FI-CODE@HIPE-2026: Efficient Person-Place Relation Classification from Distance Heuristics to Prompted Language Models

Lena Griesbeck[1,†], Moritz Hennen[1,†], Florian Babl[1] and Michaela Geierhos[1]

[1]Research Institute CODE, University of the Bundeswehr Munich, Werner-Heisenberg-Weg 39, 85579 Neubiberg, Germany

Abstract

We present our submission to the HIPE-2026 shared task on multilingual person-place relation classification in historical documents, using the FI CODE approach. This task requires the prediction of both a ternary AT relation and a binary ISAT relation for sampled person-location pairs in OCR-processed historical documents written in German, English, and French. Three systems designed for efficiency are compared: a distance heuristic over entity mention positions with no training, an encoder-based ITER relation classifier, and an in-context learning system based on a small large language model. Our experiments demonstrate that, while simple positional cues provide a competitive baseline, encoder- and prompt-LLM-based systems offer additional advantages. In the HIPE-2026 efficiency ranking, we achieved first place for French texts and second place for English texts using distance-based relation classification.

Keywords

Natural Language Processing for Historical Texts, Relation Classification, Efficient NLP, HIPE 2026

1. Introduction

Although historical newspapers contain a lot of information about people, places, and events, extracting this information is difficult because the documents may be written in different languages, the quality of optical character recognition (OCR) varies, and the relationships between entities are often expressed indirectly. The HIPE-2026 shared task[1] [1, 2] addresses these challenges by proposing to classify person-place relationships. Given pairs of person and location mentions, systems must predict both a ternary historical AT relationship and a binary, time-specific ISAT relationship.

We approached the task from an efficiency-oriented perspective. Rather than submitting a single high-capacity model, we compare three systems with different computational profiles: (i) a heuristic for distance-based relation classification (RC) without learned parameters, (ii) an encoder-based relation classifier, ITER, adapted to the task's requirements, and (iii) a small, instruction-based 2B large language model (LLM). This comparison enables us to analyze the extent to which positional cues capture the task, how well compact discriminative encoders perform when candidate entities are provided, and if small LLMs can generate valid multilingual relation predictions through structured prompts.

The remainder of this paper is structured as follows: Section 2 introduces the shared task. Sections 3, 4, and 5 describe the three submitted systems. Section 6 reports the experimental setup, followed by Section 7 with the results. Section 8 contains the conclusion.

2. Task Description

The HIPE-2026 shared task aims to classify whether a person, mentioned at least once in a historical news article, was at a specific geographical location also mentioned in the article, based on contextual

CLEF 2026: Conference and Labs of the Evaluation Forum, September 21–24, 2026, Jena, Germany

[†]These authors contributed equally.

lena.griesbeck@unibw.de (L. Griesbeck); moritz.hennen@unibw.de (M. Hennen); florian.babl@unibw.de (F. Babl); michaela.geierhos@unibw.de (M. Geierhos)

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

information from the article. This classification is performed in terms of two relations: AT and ISAT, where the ISAT relation is a hyponym of the more general AT relation: ISAT constrains the classification of the person having been at a location to the immediate temporal context of the news article. The two relations also differ by their associated label space: while ISAT can solely be classified as y_{isAt} = TRUE or y_{isAt} = FALSE, AT additionally encompasses a y_{at} = PROBABLE clause. The PROBABLE label allows systems to indicate that the classification is based on context-dependent clues, rather than on explicit textual evidence. For instance, this includes a person participating in an event which is attributed to a specific geographic location.

Mathematically, the task can be described as follows. Given a historical article $\mathbf{t} = \langle t_1 \dots t_N \rangle$ as a string of N characters, two sets of co-references for a singular person and location, $P = \{p_1, \dots\}$ and $L = \{l_1, \dots\}$, and two relation types AT and ISAT, the HIPE-2026 shared task was to design a system for the aforementioned classification task, following Equation 1.

$$system : S \times \wp(S) \times \wp(S) \times \{\text{AT}, \text{ISAT}\} \mapsto L \tag{1}$$

$$system(\mathbf{t}, P, L, x) \rightarrow \begin{cases} y \in \{\text{TRUE}, \text{FALSE}, \text{PROBABLE}\} & \text{iff. } x \equiv \text{AT} \\ y \in \{\text{TRUE}, \text{FALSE}\} & \text{otherwise} \end{cases} \tag{2}$$

The shared task includes newspaper data in German, English, and French, following the annotation and evaluation setup defined by the HIPE-2026 organizers [2]. An additional surprise test set in French is released for the final evaluation. Systems are evaluated separately for AT and ISAT using macro-recall, and a global score is calculated across both relations. The evaluation of this shared task is performed using three distinct evaluation profiles: accuracy, efficiency and generalization. While the accuracy profile measures pure system accuracy in terms of macro recall, the efficiency profile ranks systems in three dimensions: macro recall, parameter count, and model size, where the latter two are to be minimized. The score for this profile is the arithmetic mean of the ranks across these three dimensions. The generalization profile aims to compare systems based on their performance on the French literary surprise test set. The efficiency profile in particular encourages the development of resource-efficient, low-parameter, and low-file-size systems. To evaluate the different models, we utilized the official scorer[2].

3. Distance-based Relation Classification

The first attempt to minimize the number of parameters used by a model and maximize the efficiency rating of the evaluated system involved analyzing the distance between PER-LOC entity pairs.

Distances Between PER-LOC Pairs. Suppose the positions of the PER-LOC mention pairs P and L within the text are given by their start and end boundaries $\rho(p_i) = (p_i^{\langle}, p_i^{\rangle})$ and $\rho(l_j) = (l_j^{\langle}, l_j^{\rangle})$. Furthermore, let $\mathbf{t}[n : m]$ be the substring of $\mathbf{t}$ from n to m. Then, the distance Δ between the person and the location mentions P and L is given by the smallest number of whitespace characters in the substring of the historical article $\mathbf{t}$ from p to l (or from l to p vice versa), as defined in Equation 3.

$$\Delta(P, L) = \min_{p_i \in P} \min_{l_j \in L} (\text{whitespaces}(\mathbf{t}[\min(p_i^{\rangle}, l_j^{\rangle}) : \max(p_i^{\langle}, l_j^{\langle})])) \tag{3}$$

$$\text{where whitespaces}(\mathbf{t}) = \sum_{n=1}^{N} \mathbb{1}[t_n \equiv \text{“ ”}] \tag{4}$$

For further analysis, distances $\Delta(P, L)$ are grouped based on the correct label $\hat{y}_x$ for both relation classes x = AT and x = ISAT. Figure 1 visualizes this process and highlights the key finding of this procedure: At distances $\Delta > 80$ between the mentions of all *person-location* position combinations in a text, the likelihood of the true label $\hat{y}_x$ being FALSE strongly increases.

[2] https://github.com/hipe-eval/HIPE-2026-data/releases/tag/v1.0

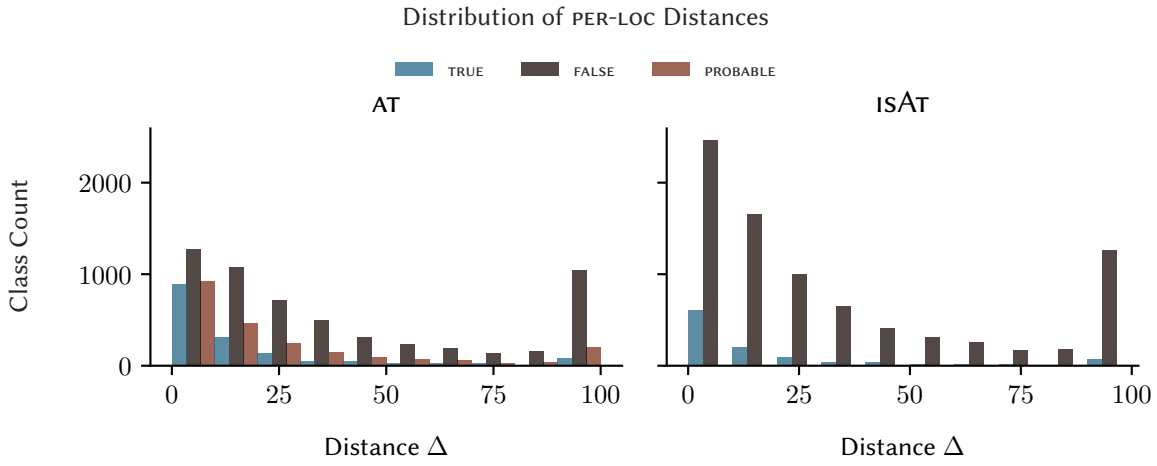

Figure 1: Distribution of distances for both relation types and target labels $\hat{y}_c$, AT and ISAT. Distances are binned into ten bins in total, and distances $\Delta > 100$ are clamped following $\Delta' = \min(\Delta, 100)$. For both relation types, a long tail of $\hat{y}_c = false$ labels can be observed, leading to the conclusion that as the distance between P and L increases, it is increasingly unlikely that there exists a relation between the *person* and the *location*.

A Naive, Distance-Based System. Based on the findings in Figure 1, a naive system is implemented. When the minimal distance between person and place pairs exceeds $\Delta(P, L) > 80$, the likelihood of $\hat{y}_x =$ FALSE tends towards a 100 % ratio. For both relation types, the system predicts the TRUE label if the minimum distance $\Delta(P, L)$ between mention pairs P and L is less than 80, and predicts FALSE otherwise, as defined in Equation 5.

$$\text{system}(P, L, c) = \begin{cases} true & \text{iff. } \Delta(P, L) < 80 \\ false & \text{otherwise} \end{cases} \tag{5}$$

This system has an inherent limitation: it never predicts the label PROBABLE. Although it is very simple in design, our preliminary experiments show that it outperforms the random baseline provided by the HIPE-2026 shared task organizers. Therefore, we chose to include this system in our official submission. Most notably, this approach works without any optimization procedure, making it very efficient in inference and in general since no training is required.

Aside from its simple nature, this system's findings in Figure 1 highlight something about the semantic nature of this relation classification task. Specifically, the distance between entity pairs – or between multiple candidate pairs, as is the case with this shared task – is correlated with the correct label $\hat{y}_x$. To improve predictions from more sophisticated systems, knowledge from this naive, distance-based system can be used to filter out possible misclassifications, especially when $\Delta(P, L) > 80$. Additionally, the threshold might be further optimized since it was solely derived visually from Figure 1.

4. Relation Classification with ITER

ITER [3] is a recent encoder-based model for named entity recognition (NER) and relation extraction (RE). It performs on par with most generative models in this domain. Our choice of this model was primarily due to its ability to leverage small discriminative Transformer models, as model efficiency is another motivating factor for this shared task, aside from pure retrieval accuracy. However, several adjustments were made to its inference process, because it is a traditional RE model that usually requires identifying named entities in an input before classifying their relationships.

ITER Architecture. ITER uses a three-step process to recognize entities and to extract their relations. The input is first passed into the encoder backbone, typically a discriminative BERT-like Transformer model, to retrieve a latent, contextualized representation of the input. Based on the obtained sequence

ITER Input Construction

Original Article:
t = George Washington and John Adams were seen in New York last week. Washington and Adams were in N.Y. for a political meeting and were seen exploring Times Square.
Given Person and Location Pair in t:
P = {George Washington, Washington}
L = {New York, N.Y.}
Shortest Mention:
$\bar{p} = \arg\min_{p_i} |p_i|$ = Washington
$\bar{l} = \arg\min_{l_j} |l_j|$ = N.Y.
Model Input:
$\bar{\mathbf{t}}$ = Washington [HEAD] N.Y. [TAIL] George Washington and John Adams were seen in New York last week. Washington and Adams were in N.Y. for a political meeting and were seen exploring Times Square.

Figure 2: Example for the sample construction of an ITER model input. The shortest mentions of people and places are prepended to the original input article and are separated by special meta tokens.

of latent vectors, the model first marks all positions $\mathbf{t}_n$ as beginning a named entity or not. The corresponding ending anchors for named entities are then identified, based on the knowledge of where their respective start anchors are located, and again based on the latent representation. These two steps identify the named entities in the input. In the third and final phase, the latent representations of start and end anchors are combined to predict whether the corresponding named entity holds any relations with any of the other identified named entities, using pairwise comparison. Assuming the number of named entities is N, $N \times N$ combinations are tested for relations.

Input Augmentation. In this shared task, the named entities have already been identified and supplied to the system as additional information. To facilitate inference for this model, the input **t** is thereby augmented to include the information about the already recognized named entities in P and L. Specifically, the *shortest mention surface form*[3] mentions of P and L, $\bar{p}$ and $\bar{l}$ with $\bar{p} = \arg\min_{p_i} |p_i|$ and $\bar{l} = \arg\min_{l_j} |l_j|$, are *prepended* to the original contents of the historical news article **t**. Figure 2 shows an illustration of this process. Given this input augmentation, the general architecture and inference procedure of ITER do not change in a meaningful way. The augmentation solely simplifies the predictive process by providing hints for the named entities to identify, trivializing the NER task.

Label Space Augmentation. Using this model for the shared task presents another key challenge to consider. Since ITER is a classical relation extraction model, it cannot assign labels to each relation. Instead, it only predicts whether a relation exists between a pair of entities. This is problematic for the AT relation, where the label space also includes the PROBABLE label. To address this issue, the prediction process is augmented. The model is trained to predict three relation types – AT, ISAT and a helper type *AT-PROBABLE*. At inference time, the assigned label for both *core* relations AT and ISAT is FALSE by default. If the model predicts that person P has a relation AT-PROBABLE with location L, the output for the AT relation is changed to PROBABLE. In the other two cases, i.e., a prediction of AT and ISAT, the output of the respective relation is changed to TRUE.

Base Model Selection. ITER requires an encoder model as its backbone; specifically, a Transformer-based model. To identify the most suitable model, we conducted preliminary experiments using only the Silver Train A dataset for the English language. We selected hmBERT [4], the TINY variant of the

[3]The decision to use only the shortest mention surface forms was made to increase the length of the input article by the smallest amount possible.

T5 [5, 6] encoder model, and both DeBERTaV3 [7] models as potential candidates. This selection is based on task harmonization (hmBERT), parameter efficiency (T5), and model recency (DeBERTaV3).

Table 1
Comparison of different backbone models for ITER in terms of macro-recall. The base variant of DeBERTaV3 outperforms its larger counterpart, as well as the very small T5 and hmBERT models. All four models were trained on the English Gold Train A dataset. The Silver Dev A dataset was used for evaluation during training and the final scores reported are based on the English Silver Train A dataset.

Model	`MacroRecall_at`	`MacroRecall_isAt`	**Macro Recall**
hmBERT	0.4748	0.5000	0.4874
T5 TINY	0.5683	0.5000	0.5341
DeBERTaV3 (base)	**0.6012**	0.5000	**0.5506**
DeBERTaV3 (large)	0.5921	0.5000	0.5460

The results of this process are shown in Table 1. In these preliminary tests, the base variant of DeBERTaV3 performed best with respect to the *macro-recall* metric and was therefore used for all subsequent experiments. ITER consistently achieves a `MacroRecall_isAt` score of 0.500 in all settings, which we investigate further and find that it stems from the fact that the model is only predicting FALSE for the ISAT relation. While this is a significant shortcoming of the model, we do not conduct any additional error analysis.

5. In-Context Learning with Large Language Models

In addition to evaluating a distance-based relation classification system and an encoder-based ITER approach, we examined instruction-tuned LLMs in an in-context learning setting. Rather than training a dedicated classifier, we framed the task as multilingual relation classification via prompting. Formally: Given a document t and a pair of entities (p, l), consisting of a person mention p and a location mention l, we instructed the model to predict the relation labels AT and ISAT directly in a structured JSON format.

Model Comparison. We evaluated several small, instruction-tuned language models with 0.8 to 7 billion parameters. The evaluated models include Qwen3.5 0.8B[4] [8], Qwen3.5-2B[5] [8], Mistral 7B Instruct[6] [9], Llama 3.2 1B Instruct[7] [10], and LFM2 1.2B Instruct[8] [11]. The results are summarized in Table 2, which reports only selected configurations from the screening stage in order to keep the comparison focused on model-level differences. The more extensive prompt-level evaluation of the ultimately selected model Qwen3.5-2B is reported separately in Appendix B. This table should be understood as a screening experiment rather than a complete multilingual comparison of all candidate models. We first evaluated the candidate models primarily in English and then added selected German runs to assess whether performance changed substantially with prompts in different languages. French was not included for all models in this exploratory stage because the goal was to identify a compact model-prompt configuration for the final multilingual submission. The final selected configuration was then evaluated for all three languages, with the prompt-level results reported in the appendix.

Generally, smaller models exhibited greater sensitivity to the phrasing of the prompt and lower multilingual robustness. Larger models, on the other hand, achieved better results, albeit at the cost of higher computational overhead and larger checkpoint sizes. Since the joint task includes an efficiency profile, we preferred models with smaller file sizes that demonstrated stable performance when executing commands in all three task languages. For the final LLM run, we chose Qwen3.5-2B because it offered a

[4]https://huggingface.co/Qwen/Qwen3.5-0.8B
[5]https://huggingface.co/Qwen/Qwen3.5-2B
[6]https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3
[7]https://huggingface.co/meta-llama/Llama-3.2-1B
[8]https://huggingface.co/LiquidAI/LFM2.5-1.2B-Instruct

Table 2
Overview of model evaluation results from the screening experiments. The table reports selected model–prompt configurations used to compare compact instruction-tuned models. Detailed prompt-level results for the final Qwen3.5-2B model are reported in Tables 7, 8, and 9.

Model	Lang	Prompt	MacroRecall_at	MacroRecall_isAt	Macro Recall
LLM-Baseline (Mistral3-3B)	en	–	0.4587	0.6564	0.5576
LiquidAI/LFM2-1.2B-Instruct	en	V2	0.3742	0.5073	0.4408
LiquidAI/LFM2-1.2B-Instruct	en	V4	0.3467	0.4842	0.4154
LiquidAI/LFM2-1.2B-Instruct	en	V5	0.3049	0.5096	0.4072
meta-llama/Llama-3.2-1B-Instruct	en	V2	0.3321	0.5054	0.4187
meta-llama/Llama-3.2-1B-Instruct	en	V3	0.3343	0.5098	0.4220
meta-llama/Llama-3.2-1B-Instruct	en	V5	0.3113	0.5098	0.4105
Qwen/Qwen-3.5-0.8B	en	V2	0.3646	0.5460	0.4553
Qwen/Qwen-3.5-0.8B	en	V3	**0.4665**	0.6055	0.5360
Qwen/Qwen-3.5-0.8B	en	V4	0.3538	0.5158	0.4348
Qwen/Qwen-3.5-0.8B	en	V5	0.3266	0.5274	0.4270
Qwen/Qwen-3.5-0.8B	de	V3	0.3714	0.5066	0.4390
Qwen/Qwen-3.5-0.8B	de	V4	0.3490	0.5211	0.4351
Qwen/Qwen-3.5-0.8B	de	V5	0.3506	0.4960	0.4233
Qwen/Qwen3.5-2B	en	V2	0.4320	**0.7090**	**0.5705**
mistralai/Mistral-7B-Instruct-v0.3	en	V1	0.4637	0.5377	0.5007
mistralai/Mistral-7B-Instruct-v0.3	en	V3	0.3798	0.6805	0.5301
mistralai/Mistral-7B-Instruct-v0.3	en	V4	0.2697	0.7061	0.4879
mistralai/Mistral-7B-Instruct-v0.3	en	V5	0.3951	0.6739	0.5345

smaller size and stable inference on a single GPU while delivering competitive multilingual performance. Our final choice prioritized efficiency over the LLM configuration with the highest score.

Prompt Engineering. To avoid additional cross-linguistic effects when following the instructions, the prompt instructions were written in the same language as the source document (i.e., English, German, or French). All prompt variants followed the same inference protocol. For each document, we reviewed all the provided person-location pairs and provided the model with a separate prompt for each pair. Each prompt contained the publication date, the full article text, and mentions of the current person and location. All prompts instructed the model to predict both relations, AT and ISAT, within a specified JSON schema. We have explicitly specified that PROBABLE is a valid label only for AT, while ISAT allows only TRUE and FALSE.

The prompt variants were developed iteratively and differed in terms of their level of detail, the calibration of the PROBABLE label, the use of examples, and the scope of the step-by-step instructions. Table 6 provides an overview of the evaluated prompt versions in Appendix A. The results for all evaluated prompt variants of the model that was ultimately selected are listed in Tables 7, 8, and 9 in Appendix B.

Overall, compact prompts demonstrated more robust performance than longer, more descriptive prompts when used with small, instruction-tailored models. In some cases, few-shot prompts improved recall for PROBABLE predictions but often reduced the accuracy of TRUE. Longer and more guided prompts often caused the model to collapse toward overproduction or suppression of the PROBABLE label, particularly for models with fewer than 1 billion parameters. The final submission uses prompt version V2, as it demonstrated the most stable overall performance in experiments with Qwen3.5-2B for English, German, and French. Among the evaluated alternatives, V7 was the closest competing prompt variant as it slightly outperformed V2 in German. But V2 achieved higher global macro recall for English and French and the better average across all three languages as well as consistently valid,

structured outputs.

Because the small LLMs in the experiment at times produced invalid JSON or added extra text to the target output, we created a parsing and repair pipeline. First, we parsed the generated output directly as JSON. If parsing failed, we tried to reconstruct the JSON object using bracket matching from the generated text. If that failed as well, we used regular expression-based recovery to extract valid relationship labels. Predictions that could not be recovered were assigned fallback labels of FALSE. After parsing, we normalized all labels based on the allowed label sets for the respective relations. Finally, we performed a consistency correction step to prevent invalid outputs where ISAT was predicted as TRUE while AT was predicted as FALSE.

The repair pipeline mainly served as a safety guard and we did not log exact recovery frequencies for the final analysis, therefore we report its impact qualitatively. Incomplete or invalid JSON occured rarely overall, but was observed more often in the smallest models, especially in combination with longer more complex prompts. The issue was most visible for models below 1B parameter size, for larger models and for the final configuration with Qwen3.5-2B and V2, structured-output failures where uncommon.

6. Experimental Setup

This section provides fine-grained information about how we set up our experiments. Additionally, we offer supplementary details on the encoder-based and in-context learning approaches.

Table 3
Overview of the HIPE-2026 shared task data for English, French, and German. A total of 721 samples are provided, containing 1,251 and 8,251 annotations for the ISAT and AT relations for human-annotated and LLM-annotated data, respectively.

Quality	Language	Split	# Samples	# ISAT		# AT			
				# TRUE	# FALSE	# TRUE	# FALSE	# PROB.	Σ
Human-Annotated	English	train	35	83	224	125	152	30	614
	French	train	35	107	371	181	269	28	956
	German	train	34	89	377	135	269	62	932
		Σ =	104	279	972	441	690	120	
					1,251			1,251	
Ensemble Voting	English	train	56	77	419	98	239	159	992
		dev	17	18	133	29	68	54	302
	French	train	317	466	3,984	636	2,727	1,087	8,900
		dev	107	127	1,371	179	952	367	2,996
	German	train	88	134	1,090	188	703	333	2,448
		dev	32	29	403	41	244	147	864
		Σ =	617	851	7,400	1,171	4,933	2,147	
					8,251			8,251	
			721		9,502			9,502	19,004

Official Shared Task Data. For the three languages of the shared task, the organizers of HIPE-2026 provide 104 human-labeled and 617 LLM-labeled historical news articles[9]. An overview of the provided data is shown in Table 3. Final submissions were generated using four unlabeled official test files released by the organizers. The test files encompassed in-domain evaluation data for all three languages – English, French and German – as well as the out-of-domain French literary surprise test set and

[9] https://github.com/hipe-eval/HIPE-2026-data/releases/tag/v1.0

comprised of 19, 19, 19 and 30 samples, respectively. The remainder of this section contains information about the experimental setup for each system introduced individually.

System 1/Distance Based Relation Classification. To calibrate the cutoff distance Δ, all 9,502 PER-LOC pairs from Gold and Silver Train A as well as Silver Dev A in all languages are used (see Table 3).

System 2/ITER. Each per-language ITER model is trained using a single NVIDIA H100 GPU. For the three languages, Gold Train A data is used for training, while Silver Dev A and Silver Train A data are used for validation during training and final post-training validation, respectively. ITER provides fine-grained control over its training hyperparameters. GELU [12] is used as the activation function for the multilayer perceptron (MLP) parameters of the model. The intermediate MLP dimension is 2,048. The MLP parameters and those of the underlying base model are trained using AdamW [13] with learning rates of 1×10^{-5} and 5×10^{-5}, respectively. A weight decay of 0.1 and 0.02 is also applied. Both learning rates decay to 10 % of their original value over the course of model training, using a cosine schedule. Optimization is performed for 80,000 steps, and evaluation is done in 5,000-step intervals. An in-depth hyperparameter search is not conducted, and hyperparameters that incentivize a low training loss are manually selected.

System 3/Prompted LLM. The final prompting run used Qwen3.5-2B with prompt version V2. Various prompts and models were evaluated using human-annotated data for all three languages.

The model is served using vLLM on a single RTX 3090 GPU with `float16` precision and a maximum context length of 4,096 tokens. The tensor parallel size is set to 1, and the GPU memory utilization is set to 0.90. Inference is performed via the OpenAI-compatible chat API. We used deterministic decoding with a temperature of 0.0, `top_p=1.0`, and `top_k=-1`; therefore, neither nucleus sampling nor top-k filtering was applied. The repetition penalty is set to 1.05, and `max_tokens` is set to 256. For models that support chat template options, thinking is disabled via `chat_template_kwargs`.

7. Results

As shown in Table 4, the official HIPE-2026 efficiency rankings reveal a clear trade-off between accuracy and efficiency across all evaluated approaches.

Table 4
Official HIPE-2026 efficiency profile rankings across Newspaper test sets, grouped by language. The distance-based system outperforms both the encoder- and LLM-based methods, primarily due to its superior parameter and model size rankings. In total, the evaluation encompasses and ranks a total of 45 systems by 15 participants, with three systems per participant.

Approach	Macro Recall	Acc. Rank	Param. Rank	Size Rank	Eff. Score	Rank
German						
Distance-based RC	**0.4678**	32	1	1	11.33	5
ITER	0.4162	41	6	5	17.33	15
LLM	0.4378	37	18	19	24.67	29
English						
Distance-based RC	**0.4624**	32	1	1	11.33	2
ITER	0.4313	40	7	6	17.67	13
LLM	0.4466	35	19	20	24.67	26
French						
Distance-based RC	0.4902	25	1	1	9.00	1
ITER	0.4333	37	6	5	16.00	10
LLM	**0.5091**	24	18	19	20.33	17

The distance-based RC achieved the highest efficiency scores across all languages, ranking first for French, second for English, and fifth for German. Its strong performance and high ranking are primarily due to its lack of model parameters. While the distance-based approach did not always achieve the highest accuracy, its performance surpassed that of the small LLM. For English and German, the approach achieved higher accuracy than ITER while requiring practically no additional parameters. For French, however, the LLM-based approach achieved the highest overall accuracy, outperforming the distance-based RC by about 1.9 %. However, this improvement came at a comparatively high computational cost, resulting in a significantly lower efficiency ranking of 17 compared to first place for distance-based RC. While ITER consistently occupies a middle position and outperforms the LLM in the overall efficiency ranking due to its lower parameter and disk size count, its accuracy lags behind both alternatives in all languages. The accuracy rankings in Table 5 further emphasize the distance-based approach's competitive performance across most languages.

Table 5
Official HIPE-2026 accuracy profile rankings across Newspaper test sets, grouped by language. For English and German Test A, as well as Test B, the distance-based measure outperforms our other systems in terms of accuracy. The only exception is the main French test set, where the LLM system ranks 26th. In total, the evaluation encompasses and ranks a total of 45 systems by 15 participants, with three systems per participant.

Approach	Macro Recall	MacroRecall_at	MacroRecall_isAt	Rank
German				
Distance-based RC	**0.4678**	0.3906	0.5450	35
ITER	0.4162	0.3324	0.5000	44
LLM	0.4378	0.3619	0.5136	40
Random Baseline	0.4058	0.3211	0.4906	45
English				
Distance-based RC	**0.4624**	0.3781	0.5467	35
ITER	0.4313	0.3626	0.5	43
LLM	0.4466	0.3251	0.5680	38
Random Baseline	0.3733	0.2532	0.4935	47
French				
Distance-based RC	0.4902	0.4001	0.5802	28
ITER	0.4333	0.3667	0.5	40
LLM	**0.5091**	0.4112	0.6070	27
Random Baseline	0.4355	0.3449	0.5262	49
French (Test B)				
Distance-based RC	**0.3755**	0.3755	–	34
ITER	0.3580	0.3580	–	42
LLM	0.3546	0.3546	–	43
Random Baseline	0.3628	0.3628	–	39

8. Conclusion

We present three efficiency-oriented systems for the HIPE-2026 person-location relationship classification task. The distance-based RC demonstrates that simple positional information is an effective indicator of non-existent relations and offers a valuable baseline that does not require training. As for ITER, this task reveals the limitations of methods that use small embedding models compared to billion-parameter language models. Although it is very efficient in terms of disk size and parameter count, this approach does not keep up with the distance-based baseline, even in terms of raw retrieval accuracy, and it ranks third for all languages. The prompt-based LLM system shows that small, instruction-tuned models can make valid predictions with prompting for AT and ISAT, but performance is still sensitive to prompt design and poor in the PROBABLE class. Overall, our results suggest that systems can benefit

from paying attention to structural text cues for pre-selection when the focus is on efficiency.

Declaration on Generative AI

The authors have not employed any generative AI tools.

References

[1] 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.

[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] M. Hennen, F. Babl, M. Geierhos, ITER: Iterative transformer-based entity recognition and relation extraction, in: Y. Al-Onaizan, M. Bansal, Y. Chen (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024, Findings of ACL, Association for Computational Linguistics, 2024, pp. 11209–11223. URL: https://doi.org/10.18653/v1/2024.findings-emnlp.655.

[4] S. Schweter, L. März, K. Schmid, E. Çano, hmbert: Historical multilingual language models for named entity recognition, in: G. Faggioli, N. Ferro, A. Hanbury, M. Potthast (Eds.), Proceedings of the Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, Bologna, Italy, September 5th - to - 8th, 2022, CEUR Workshop Proceedings, CEUR-WS.org, 2022, pp. 1109–1129. URL: https://ceur-ws.org/Vol-3180/paper-87.pdf.

[5] Y. Tay, M. Dehghani, J. Rao, W. Fedus, S. Abnar, H. W. Chung, S. Narang, D. Yogatama, A. Vaswani, D. Metzler, Scale efficiently: Insights from pre-training and fine-tuning transformers, CoRR abs/2109.10686 (2021). URL: https://arxiv.org/abs/2109.10686. arXiv:2109.10686.

[6] 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, CoRR abs/1910.10683 (2019). URL: http://arxiv.org/abs/1910.10683. arXiv:1910.10683.

[7] P. He, J. Gao, W. Chen, Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing, in: The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023, OpenReview.net, 2023. URL: https://openreview.net/forum?id=sE7-XhLxHA.

[8] Qwen Team, Qwen3.5: Towards native multimodal agents, 2026. URL: https://qwen.ai/blog?id=qwen3.5.

[9] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, W. E. Sayed, Mistral 7b, 2023. URL: https://arxiv.org/abs/2310.06825. arXiv:2310.06825.

[10] Llama Team, The llama 3 herd of models, CoRR abs/2407.21783 (2024). URL: https://doi.org/10.48550/arXiv.2407.21783. doi:10.48550/ARXIV.2407.21783. arXiv:2407.21783.

[11] A. Amini, A. Banaszak, H. Benoit, A. Böök, T. Dakhran, S. Duong, A. Eng, F. Fernandes, M. Härkönen, A. Harrington, R. Hasani, S. Karwa, Y. Khrustalev, M. Labonne, M. Lechner, V. Lechner, S. Lee, Z. Li, N. Loo, J. Marks, E. Mosca, S. J. Paech, P. Pak, R. N. Parnichkun, A. Quach, R. Rogers, D. Rus,

N. Saxena, B. Schlager, T. Seyde, J. T. H. Smith, A. Tadimeti, N. Tumma, Lfm2 technical report, 2025. URL: https://arxiv.org/abs/2511.23404. arXiv:2511.23404.

[12] D. Hendrycks, K. Gimpel, Bridging nonlinearities and stochastic regularizers with gaussian error linear units, CoRR abs/1606.08415 (2016). URL: http://arxiv.org/abs/1606.08415. arXiv:1606.08415.

[13] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, in: 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019, OpenReview.net, 2019. URL: https://openreview.net/forum?id=Bkg6RiCqY7.

A. Prompt Versions

Table 6
Overview of evaluated prompt variants, their design rationale, and observed behavior during development. The full prompt templates used in the experiments are available in an accompanying GitHub Gist.

Prompt	Type	Main design idea	Observed behavior
V1	Long rule-based prompt	Detailed definitions for both relations, explicit warnings against co-mention, nationality, and titles; asks the model to identify evidence before returning JSON.	Reasonable but verbose; not the strongest small-model prompt.
V2	Short compact prompt	Minimal multilingual prompt with direct definitions, concise warnings, and strict JSON output.	Most stable Qwen3.5-2B prompt on gold diagnostics; final efficiency prompt.
V3	Very compact prompt	Shorter classification prompt with compressed definitions and JSON schema.	Worked well for Qwen3.5-0.8B in early English runs, but not consistently best later.
V4	Few-shot prompt	Long multilingual prompt with hand-written examples for label combinations, including PROBABLE.	Encouraged PROBABLE, but often overcorrected and reduced TRUE recall.
V5	Strict/guided-style prompt	More constrained JSON-oriented prompt emphasizing strict labels and concise evidence.	Often conservative; for Qwen it tended to produce no PROBABLE.
V6	Step/order prompt	Asks the model to decide `at` first and `isAt` second, with strict consistency instructions.	More conservative; did not improve over V2.
V7	Decision-tree prompt	Silently classify `at` evidence as DIRECT, INDIRECT, or NONE, then map to TRUE, PROBABLE, or FALSE.	Most competitive prompt compared to V2.

B. Prompt Results

Table 7
Prompt results for English.

Model: Qwen 3.5 2B – EN	isAt		at		Global
Prompt Version	**Accuracy**	**Macro Recall**	**Accuracy**	**Macro Recall**	**Macro Recall**
V1	0.7296	0.5341	0.5863	0.4255	0.4798
V2	0.6971	**0.7090**	0.5798	0.4320	**0.5705**
V3	0.6808	0.5992	0.5244	0.4009	0.5001
V4	0.4332	0.5547	0.1857	0.3682	0.4614
V5	**0.7492**	0.6537	0.6059	0.4306	0.5244
V6	0.7459	0.6325	**0.6319**	0.4515	0.5420
V7	0.6645	0.6487	**0.6319**	**0.4614**	0.5551

Table 8
Prompt results for German.

Model: Qwen 3.5 2B – DE	isAt		at		Global
Prompt Version	**Accuracy**	**Macro Recall**	**Accuracy**	**Macro Recall**	**Macro Recall**
V1	0.8090	0.5000	0.5773	0.3333	0.4167
V2	0.7897	0.5696	**0.6052**	0.3839	0.4767
V3	0.7639	0.5365	0.5815	**0.3875**	0.4620
V4	0.5966	**0.6436**	0.3884	0.3777	**0.5105**
V5	0.8090	0.5000	0.5772	0.3333	0.4167
V6	0.8069	0.4987	0.5751	0.3321	0.4154
V7	**0.8197**	0.5882	0.5987	0.3716	0.4799

Table 9
Prompt results for French.

Model: Qwen 3.5 2B – FR	isAt		at		Global
Prompt Version	**Accuracy**	**Macro Recall**	**Accuracy**	**Macro Recall**	**Macro Recall**
V1	0.7762	0.5000	0.5565	0.3356	0.4178
V2	0.6109	**0.5565**	0.5230	0.3622	**0.4593**
V3	0.7594	0.5191	0.5439	0.3475	0.4333
V4	0.5921	0.5510	0.3243	0.3005	0.4258
V5	0.7678	0.5312	**0.5711**	0.3479	0.4396
V6	**0.7782**	0.5346	**0.5711**	0.3455	0.4401
V7	0.4812	0.5228	0.5084	**0.3686**	0.4457