-FR	17/46	0.6000	0.6000	0.6042	—	—

Table 9
Official test results — binary evaluation (PROBABLE "`at`" → TRUE); columns as in Table 8, including the Accuracy/Generalization profile distinction by **Test Set**.

Run	Model	Config	Test Set	Rank	Score	at Rec	at Acc	isAt Rec	isAt Acc
run1	Gemma-4-26B (MoE)	en / 3-shot	DE	40/46	0.5329	0.5435	0.6471	0.5224	0.7311
		en / 5-shot	EN	36/47	0.5382	0.5559	0.4444	0.5205	0.6111
		en / 3-shot	FR	42/46	0.5044	0.5059	0.5672	0.5030	0.6597
		en / 3-shot	Surprise-FR	**3/44**	**0.8804**	0.8804	0.8917	—	—
run2	Gemma-3-27b	native / 0-shot	DE	24/46	0.6231	0.7102	0.7185	0.5360	0.7311
		en / 0-shot	EN	**21/47**	**0.6524**	0.6330	0.6852	0.6718	0.7346
		en / 0-shot	FR	**14/46**	**0.7409**	0.7462	0.7227	0.7356	0.7773
		en / 0-shot	Surprise-FR	26/44	0.6473	0.6473	0.5771	—	—
run3	Mistral-Small-24B	native / 0-shot	DE	**23/46**	**0.6381**	0.6312	0.6807	0.6450	0.7899
		native / 0-shot	EN	28/47	0.5904	0.6193	0.5988	0.5615	0.6481
		native / 0-shot	FR	23/46	0.6538	0.7584	0.7437	0.5492	0.6891
		native / 0-shot	Surprise-FR	20/44	0.7262	0.7262	0.6813	—	—

```
European newspapers.

Read carefully, expect OCR noise, and base every judgment only on the document text and
metadata given here.

Task:
- Classify relation "at" for the person and place. Allowed labels: TRUE, PROBABLE, FALSE
- Classify relation "isAt" for the person and place. Allowed labels: TRUE, FALSE

Relation meanings:
- "at" = the article gives textual evidence that the person was at the location at some
  point before publication.
- "isAt" = the article supports that the person was at the location in the article's
  immediate temporal horizon, meaning current, ongoing, or very recent relative to the
  publication date.

Decision guidance:
- Use TRUE for "at" only when the text gives clear evidence of presence.
- Use PROBABLE for "at" only when the text gives indirect, partial, or weak evidence that
  still points toward presence.
- Use FALSE for "at" when the article does not support presence.
- Use TRUE for "isAt" only when the article clearly places the person there in a current,
  ongoing, or very recent frame.
- Never use PROBABLE for "isAt". It must be TRUE or FALSE.
- If "at" is FALSE, then "isAt" must also be FALSE.
```

```
- If "isAt" is TRUE, then "at" must also be TRUE.

Rules:
- Do not use external knowledge.
- Do not infer more than the text warrants.
- Prefer FALSE when evidence is missing or too uncertain.
- Return JSON only.
- Use the exact person and place strings given for the current pair.
- Be robust to OCR noise and line-break artifacts in the text and entity mentions.
- Treat clearly equivalent mention surfaces as the same entity even if they differ by
  escaped line breaks, hyphenation, or minor OCR spelling noise.
- First write ‘at_explanation‘, then decide ‘at‘.
- Then write ‘isAt_explanation‘, then decide ‘isAt‘.
- Keep each explanation concise, at most 100 words.
- Each explanation must mention only evidence from the article, not your reasoning process.
- The document is in English. Entity mentions may include historical spelling and archaic
  orthographic variants.

Document language: en
Publication date: {publication_date}

Article text:
{article_text}

{pair_context}
```

Appendix C. Experimental Setup and Hyperparameters

All inference was performed on NVIDIA A100-SXM4-80GB GPUs using the HuggingFace Transformers library. Table 10 summarises the hardware and generation hyperparameters applied uniformly across all models and configurations. No training, gradient computation, or weight updates were performed at any stage.

Table 10
Hardware and generation hyperparameters used for all inference runs.

Parameter	Value
GPU	NVIDIA A100-SXM4-80GB
GPU allocation	1 per model (`device_map=auto`)
Numerical precision	bfloat16
Decoding strategy	Greedy (temperature = 0.0)
Maximum new tokens	512
Batch size	1 (OOM-safe retry with split batches)
Framework	HuggingFace Transformers + accelerate 1.11.0
PyTorch version	2.5.1 + CUDA 12.1
Python	3.10
Conda environment	hipe2026
Quantisation	None (full bfloat16)

Models were loaded with `torch_dtype=bfloat16` and `device_map="auto"`, distributing layers across available GPUs as needed. For the 24B–27B models on a single A100-80GB, all weights fit on one GPU with sufficient headroom for KV cache. 4-bit and 8-bit quantisation modes were available as command-line options but were not used for any submitted run.

References

[1] P. Joshi, S. Santy, A. Budhiraja, K. Bali, M. Choudhury, The state and fate of linguistic diversity and inclusion in the NLP world, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 6282–6293.

[2] J. Opitz, C. Raclé, A. Michail, M. Romanello, E. Boros, S. Gabay, M. Ehrmann, 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] J. Opitz, C. Raclé, A. Michail, M. Romanello, M. Ehrmann, 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.
[4] M. Ali, R. Speck, H. M. Zahera, M. Saleem, D. Moussallem, A.-C. N. Ngomo, Multilingual relation extraction-a survey, IEEE Access (2025).
[5] R. Keinan, E. Margalit, D. Bouhnik, Emotional analysis in a morphologically rich language: Enhancing machine learning with psychological feature lexicons, Electronics 14 (2025) 3067.
[6] L. Hennig, P. Thomas, S. Möller, Multitacred: A multilingual version of the tac relation extraction dataset, in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2023, pp. 3785–3801.
[7] P. Pakray, A. Gelbukh, S. Bandyopadhyay, Natural language processing applications for low-resource languages, Natural Language Processing 31 (2025) 183–197.
[8] P.-L. H. Cabot, S. Tedeschi, A.-C. N. Ngomo, R. Navigli, Redfm: a filtered and multilingual relation extraction dataset, in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2023, pp. 4326–4343.
[9] N. Zhang, S. Deng, Z. Sun, G. Wang, X. Chen, W. Zhang, H. Chen, Long-tail relation extraction via knowledge graph embeddings and graph convolution networks, 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), 2019, pp. 3016–3025.
[10] M. Mintz, S. Bills, R. Snow, D. Jurafsky, Distant supervision for relation extraction without labeled data, in: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, 2009, pp. 1003–1011.
[11] M. Ehrmann, M. Romanello, S. Bircher, S. Clematide, Extended overview of CLEF HIPE 2020: Named entity processing on historical newspapers, in: Working Notes of CLEF 2020 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, 2020.
[12] M. Ehrmann, M. Romanello, A. Doucet, S. Clematide, HIPE-2022: Labeling, benchmarking and evaluating named entity recognition and linking in historical newspapers, Journal of Data Mining & Digital Humanities (2022).
[13] D. Vrandečić, M. Krötzsch, Wikidata: a free collaborative knowledgebase, Communications of the ACM 57 (2014) 78–85.
[14] Gemma Team, Google DeepMind, Gemma 4, https://deepmind.google/models/gemma/gemma-4/, 2026. Accessed 2026-07-02.
[15] G. Team, Gemma 3 technical report, 2025. arXiv:2503.19786.
[16] A. Üstün, V. Aryabumi, Z. Yong, W.-Y. Ko, D. D'souza, G. Onilude, N. Bhandari, S. Singh, H.-L. Ooi, A. Kayid, et al., Aya model: An instruction finetuned open-access multilingual language model, in: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2024, pp. 15894–15939.
[17] J. Etxaniz, G. Azkune, A. Soroa, O. L. de Lacalle, M. Artetxe, Do multilingual language models think better in english?, in: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers), 2024, pp. 550–564.
[18] I. Mondshine, T. Paz-Argaman, R. Tsarfaty, Beyond english: The impact of prompt translation strategies across languages and tasks in multilingual llms, arXiv preprint arXiv:2502.09331 (2025).
[19] DeepSeek-AI, et al., Deepseekmoe: Towards ultimate expert specialization in mixture-of-experts

language models, arXiv preprint arXiv:2401.06066 (2024).
[20] S. Wang, Z. Chen, B. Li, K. He, M. Zhang, J. Wang, Scaling laws across model architectures: A comparative analysis of dense and MoE models in large language models, in: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, 2024, pp. 5583–5595.
[21] G. Zhou, J. Su, J. Zhang, M. Zhang, Exploring various knowledge in relation extraction, in: Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05), 2005, pp. 427–434.
[22] X. Zhao, Y. Deng, M. Yang, L. Wang, R. Zhang, H. Cheng, W. Lam, Y. Shen, R. Xu, A comprehensive survey on relation extraction: Recent advances and new frontiers, ACM Computing Surveys 56 (2024) 1–39.
[23] Y. Lin, S. Shen, Z. Liu, H. Luan, M. Sun, Neural relation extraction with selective attention over instances, in: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2016, pp. 2124–2133.
[24] G. R. Doddington, A. Mitchell, M. A. Przybocki, L. A. Ramshaw, S. M. Strassel, R. M. Weischedel, et al., The automatic content extraction (ace) program-tasks, data, and evaluation., in: Proceedings of LREC 2004, volume 2, 2004, pp. 837–840.
[25] G. Stoica, E. A. Platanios, B. Póczos, Re-tacred: Addressing shortcomings of the tacred dataset, in: Proceedings of the 35th AAAI Conference on Artificial Intelligence, 2021, pp. 13843–13850.
[26] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2019).
[27] P.-L. Huguet Cabot, R. Navigli, REBEL: Relation extraction by end-to-end language generation, in: Findings of the Association for Computational Linguistics: EMNLP 2021, 2021, pp. 2370–2381.
[28] R. Keinan, A. D. N. Cohen, R. Tsarfaty, From named entities to relations: End-to-end biomedical information extraction, 2025.
[29] 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, 2019, pp. 764–777.
[30] W. Zhou, K. Huang, T. Ma, J. Huang, Document-level relation extraction with adaptive thresholding and localized context pooling, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 35, 2021, pp. 14612–14620.
[31] 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, 2020, pp. 8440–8451.
[32] M. Artetxe, S. Ruder, D. Yogatama, On the cross-lingual transferability of monolingual representations, arXiv preprint arXiv:1910.11856 (2020).
[33] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, Advances in Neural Information Processing Systems 33 (2020) 1877–1901.
[34] S. Wadhwa, S. Amir, B. C. Wallace, Revisiting relation extraction in the era of large language models, in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2023, pp. 15566–15589.
[35] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, et al., Gpt-4 technical report, arXiv preprint arXiv:2303.08774 (2023).
[36] O. Sainz, O. Lopez de Lacalle, G. Labaka, A. Barrena, E. Agirre, Label verbalization and entailment for effective zero and few-shot relation extraction, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 2021, pp. 1199–1212.
[37] L. Xue, D. Zhang, Y. Dong, J. Tang, Autore: Document-level relation extraction with large language models, in: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations), 2024, pp. 211–220.
[38] U. Shaham, J. Herzig, R. Aharoni, I. Szpektor, R. Tsarfaty, M. Eyal, Multilingual instruction tuning with just a pinch of multilinguality, in: Findings of the Association for Computational Linguistics:

ACL 2024, 2024, pp. 2304–2317.
[39] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, D. Zhou, Chain-of-thought prompting elicits reasoning in large language models, arXiv preprint arXiv:2201.11903 (2022).
[40] D. Jinensibieke, M. Maimaiti, W. Xiao, Y. Zheng, X. Wang, How good are llms at relation extraction under low-resource scenario? comprehensive evaluation, ArXiv abs/2406.11162 (2024).
[41] A. Swarup, T. Pan, R. Wilson, A. Bhandarkar, D. Woodard, LLM4RE: A data-centric feasibility study for relation extraction, in: Proceedings of the 31st International Conference on Computational Linguistics, 2025, pp. 6670–6691.
[42] R. Keinan, Y. HaCohen-Kerner, JCT at SemEval-2023 tasks 12a and 12b: Sentiment analysis for tweets written in low-resource african languages using various machine learning and deep learning methods, resampling, and hyperparameter tuning, in: Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023), 2023, pp. 365–378.
[43] R. Keinan, Text mining at SemEval-2024 task 1: Evaluating semantic textual relatedness in low-resource languages using various embedding methods and machine learning regression models, in: Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), 2024, pp. 420–431.
[44] R. Keinan, Sexism identification in social networks using tf-idf embeddings, preproccessing, feature selection, word/char n-grams and various machine learning models in spanish and english., in: CLEF (Working Notes), 2024, pp. 1058–1069.
[45] R. Keinan, Tackling sexism in social media: Multilingual AI solutions, Journal of Electrical and Electronics Engineering 3 (2024) 01–07.
[46] S. Schweter, L. März, hmBERT: Historical multilingual language models for named entity recognition, in: Working Notes Papers of the CLEF 2021 Evaluation Labs, CEUR Workshop Proceedings, 2021.
[47] R. Keinan, E. Margalit, D. Bouhnik, Analysis of user trends in digital health communities using big data mining, Plos one 19 (2024) e0290803.
[48] R. Keinan, E. Margalit, D. Bouhnik, Impacts of a public health crisis on health-centered online social networks, Informing Science 28 (2025).
[49] X. Li, F. Yin, Z. Sun, X. Li, A. Yuan, D. Chai, M. Zhou, J. Li, Entity-relation extraction as multi-turn question answering, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 1340–1350.