Extended Overview of HIPE-2026: Evaluating Accurate and Efficient Person–Place Relation Extraction from Multilingual Historical Texts

HIPE at CLEF 2026

Juri Opitz[1,*], Maud Ehrmann[5,*], Corina Raclé[1], Andrianos Michail[1], Matteo Romanello[1,3], Emanuela Boros[2], Simon Gabay[4] and Simon Clematide[1,*]

[1]University of Zurich, Switzerland
[2]École Polytechnique Fédérale de Lausanne (EPFL), Switzerland
[3]Swiss Federal Institute of Technology Zurich (ETHZ), Switzerland
[4]University of Geneva, Switzerland

Abstract

Was this person ever at that place, and if so, when? Answering such questions from noisy, multilingual historical documents is the central challenge of the HIPE-2026 shared task, the third edition of the HIPE-eval evaluation series. Moving from named entity recognition and linking (HIPE-2020, HIPE-2022) to reasoning about relationships between person and location entities, HIPE-2026 targets two temporally grounded relation types: **at**, indicating that a person was present at a location at some point prior to a document's publication date, and **isAt**, indicating presence contemporaneous with that date. This paper presents the results of the evaluation campaign, which confronted 17 participating teams with the challenges of historical language variation, OCR noise, and indirect contextual cues across three languages: French, German, and English. The datasets include historical newspaper texts from the nineteenth and twentieth centuries, as well as a surprise-domain generalization set drawn from early modern French literary texts. A distinctive feature of HIPE-2026 is its three-fold evaluation framework, which assesses predictive accuracy, computational efficiency, and cross-domain generalization, reflecting the practical demands of large-scale historical document processing in the cultural heritage domain. Across more than 40 submitted runs, results reveal a wide range of strategies, from state-of-the-art large language models to lightweight task-specific classifiers, and highlight the trade-offs between accuracy, efficiency, and robustness inherent to historical relation extraction at corpus scale. A systematic error analysis across several data dimensions — OCR quality, publication period, textual proximity, and entity-linking status — finds none of these surface properties to be a robust driver of difficulty, pointing instead to genuine semantic and abductive reasoning as the central bottleneck. System descriptions, datasets, and findings are presented and discussed, offering a detailed picture of the current state of temporally grounded relation extraction for historical documents.

Note: *This is an extended version of the LNCS shared task overview paper [1], with additional material and analyses.*

Keywords

Relation Extraction, Multilingual NLP, Historical Texts, Digital Humanities, Evaluation Campaign

1. Introduction

Historical documents are traces of the past, recording, among other things, what people did, where they went, and how they related to one another. The large-scale digitization of historical newspapers and other cultural heritage collections has made these traces accessible at unprecedented scale, but typically yields texts that are noisy and marked by historical language variation, posing significant challenges for

CLEF 2026 Working Notes, 21 – 24 September 2026, Jena, Germany
*Corresponding author.
jurialexander.opitz@uzh.ch (J. Opitz); maud.ehrmann@epfl.ch (M. Ehrmann); simon.clematide@uzh.ch (S. Clematide)
https://www.juriopitz.com (J. Opitz)
0000-0001-6892-4574 (J. Opitz); 0000-0001-9900-2193 (M. Ehrmann); 0009-0003-8842-8190 (C. Raclé); 0009-0004-1025-7851 (A. Michail); 0000-0002-7406-6286 (M. Romanello); 0000-0001-6299-9452 (E. Boros); 0000-0001-9094-4475 (S. Gabay); 0000-0003-1365-0662 (S. Clematide)
© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

automatic processing [2, 3, 4]. Robust information extraction from these sources – which span multiple languages, document types, and historical periods – is therefore a prerequisite for many downstream uses in digital history, computational humanities, and the social sciences [5, 6, 7, 8, 9, 10, 11]. Advancing and benchmarking such methods is the driving purpose of the HIPE evaluation series[1].

HIPE-2026[2] is the third edition of the series. While HIPE-2020 and HIPE-2022 focused on named entity recognition and linking in multilingual historical documents [12, 13, 14, 15], HIPE-2026 moves toward deeper document understanding, advancing from the identification of entities to reasoning about the relationships between them [16, 1]. The central question is whether a text provides evidence for a relation of physical presence between a person and a place, and whether that relation holds near the publication date of the document. In short: *Who was where, and when?* Such temporally grounded person–place relations underpin the reconstruction of individuals' geographical and temporal trajectories, populating biographical knowledge graphs, supporting prosopographical research, and enabling spatial analysis of historical mobility. Advancing their extraction from historical documents is the core objective of HIPE-2026, in support of digital humanities scholarship.

To this end, the task is deliberately designed to go beyond entity co-occurrence. On the one hand, apparent connections may be spurious: a person and a place mentioned in the same article may be unrelated, or linked only by affiliation rather than physical presence. On the other hand, genuine connections may be hidden: relevant evidence can be indirect (e.g. a person's presence implied through an event whose location must be inferred from context), cross-sentential, or degraded by OCR errors. Beyond the quality of evidence, the task also requires temporal reasoning: systems must distinguish between a person's presence at a place at some point in the past and their presence near the time of publication — a distinction that is often implicit in the text. HIPE-2026 therefore evaluates not only extraction of explicit locative statements, but also temporally grounded reasoning over text where evidence must be weighed rather than simply detected.

Meeting this challenge reliably and at scale requires methods and models that are fit for cultural heritage processing: accurate enough to support scholarly use, scalable enough to process large historical collections, and versatile enough to generalize across domains and document types. For this reason, the campaign includes three complementary evaluation profiles: an *accuracy* profile on historical newspaper data, an *efficiency* profile that combines accuracy with computing resource metadata, and a *generalization* profile on a surprise French test set of literary and historiographical documents. The campaign attracted 17 participating teams, who submitted more than 40 runs across these profiles.

This lab overview defines the task (Section 2), describes the data (Section 3), presents the evaluation framework (Section 4), summarizes the participating systems (Section 5), and reports the official results (Section 6). Compared to the condensed overview [1], this extended version expands the data description and discussion, and substantially extends the error analysis (Section 7) with, inter alia, a study on the utility of the silver data and a systematic analysis of the effect of OCR quality, publication period, person–place proximity, and Wikidata entity-linking status. We conclude with the main lessons from the campaign (Section 8).

All shared task artifacts are publicly available: datasets at https://github.com/hipe-eval/hipe-2026-data and evaluation toolkit (submissions and results) at https://github.com/hipe-eval/hipe-2026-eval.

2. Task Definition

HIPE-2026 frames person–place relation extraction as a document-level classification task. Given a historical document and a set of candidate (person, location) pairs extracted from it, systems must determine what kind of relationship, if any, the text supports for each candidate pair.

For each document, systems are given access to the full text, metadata including language and publication date, and up to 16 sampled candidate pairs. Each candidate pair consists of one person entity and one location entity drawn from the same document. Person and location entities are represented by

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

[2] https://github.com/hipe-eval/hipe-2026

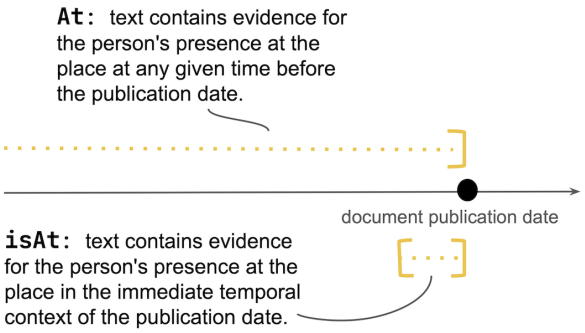

Figure 1: Schematic representation of the two person-place relation types and their temporal scope.

clustered mentions and may carry external Wikidata identifiers when a matching entry exists. Systems must assign labels for two relation types, `at` and `isAt` (see also Figure1).

- The `at` relation captures whether the document supports the claim that the person was at the place at some point before the publication date. It is a three-class task. The label `TRUE` indicates explicit textual evidence, `PROBABLE` indicates a plausible inference from contextual cues, and `FALSE` indicates absence of evidence or contradiction.
- The `isAt` relation is a narrower temporal refinement of `at`. It captures whether the person is located at the place in the immediate temporal horizon of the document, operationalized in the guidelines as roughly up to one month before publication [17]. It is a binary task with labels `TRUE` and `FALSE`.

Conceptually, `isAt=TRUE` presupposes a positive or at least plausible `at` relation. For example, if a text reports that a person is currently visiting Berlin, then both `at` and `isAt` should be positive. If the same text only mentions that the person visited Berlin the previous year, `at` may be positive while `isAt` should be `FALSE`. The official scorer nevertheless evaluates both relation types independently, so that inconsistent predictions can be measured rather than rejected.

The `PROBABLE` label is central to the task design. It reflects an abductive view of interpretation, in which relevant meaning can be supported by the assumptions needed to make discourse coherent rather than by explicit surface statements alone [18]. In historical person–place relation extraction, such cases include event participation, institutional roles, travel reports, and geographically grounded narrative contexts.

3. Data

3.1. Domains and Underlying Datasets

HIPE-2026 uses two evaluation domains, each built on top of an existing annotated dataset that provides person and location entities; the `at`/`isAt` relation annotations are added on top of these. Domain A contains historical newspaper articles in German, English, and French, spanning the nineteenth and twentieth centuries, derived from the HIPE-2022 newspaper data[3], which provides human-annotated named entities and Wikidata entity links [14]; further details on the underlying datasets are available in the HIPE-2022 participation guidelines [19].

Domain B is a "surprise"[4] test set consisting of French literary texts and historiographical works from the sixteenth to eighteenth centuries, included to assess out-of-domain generalization; its entity

[3]In its version v2.1 available here: https://github.com/hipe-eval/HIPE-2022-data/tree/main/data/v2.1
[4]We use the term "surprise" to indicate that participants are not provided with information or examples on/from this dataset prior to evaluation, as we wish to utilize it for testing the out-of-domain generalization of systems.

annotations and Wikidata links come from the FREEM$_{NER}$ corpus (version 6) [20, 21], which includes an extended and improved version of *Presto* core [22].

Table 1 gives an overview of the underlying datasets of HIPE-2026: Domain A spans both national and regional press outlets across the three languages, while Domain B comprises a small set of French literary and historiographical works.

Table 1
Overview of the corpora used for HIPE-2026, including newspaper corpora and literary/historiographical texts, with their full publication periods and the document counts for our sampled gold and silver datasets.

Dataset	Publication Title	Time Period	Gold	Silver
Newspapers				
hipe2020	Alexandria Gazette & Daily Advertiser	1817–1822	5	8
	Chariton Courier	1878–Today	4	5
	Cherokee Phoenix, and Indians' Advocate	1829–1834	3	3
	Escher Tageblatt	1913–1950	2	3
	Gazette de Lausanne	1804–1991	9	91
	Gazette of the United States, & Daily Advertiser	1776–1849	3	7
	Gazette of the United-States	1789–1793	3	6
	Jamestown Alert	1878–1882	2	5
	L'Express	1738–2018	15	56
	L'Impartial	1881–2018	2	12
	Luxemburger Wort	1882–1950	16	32
	Neue Zürcher Zeitung	1780–1950	27	74
	Pullman Herald	1888–1989	4	4
	Putnam County Herald	1903–1922	7	5
	Tabor City Tribune	1946–1991	5	3
	The Delaware Gazette	1805–1810	4	9
	The Detroit Tribune	1935–1966	6	6
	The Elk Mountain Pilot	1880–1924	3	5
	The Independent	1908–1936	3	3
	The North-Carolinian	1839–1861	2	4
letemps	Gazette de Lausanne	1804–1991	18	171
	Journal de Genève	1826–1998	4	33
newseye	Arbeiter Zeitung	1889–1985		1
	Ce soir	1937–1953		3
	La Fronde	1897–1929		5
	La Presse	1836–1952	1	12
	Le Matin	1884–1944		10
	L'OEuvre	1904–1944	3	18
	Marianne	1932–1940		2
	Marie-Claire	1937–1944	1	5
	Neue Freie Presse	1864–1939	1	2
	Paris-soir	1923–1949		2
	Regards	1932–1960	1	4
sonar	Berliner Börsen-Zeitung	1857–1939		3
	Berliner Tageblatt und Handels-Zeitung	1872–1932	1	2
	Deutsches Nachrichtenbüro	1936–1940	2	
	Neueste Mittheilungen	1882–1894	1	1
	Provinzial-Correspondenz	1863–1884	2	1
	Vossische Zeitung	1911–1934	1	1
Literary and Historiographical Texts				
surprise	Essai sur l'histoire générale et sur les moeurs et l'esprit des nations	1756	15	
	Gargantua and Pantagruel	1542	2	
	La Fausset Des Vertus Humaines	1678	1	
	Le Moyen de parvenir	1616	1	
	Les nouvelles récréations et joyeux devis	1561	1	
	Lettres de Bussy-Rabutin	1666	1	
	Lettres de Peiresc	1637	2	
	Voyage de la Pérouse	1797	7	

3.2. Data Construction and Annotation

Person–location candidate pairs were first sampled from the two domains and then annotated with `at` and `isAt` relation labels.

3.2.1. Candidate Pair Construction

Several preprocessing steps were applied to both domains before sampling candidate pairs. After conversion to the HIPE-2026 format[5], entity mentions without a Wikidata identifier, referred to as NIL mentions, were heuristically clustered at the document level if their string similarity (Levenshtein) exceeded 75%. Documents longer than 12,000 characters and statistical outliers in the number of locations or persons, defined as values more than three standard deviations above the mean (computed separately per domain), were then removed. Finally, from the remaining documents, person–location candidate pairs were selected from the Cartesian product, with up to 16 pairs per document. Pairs were sampled with probability proportional to $\exp(-d/T)$, where d is the text distance and $T = 50$ controls the degree of locality, favoring textually close pairs while still allowing occasional long-range ones.

3.2.2. Annotation Process

The `at` relationships were annotated on top of these candidate pairs; Test A evaluates both `at` and `isAt`, while Test B evaluates only `at`. Two annotation tiers were produced. The training and validation sets of Domain A were first annotated by majority vote over three runs of GPT-4.1; this automatically produced annotation is referred to as *silver* data throughout this paper. For each language, a subset of articles was subsequently human-annotated and released as *gold* Train A data, which we treat as the more reliable annotation tier. A sample of 23 human-annotated articles was released early in the evaluation campaign as illustrative examples for participants. The Domain A test set consists of 19 human-annotated articles per language, kept hidden from participants during the evaluation campaign. Since Domain B is used exclusively as test data, it has no silver counterpart; 30 documents were sampled and human-annotated for `at`, roughly matching the size of each Domain A language-specific test set. Human annotations were carried out by eight trained annotators, with at least two native-language annotators per language; in a final round, disagreements were resolved through discussion and adjudication among annotators. The competition data (version 1.0) is archived on Zenodo[6] and released on GitHub.[7]

3.3. Dataset Statistics

We report on the data's composition by language and split (Tables 2–4), and compare overlapping LLM- and human-annotated documents (Table 5). Table 2 reports document counts, pair counts, and label distributions, broken down by source dataset — the disaggregated view underlying the language-level summaries in Tables 3 and 4 below. Across languages and splits, `hipe2020` is consistently the largest or sole contributing source, with `sonar`, `newseye`, and `letemps` contributing smaller and more unevenly distributed shares.

Tables 3 and 4 summarize the released gold and silver splits. The Test A gold sets are balanced by document count, with 19 documents per language, but not by number of candidate pairs: English contains fewer pairs per document than German and French. Across the gold splits, `FALSE` is the most frequent `at` label, while `PROBABLE` is comparatively rare (motivating the use of macro recall as the official metric (Section 4)).

The silver training data is substantially larger, especially for French, and was released to support model development and adaptation. Table 5 compares overlapping documents annotated both by LLM majority vote and by humans. Across all three languages, the LLM annotations assign `PROBABLE` more frequently and `TRUE` less frequently than the human annotations. For `isAt`, the LLM annotations similarly contain

[5] https://github.com/hipe-eval/HIPE-2026-data/blob/main/schemas/hipe-2026-data.schema.json
[6] https://zenodo.org/records/21165041
[7] https://github.com/hipe-eval/HIPE-2026-data/releases/tag/v1.0

Table 2
Dataset statistics by language, split, and source dataset. Label columns give counts for TRUE, PROBABLE, and FALSE for at, and TRUE and FALSE for isAt.

Lang.	Split	Dataset	Docs	Pairs	at_T	at_P	at_F	isAt_T	isAt_F
de	Train Silver	hipe2020	80	1096	156	309	631	104	992
		sonar	6	96	22	20	54	22	74
		newseye	2	32	10	4	18	8	24
		ALL	88	1224	188	333	703	134	1090
	Train Gold	hipe2020	30	402	120	54	228	86	316
		sonar	4	64	15	8	41	3	61
		ALL	34	466	135	62	269	89	377
	Dev	hipe2020	29	398	34	134	230	25	373
		sonar	2	18	7	5	6	4	14
		newseye	1	16	0	8	8	0	16
		ALL	32	432	41	147	244	29	403
	Test	hipe2020	15	178	34	34	110	50	128
		sonar	3	44	13	4	27	15	29
		newseye	1	16	7	0	9	2	14
		ALL	19	238	54	38	146	67	171
fr	Train Silver	hipe2020	127	1805	256	465	1084	193	1612
		letemps	146	2019	308	520	1191	217	1802
		newseye	44	626	72	102	452	56	570
		ALL	317	4450	636	1087	2727	466	3984
	Train Gold	hipe2020	17	246	107	20	119	71	175
		letemps	17	216	74	8	134	36	180
		newseye	1	16	0	0	16	0	16
		ALL	35	478	181	28	269	107	371
	Dev	hipe2020	32	470	59	126	285	47	423
		letemps	58	770	77	190	503	51	719
		newseye	17	258	43	51	164	29	229
		ALL	107	1498	179	367	952	127	1371
	Test	hipe2020	9	117	58	10	49	55	62
		newseye	5	61	19	9	33	18	43
		letemps	5	60	8	0	52	8	52
		ALL	19	238	85	19	134	81	157
	Surprise Test	presto_max_v6	30	480	130	60	290	0	480
en	Train Silver	hipe2020	56	496	98	159	239	77	419
	Train Gold	hipe2020	35	307	125	30	152	83	224
	Dev	hipe2020	17	151	29	54	68	18	133
	Test	hipe2020	19	162	75	28	59	65	97

fewer TRUE labels. This suggests a more conservative or uncertainty-oriented annotation behaviour, in which positive relations are often weakened to PROBABLE or FALSE. Consequently, the silver data should be treated as useful additional supervision, but not as a direct substitute for the human-annotated gold data.

Table 3
Overview of gold training and test data. Label columns report counts for `TRUE` (T), `PROBABLE` (P), and `FALSE` (F); `isAt` has only T/F labels. Test B is evaluated only for `at`.

Split	Lang.	Docs	Pairs	Pairs/doc	Years	at			isAt	
						T	P	F	T	F
Gold Train A	de	34	466	13.71	1818–1948	135	62	269	89	377
	en	35	307	8.77	1790–1960	125	30	152	83	224
	fr	35	478	13.66	1804–2018	181	28	269	107	371
Gold Test A	de	19	238	12.53	1808–1948	54	38	146	67	171
	en	19	162	8.53	1800–1960	75	28	59	65	97
	fr	19	238	12.53	1804–1998	85	19	134	81	157
Gold Test B	fr	30	480	16.00	1542–1797	130	60	290	0	480

Table 4
Overview of silver training and validation data. Label columns report counts for `TRUE` (T), `PROBABLE` (P), and `FALSE` (F); `isAt` has only T/F labels.

Split	Lang.	Docs	Pairs	Pairs/doc	Years	at			isAt	
						T	P	F	T	F
Silver Train A	de	88	1224	13.91	1798–1948	188	333	703	134	1090
	en	56	496	8.86	1790–1960	98	159	239	77	419
	fr	317	4450	14.04	1798–2018	636	1087	2727	466	3984
Silver Dev A	de	32	432	13.50	1818–1948	41	147	244	29	403
	en	17	151	8.88	1790–1920	29	54	68	18	133
	fr	107	1498	14.00	1798–1988	179	367	952	127	1371

Table 5
Comparison of LLM majority-vote annotations and human annotations on overlapping documents. Label columns report counts and row-wise percentages (format: count / percentage) for `TRUE` (T), `PROBABLE` (P), and `FALSE` (F); `isAt` has only T/F labels.

Split	Lang.	Docs	Pairs	at			isAt	
				T	P	F	T	F
de	LLM	24	331	69 / 20.85	87 / 26.28	175 / 52.87	49 / 14.80	282 / 85.20
	Human			100 / 30.21	50 / 15.11	181 / 54.68	68 / 20.54	263 / 79.46
en	LLM	33	290	57 / 19.66	88 / 30.34	145 / 50.00	43 / 14.83	247 / 85.17
	Human			120 / 41.38	28 / 9.66	142 / 48.97	79 / 27.24	211 / 72.76
fr	LLM	24	340	47 / 13.82	77 / 22.65	216 / 63.53	37 / 10.88	303 / 89.12
	Human			113 / 33.24	11 / 3.24	216 / 63.53	71 / 20.88	269 / 79.12

3.4. Illustrative Examples

We illustrate the annotations and the reasoning behind them with examples from both domains and, for Domain A, across languages.

3.4.1. Newspaper Examples (Domain A)

We first provide an original English test set article in Figure 2, illustrating explicit, probable, and absent evidence within a single document; person mentions are shown in bold and location mentions in italics.

The court clerk Tom Scarlett must be present in Putnam County shortly before the publication of the article to issue the bill of sale; therefore both `at` and `isAt` are set to `TRUE`. However, his

NOTICE. By virtue of a bil of sale issued by **Tom Scarlett, Circuit Court Clerk for Putnam County**, *.ennessee*, dated on the 6tU. day of January, 1920, I will expore to sale to tne highest bidder for cash on the 31st day of January, 1920, at 1:0; o'clock, P. M. at the Courthouse door in the Town of *Cookeville*, the following property, to-wit: A house and lot located in the 19th Civil District of *Putnam County, Tennessee*, containing one acre, more or less, bounded on the east by the lands of **J. R. Watson**; on the north by the T. C. R. K. Co.; on the west by the lands of **Phy**; on the south by the lands of **J. R. Watson**; levied on as the lands of **D. Rittenberry**, to satisfy a Judgment in favor of **P. L. Ramsey** ag-ih st him, with interest and all the cost.i of the case. This Jan. 6th, 1920. **L. F. MILLER, Sheriff.**

Person	Location	at	at T/P/F	isAt	isAt T/F
Tom Scarlett	Cookeville	FALSE	**35.6**/**35.6**/28.9	FALSE	40.0/**60.0**
Tom Scarlett	Putnam County	TRUE	**53.3**/28.9/17.8	TRUE	**60.0**/40.0
J. R. Watson	Tennessee	TRUE	**44.4**/33.3/22.2	TRUE	22.2/**77.8**
D. Rittenberry	Putnam County	TRUE	37.8/**40.0**/22.2	FALSE	26.7/**73.3**
P. L. Ramsey	Tennessee	PROBABLE	28.9/**51.1**/20.0	TRUE	26.7/**73.3**
L. F. Miller	Putnam County	TRUE	**46.7**/37.8/15.6	TRUE	**53.3**/46.7

Figure 2: Illustrative example from Test A: a newspaper excerpt published in the *Putnam County Herald* on 29 January 1920, together with its person–place relation annotations. Distribution columns report the percentages of submitted predictions over 45 predictions: T/P/F for `at` (`TRUE`/`PROBABLE`/`FALSE`) and T/F for `isAt` (`TRUE`/`FALSE`); bold values indicate the majority vote.

exact location within the county is unspecified, as he may be present at any courthouse rather than specifically in Cookeville. Consequently, both `at` and `isAt` are labeled as `FALSE` for this relation. The lands of J.R. Watson, bordering the estate being sold, are indicated to lie within Tennessee, where J.R. Watson most likely also is present. The `at` label captures whether the document supports a past or present person–place relation. Given a positive `at` relationship, `isAt` assesses whether the entity can be assumed to be physically present at that location within the time horizon of the article. The bill of sale is issued for the lands of D. Rittenberry, implying that these lands lie within Putnam County; accordingly `at` is set to `TRUE`. However, being involved in legal proceedings, D. Rittenberry may not necessarily be present in Putnam County, thus `isAt` is marked as `FALSE`. The judgment was ruled in favor of P.L. Ramsey, suggesting that he is likely present in Tennessee. However, he might be a creditor residing outside the state, therefore `at` is set to `PROBABLE`. If he indeed is located in Tennessee, he is most likely a fellow citizen of Putnam County, consequently `isAt` is marked as `TRUE`. Note the induced complexity due to long-range pronoun resolution: As indicated by the use of first-person pronouns and the signature of L. F Miller at the end of the article, L. F Miller is the sheriff of Putnam County and assumed present there, thus both `at` and `isAt` are set to `TRUE`. Notably, "I" does not refer to Tom Scarlett, despite his being textually closer.

To illustrate the multilingual setting of Test A, we additionally provide German and French examples with English translations.

The annotations in Figure 3 illustrate how employment and institutional roles support graded inference: presence at a place is rarely stated outright, but can be inferred with varying confidence from what a person is described as doing there. The letters from Paris state that H. v. Vitrolle has been dismissed from his position as Minister of State. This implies that he previously served as Minister of State in France; accordingly, we label both `at` and `isAt` as `TRUE` for the H. v. Vitrolle–France relation. Since the letters originate from Paris and H. v. Vitrolle was employed by the French government, it is reasonable to infer that he was present in Paris. However, he could also have served as Minister of State while primarily representing another region of France. Therefore, we label `at` as `PROBABLE` and `isAt` as `TRUE` for the H. v. Vitrolle–Paris relation, as employment in Paris implies that he was most likely present there within the temporal scope of the article. The text further states that H. v. Vitrolle is the grandson of an innkeeper in the Alps. However, this does not imply that he himself has ever been to the Alps; consequently, no positive relation is annotated. H. Lasitte is identified as the secretary

[German Original] In Briefen aus *Paris* vom 3. Aug. heißt es: „Eine Ordonnanz hat also die Absetzung des **H. von Vitrolles**, als Staatsministers, ausgesprochen. So hat nun dieser Mann, der, aus niedern Volksklassen stammend (er ist der Enkel eines Gastwirths in den *Alpen*), Sekretair des Ra thes der Minister, [...] nach und nach alle seine Aemter und Einkünfte verloren; und alle diese Opfer hat er dem Vergnügen zu intriguiren gebracht! Man gibt ihn als den Verfasser der Adresse an, [...] In dieser Adresse ward ganz ernsthaft die Frage erörtert, ob das gegenwär tige Ministerium *Frankreich* retten könne? Sie entschied, daß es dessen nicht fähig sey, und ersuchte die Mächte, sich ins Mittel zu legen, um dem Könige von *Frankreich* an dere Minister zu geben, die mehr nach dem Geschmacke des **H. v. Vitrolle** wären. — Sein Sekretair, **H. Lasitte**, ist in Verhaft; man hofft, daß seine Erklärungen, und das Verhör, welches **H. v. Vitrolle** unstreitig selbst zu be stehn haben wird, endlich die Urheber jener Adresse aus Tagslicht ziehn, und uns belehren werden, auf welche Klasse von Franzosen sich der öffentliche Unwille werfen muß. Die Dienstentsetzung des **H. v. Vitrolle** hat auf die Gemüther der Freunde des Königs und ihres Landes die beßte Wir kung hervorgebracht. Mit Vergnügen gewahrt man, wie die Regierung eine Kraft entwickelt, welche zur Erhaltung chen Ruhe nothwendig ist. Eine neue Handlung der Streuge bezeichnet neuerdings die Linie, die sie sich vorgezeichnet hat. Sie entsetzte **H. Agier, Substituten des Königlichen Anwalds beym Pariser-Gerichtshofe**, seiner Stelle; [...]

[English Translation] In letters from *Paris* of 3 August it is said: "An ordinance has therefore pronounced the dismissal of **M. de Vitrolles** from his office as Minister of State. Thus this man, who, coming from the lower classes of the people (he is the grandson of an innkeeper in the *Alps*), became secretary to the Council of Ministers, [...] has now gradually lost all his offices and all his revenues; and all these sacrifices he has made for the pleasure of intriguing! He is said to be the author of the address [...] In this address the question was quite seriously discussed whether the present ministry could save *France*. It concluded that it was incapable of doing so, and requested the powers to intervene in order to provide the King of *France* with other ministers more to the taste of **M. de Vitrolles**. **His secretary, M. Lasitte**, is under arrest; it is hoped that his statements, and the examination which **M. de Vitrolles** himself will undoubtedly have to undergo, will at last bring to light the authors of that address, and will inform us upon which class of Frenchmen the public indignation ought to fall. The dismissal of **M. de Vitrolles** has produced the best effect upon the minds of the friends of the King and of their country. It is with satisfaction that one observes the government displaying a firmness which is necessary for the preservation of public tranquillity. A new act of severity once again marks the course which it has prescribed for itself. It has removed **M. Agier, deputy to the King's Attorney at the Paris Court of Justice**, from his office; [...]"

Person	Location	`at`	`at` T/P/F	`isAt`	`isAt` T/F
H. v. Vitrolle	Frankreich	TRUE	**72.7**/13.6/13.6	TRUE	47.7/**52.3**
H. v. Vitrolle	Paris	PROBABLE	36.4/**38.6**/25.0	TRUE	31.8/**68.2**
H. v. Vitrolle	Alpen	FALSE	18.2/25.0/**56.8**	FALSE	4.5/**95.5**
Sekretair, H. Lasitte	Frankreich	TRUE	**56.8**/29.5/13.6	TRUE	40.9/**59.1**
Sekretair, H. Lasitte	Paris	PROBABLE	**40.9**/34.1/25.0	TRUE	34.1/**65.9**
H. Agier	Paris	TRUE	**63.6**/15.9/20.5	TRUE	47.7/**52.3**

Figure 3: Illustrative example from Test A (original German at the top and English translation below) and its person–place relation annotations. Distribution columns report the percentages of submitted `at` predictions over 44 predictions: T/P/F for `at` (`TRUE`/`PROBABLE`/`FALSE`) and T/F for `isAt` (`TRUE`/`FALSE`); bold values indicate the majority vote.

of H. v. Vitrolle. It is therefore reasonable to assume that he was present in the same locations, and the corresponding annotations are identical. Further, H. Agier is stated to be employed at the Parisian Court, therefore we mark both relations as `TRUE`.

The French example in Figure 4 illustrates how an event's stated location can transfer to the people who took part in it, and how physical evidence — a horse-drawn cart's typical range — can support an inference about someone's likely location even without direct confirmation in the text. The text explicitly states that the group of mountaineers had gone to Mont Blanc; therefore, we set both `at` and `isAt` to `TRUE`. The accident occurred on the path from the Grand-Mulet glacier to the Aiguille du Midi station, implying that Aimé Scheller must have been at Grand-Mulet shortly before the accident. The text further states that, following the accident, the injured Aimé Scheller was taken to the station. Consequently, we also set both `at` and `isAt` to `TRUE` for the Aimé Scheller–Aiguille du Midi relation. The second reported accident states that a dog-drawn cart belonging to M. François Savoy, a resident of Bossonens in Fribourg, collided with a motorcycle. We therefore label the M. François Savoy–Bossonens relation as `TRUE` for both `at` and `isAt`, while the M. François Savoy–Genève relation is labeled

[French Original] Les accidents
Un alpiniste genevois atteint par nn bloc de rocher a lé crâne fracturé et menrt *Genève*, 27 août. Dimanche, quatre alpinistes genevois membres du Club des grimpeurs s'étaient rendus au *Mont-Blanc* lorsque, vers 17 heures, au chemin qui conduit du glacier du *Grand-Mulet* à la station de l'*Aiguille du Midi*, l'un d'eux, **Aimé Scheller, 36 ans, sertisseur**, fut soudain atteint par un bloc qui s'était détaché et Sii lui fit une grave blessure à la tête, eux de ses compagnons demeurèrent auprès du blessé tandis que l'autre allait quérir du secours. Une colonne, organisée aussitôt, ramena le blessé à la station. **M. Scheller** reçut les soins d'un médecin français et fut descendu qu'aux *Bossons* en automobile. De là, il fut dirigé sur l'hôpital cantonal de *Genève* où il est mort lundi matin des suites d'une fracture du crâne. Il laisse une femme et une fillette.
Un char à chien contre une motocyclette * Un petit char attelé d'un chien et appairtenant à **M. François Savoy**, demeurant à *Bossonens (Èribourg)*, s'est jeté devant une motocyclette montée par **M. Max Veraaud, pasteur à Mézières (Vaud)**, ayant en croupe son frère. La moto fut renversée avec ses deux occupants ; Je Téservoir à benzine s'étant débouché, l'eesence prit feu au contact du phare à acétylène et se communiqua aux vêtements de **M. Vernaud**. Une application de sacs mouillés l'éteignit. **M. Vernaud** a les sourcils et les cheveux légèrement brûlés et des contusions. La motocyclette, fort endommagée, a dû être conduite au garage.

[English Translation] Accidents
A Geneva mountaineer struck by a rock block, skull fractured and dies *Geneva*, August 27. Sunday, four Geneva mountaineers, members of the Climbing Club, had gone to *Mont Blanc* when, around 5 p.m., on the path leading from the Glacier du *Grand Mulet* to the station of the *Aiguille du Mid*i, one of them, **Aimé Scheller, 36, setter**, was suddenly struck by a block of rock that had detached itself and caused him a serious head injury. Two of his companions remained with the injured man while the other went to seek help. A rescue party, organized immediately, brought the injured man back to the station. **Mr. Scheller** received care from a French doctor and was taken down to Les *Bossons* by car. From there he was taken to the *Geneva* cantonal hospital, where he died Monday morning from the effects of a skull fracture. He leaves behind a wife and a little girl.
A dog cart against a motorcycle * A small cart drawn by a dog and belonging to **Mr. François Savoy**, residing in *Bossonens (Fribourg)*, suddenly moved into the path of a motorcycle ridden by **Mr. Max Vernaud, pastor in Mézières (Vaud)**, who had his brother as passenger. The motorcycle was overturned with its two occupants; the petrol tank was punctured, and the gasoline caught fire on contact with the acetylene headlamp and spread to **Mr. Vernaud**'s clothing. The fire was extinguished by the application of wet sacks. **Mr. Vernaud** suffered slightly burned eyebrows and hair and some bruises. The motorcycle, heavily damaged, had to be taken to a garage.

Person	Location	`at`	`at` T/P/F	`isAt`	`isAt` T/F
Aimé Scheller	Mont Blanc	TRUE	**79.5**/11.4/9.1	TRUE	**54.5**/45.5
Aimé Scheller	Aiguille du Midi	TRUE	**72.7**/11.4/15.9	TRUE	40.9/**59.1**
M. François Savoy	Bossonens	TRUE	**54.5**/13.6/31.8	TRUE	**62.5**/37.5
M. François Savoy	Genève	FALSE	27.3/4.5/**68.2**	FALSE	13.6/**86.4**
M. Max Veraaud	Bossonens	PROBABLE	36.4/6.8/**56.8**	TRUE	22.7/**77.3**
M. Max Veraaud	Èribourg	PROBABLE	36.4/11.4/**52.3**	TRUE	27.3/**72.7**

Figure 4: Illustrative example from Test A (original French at the top and English translation below) and its person–place relation annotations (Document ID: `GDL-1928-08-28-a-i0014`). Distribution columns report the percentages of submitted predictions over 44 predictions: T/P/F for `at` (`TRUE`/`PROBABLE`/`FALSE`) and T/F for `isAt` (`TRUE`/`FALSE`); bold values indicate the majority vote.

`FALSE` for both. An interesting case is the relation involving M. Max Vernaud, who is identified as the motorcycle rider. A dog-drawn cart is unlikely to have traveled far from its owner's residence in Bossonens, making it reasonable to infer that the collision occurred in or near Bossonens. Accordingly, we label both the M. Max Vernaud–Bossonens and M. Max Vernaud–Fribourg relations as `PROBABLE` for `at` and `TRUE` for `isAt`.

3.4.2. Literary and Historiographical Examples (Domain B)

Domain B differs substantially from Domain A in genre and period; Figure 5 shows an excerpt from *Gargantua and Pantagruel*, a sixteenth-century series of novels by François Rabelais, illustrating how narrative movement toward a destination must be distinguished from movement already completed, and how genuine ambiguity about a journey's origin can leave a relation only `PROBABLE` rather than resolvable to `TRUE` or `FALSE`.

[French Original] Si de ce vous esmerveillez : esmerveillez vous dadvantaige de la queue des beliers de *Scythie* : que pesoit plus de trente livres, & des moutons de Surie, esquelz fault (si **Tenaud** dict vray) affuster une charrette au cul, pour la porter tant elle est longue & pesante. Vous ne l' avez pas telle vous aultres paillardes de plat pays. Et [la jument] fut amenee par mer en troys carracques & un brigantin jusques au *port de Olone* en Thalmondoys Lors que **Grandgousier** la veit, Voicy (dist il) bien le cas pour porter mon filz a *Paris*. Or ca de par dieu, tout yra bien. Il sera grand clerc on temps advenir. Si n' estoient messieurs les bestes, nous vivrions comme clercs. Au lendemain apres boyre (comme entendez) prindrent chemin, **Gargantua** son precepteur **Ponocrates** & ses gens, ensemble eulx **Eudemon** le jeune paige. Et par ce que c' estoit en temps serain & bien attrempé, son pere luy feist faire des botes fauves. **Babin** les nomme brodequins. Ainsi joyeusement passerent leur grand chemin : & tousjours grand chere : jusques au dessus de *Orleans*.

[English Translation] If this astonishes you, be yet more astonished at the tails of the rams of *Scythia*, which weighed more than thirty pounds, and at the sheep of Syria, to which one must (if **Tenaud** speaks true) attach a cart to their hindquarters to carry it, so long and heavy is it. You have nothing of the sort, you hussies of the flatlands. [The mare] was brought by sea in three carracks and a brigantine as far as the *port of Olonne* in Talmondais. When **Grandgousier** saw it, he said: "Here is just the thing to carry my son to *Paris*. Come now, by God, all will be well. He will be a great scholar in time to come. Were it not for these gentlemen the beasts, we would live like scholars." The next morning, after drinking (as you will understand), they took to the road: **Gargantua**, his tutor **Ponocrates** and his men, together with **Eudemon**, the young page. And since the weather was fair and mild, his father had him fitted with tawny boots — which **Babin** calls buskins. And so they passed merrily along their way, always in good cheer, as far as the outskirts of *Orléans*.

Person	Location	`at`	`at` T/P/F
Tenaud	Scythie	FALSE	15.9/4.5/**79.5**
Grandgousier	port de Olone	PROBABLE	**59.1**/13.6/27.3
Gargantua	Paris	FALSE	15.9/27.3/**56.8**
Ponocrates	Orleans	TRUE	**47.7**/13.6/38.6
Eudemon	Orleans	TRUE	**47.7**/13.6/38.6
Babin	Orleans	FALSE	27.3/27.3/**45.5**

Figure 5: Illustrative example from Test B: an excerpt from Rabelais' *Gargantua and Pantagruel* (original French at the top and English translation below) and its person–place relation annotations. The distribution column reports the percentages of submitted `at` predictions over 44 predictions as T/P/F (`TRUE`/`PROBABLE`/`FALSE`); bold values indicate the majority vote.

The annotations of the surprise text may be explained as follows: There is no evidence Tenaud ever was in Scythia, only that he described Scythian rams, therefore `at` is marked as `FALSE`. It is unknown where Grandgousier is located. It is likely that he saw this mare at the port of Olonne, but the mare also could have been brought to his location, so we set `at` to `PROBABLE` to reflect this uncertainty. Gargantua is on his way to Paris but has not arrived, therefore `at` is `FALSE` for this relationship. Gargantua and his companions Ponocrates and Eudemon are however in the lands beyond Orléans, which indicates them having passed through Orléans. Thus, we mark their relationships with Orléans as `TRUE`. Babin is only mentioned in reference to what he calls a specific type of boots, with no previous mention or any indication of being part of the travels, consequently we set `at` to `FALSE`.

Because the surprise set also contains historiographical texts, we include a second Test B example from a historical account.

The annotations provided in Figure 6 can be interpreted as follows. Although the text does not explicitly state that Gregory VII was a European ruler, this can reasonably be inferred from the historical context. Accordingly, the Gregory VII–Europe relation is labeled as `PROBABLE` for `at`. The excerpt states that Henry IV was buried in the cathedral of Liège. Since burial implies that the individual was physically present at the location at some point, the Henry IV–Liège relation is marked as `TRUE` for `at`. Despite the textual proximity, there is no evidence that Frédéric I was ever present in Rome; consequently, no positive relation is annotated. The text strongly suggests that Pascal II refers to the pope, although some ambiguity remains, as the name could instead refer to an envoy acting on the pope's behalf. We therefore label the Pascal II–Rome relation as `PROBABLE` for `at`. Furthermore, Pascal II is described as

[French Original] Deux légats l'y déposent; deux députés de la diéte, envoyés par son fils, lui arrachent les ornemens impériaux. Bientôt après, échapé de sa prison, pauvre, errant et sans secours, il mourut à *Liége* plus misérable encor que **Grégoire VII** et plus obscurément, après avoir si longtems tenu les yeux de l'*Europe* ouverts sur ses victoires, sur ses grandeurs, sur ses infortunes, sur ses vices et ses vertus. Il s'écriait en mourant : Dieu des vengeances, vous vengerez ce parricide. de tout tems les hommes ont imaginé que Dieu exauçait les malédictions des mourans, et surtout des péres. Erreur utile et respectable, si elle arrêtait le crime. Une autre erreur plus généralement répandue parmi nous faisait croire que les excommuniés étaient damnés. Le fils d'**Henri IV** mit le comble à son impieté en affectant la piété atroce de déterrer le corps de son pére inhumé dans la *cathédrale de Liége*, et de le faire porter dans une cave à *Spire*. Ce fut ainsi qu' il consomma son hypocrisie dénaturée.
CHAPITRE 37 De l' **empereur Henri V** et de *Rome*, jusqu' à **Fréderic I**.
Ce même **Henri V** qui avait détrôné et exhumé son pére, une bulle du pape à la main, soûtint les mêmes droits de **Henri IV** contre l' église, dès qu'il fut maître. Déja les papes savaient se faire un apui des rois de France contre les empereurs. Les prétentions de la papauté attaquaient, il est vrai, tous les souverains ; mais on ménageait par des négociations ceux qu'on insultait par des bulles. Les rois de France ne prétendaient rien à *Rome*. Ils étaient voisins et jaloux de l'*Allemagne*. Ils étaient donc les alliés naturels des papes. Aussi **Pascal II** vint en *France*, et implora le secours du **roi Philippe** : ses successeurs en usèrent souvent de même. Les domaines que possédait le st siége, le droit qu'il reclamait en vertu des prétendues donations de **Pepin** et de **Charlemagne**, la donation réelle de la comtesse Matilde, ne faisaient point encor du pape un souverain puissant. Toutes ces terres étaient ou contestées ou possédées par d' autres.

[English Translation] Two legates deposed him there; two deputies of the Diet, sent by his son, tore the imperial insignia from him. Soon afterward, having escaped from his prison, poor, wandering, and without assistance, he died at *Liège*, even more miserable than **Gregory VII**, and more obscurely, after having for so long held the eyes of *Europe* fixed upon his victories, his greatness, his misfortunes, his vices, and his virtues. As he was dying, he cried out: "God of vengeance, you will avenge this parricide." At all times men have imagined that God granted the curses of the dying, and especially those of fathers. An error useful and respectable, if it restrained crime. Another error, more generally widespread among us, held that the excommunicated were damned. The son of **Henry IV** carried his impiety to the utmost by affecting an atrocious piety: he had the body of his father, buried in the *cathedral of Liège*, disinterred and carried into a vault at *Speyer*. Thus did he consummate his unnatural hypocrisy.
CHAPTER 37 Of the **Emperor Henry V** and of *Rome*, down to **Frederick I**
This same **Henry V**, who had dethroned and disinterred his father, with a papal bull in his hand, upheld the same rights of **Henry IV** against the Church as soon as he became master. Already the popes knew how to make use of the kings of France as a support against the emperors. The claims of the papacy did indeed attack all sovereigns; but through negotiations they spared those whom they insulted through bulls. The kings of France laid no claim to *Rome*. They were neighbors of *Germany* and jealous of her. They were therefore the natural allies of the popes. Thus **Paschal II** came to *France* and implored the assistance of **King Philip**; his successors often did the same. The territories possessed by the Holy See, the rights which it claimed by virtue of the alleged donations of **Pepin** and **Charlemagne**, and the genuine donation of Countess Matilda, still did not make the pope a powerful sovereign. All these lands were either disputed or possessed by others.

Person	Location	`at`	`at` T/P/F
Grégoire VII	Europe	PROBABLE	**40.9**/31.8/27.3
Henri IV	cathédrale de Liège	TRUE	**63.6**/4.5/31.8
Fréderic I	Rome	FALSE	22.7/11.4/**65.9**
Pascal II	Rome	PROBABLE	25.0/20.5/**54.5**
roi Philippe	France	TRUE	45.8/**54.2**/0.0
Charlemagne	France	PROBABLE	4.2/37.5/**58.3**

Figure 6: Illustrative example from Test B (original French at the top and English translation below) and its person–place relation annotations (Document ID: `1756_Voltaire_HistoireGenerale_chunk204`). The distribution column reports the percentages of submitted `at` predictions over 44 predictions as T/P/F (`TRUE`/`PROBABLE`/`FALSE`); bold values indicate the majority vote.

imploring the assistance of King Philippe. Given the statement that the French were natural allies of the popes, it is reasonable to infer that Philippe was King of France; accordingly, we label this relation as `TRUE` for `at`. Finally, although the text does not explicitly state that Charlemagne was present in France, this can be inferred from the historical context. We therefore label the Charlemagne–France relation as `PROBABLE` for `at`.

4. Evaluation Framework

HIPE-2026 adopts three complementary evaluation profiles, accuracy, efficiency, and generalization – each targeting a different dimension of system performance.

4.1. Accuracy Profile

The official metric is macro-averaged recall, also known as balanced accuracy. For a label ℓ, recall is computed as

$$\text{Recall}(\ell) = \frac{\#\text{correct predictions with gold label } \ell}{\#\text{instances with gold label } \ell}.$$

The macro-averaged recall (MR) of a relation type is the arithmetic mean of the recalls of its labels. This choice gives equal weight to rare and frequent labels [23, 24], which is important for less frequent labels such as `at=PROBABLE` and `isAt=TRUE`.

For Test A, `at` is evaluated as a three-class problem and `isAt` as a two-class problem. The language-specific profile score is

$$\text{Score}_A = \frac{\text{MacroRecall}_{\texttt{at}} + \text{MacroRecall}_{\texttt{isAt}}}{2}.$$

The overall Test A ranking averages this score across German, English, and French for runs that submitted all three language files. Partial submissions are reported only in language-specific tables. For Test B, only `at` is evaluated, so the profile score is simply $\text{MacroRecall}_{\texttt{at}}$.

4.2. Efficiency Profile

The efficiency profile ranks each run by combining its accuracy rank with resource ranks derived from two computing-resource metadata collected from participants: model parameter count and deployed model size. We use ranks rather than absolute values to keep the two resource dimensions comparable and the evaluation robust to outliers. The official efficiency score is the mean of these three ranks:

$$R_{\mathit{eff}} = \frac{r_{acc} + r_{params} + r_{size}}{3}.$$

Lower efficiency scores are better. Runs with missing resource metadata are assigned the worst resource rank.

4.3. Generalization Profile

The generalization profile tests system accuracy on the surprise Test B, assessing the ability of systems to transfer beyond the historical newspaper domain to literary and historiographical texts of a different period and style. Only the `at` relation is evaluated, as a three-class problem, and the profile score is $\text{MacroRecall}_{\texttt{at}}$.

4.4. Scoring and Ranking

Teams were permitted to submit up to three runs per language, with predictions for both Test A and Test B (see the participation guidelines [25] for full submission rules). All evaluation profiles are scored using the evaluation toolkit available at https://github.com/hipe-eval/hipe-2026-eval. The toolkit validates submissions against the HIPE-2026 JSON schema and produces per-run diagnostic reports alongside the official scores. Systems are ranked within each profile independently; a run must cover all three languages of Test A to appear in the overall accuracy and efficiency rankings.

5. System Descriptions

Seventeen teams submitted at least one official run, alongside two organizer-provided baselines. The evaluation repository contains 185 prediction files in total. Most teams submitted complete Test A runs covering all three languages as well as Test B runs for the surprise set; one team submitted only an English Test A run.

5.1. Organizer Baselines

HIPE-Random-Baseline. This baseline assigns labels by uniform random sampling from the label inventory of each relation type: `TRUE`, `PROBABLE`, and `FALSE` for `at`, and `TRUE` and `FALSE` for `isAt`. It does not use document text, metadata, entity information, or training data.

HIPE-Ministral-Baseline. The LLM baseline is a minimal zero-shot system for person–place relation qualification, designed as a simple and reproducible reference point for the shared task. The aim is to test how far a small local instruction-following model can get on the task when given only a clear task definition and the metadata already available in the shared-task format, without task-specific fine-tuning, few-shot examples, retrieval augmentation, external knowledge sources, or ensemble strategies. Its design therefore isolates the behavior of a single generative model under deterministic decoding, rather than the effect of additional components or strategies.

This baseline uses Ministral-3-3B-Instruct-2512[8], a 3-billion-parameter instruction-following language model, executed locally in a GGUF setup with greedy decoding, i.e. temperature set to 0.0 [26]. It formulates relation prediction as a prompted classification problem using a single instruction template: for each pre-identified person–location pair, the model assigns labels for the two target relations, `at` and `isAt`. The system operates in a straightforward document-level loop over candidate pairs, produces structured JSON outputs, and is implemented modularly, allowing easy modification of the prompt, decoding settings, and inference routine.

The prompt casts the model as a historian working on person–location relations in noisy multilingual European newspapers and explicitly warns it that the input may contain OCR errors. For each candidate pair, the model receives the article text, the person–location pair to classify, the document language and publication date, and the available entity metadata. For both the person and the location entity, this metadata includes the internal entity identifier, the Wikidata QID when available, and the list of clustered mention strings. The model is instructed to produce a short text-grounded explanation before each decision: first an `at_explanation` followed by the `at` label, and then an `isAt_explanation` followed by the `isAt` label. These explanations are intended to support error inspection and debugging rather than to serve as additional evaluated outputs. The final response must follow the required JSON shape, which preserves the exact person and place strings and contains the two explanations together with the two relation labels. We do not reproduce the full prompt here for reasons of space, but make it available in the public baseline repository.[9]

5.2. Participating Systems

Awakened. The team explored both prompting-based and supervised approaches for multilingual person–place relation extraction [27]. Two submitted runs relied on Claude Sonnet 4 with carefully designed prompts and few-shot examples, while incorporating external knowledge from Wikidata, including biographical and geographical information associated with linked entities. Additional post-processing rules enforced temporal consistency using birth and death dates. A third run employed a fine-tuned XLM-RoBERTa-large model augmented with knowledge-graph-derived and pattern-based

[8] https://huggingface.co/mistralai/Ministral-3-3B-Instruct-2512

[9] The main prompt is available at https://github.com/hipe-eval/HIPE-2026-llm-baseline/blob/main/prompts/classify_pair.txt. The metadata insertion and detailed parametrization are given at https://github.com/hipe-eval/HIPE-2026-llm-baseline/blob/main/src/hipe2026_mistral_baseline/prompting.py.

Table 6
Participating teams and coarse methodological characteristics of their official submissions. A check mark indicates that at least one submitted run from the team used the corresponding component. *Supervised* includes both classical supervised methods and task-specific adaptation of pretrained models, including fine-tuning, parameter-efficient fine-tuning, and distillation. We use "–" to indicate missing information.

Team	Country	Runs	LLM	Superv.	PLM enc.	Feat./rules	Ext. knowl.
Awakened	RO	3	✓	✓	✓	✓	✓
BIU_NLP	IL	3	✓				
DS@GT	US	3		✓		✓	✓
FI-CODE	DE	3	✓	✓	✓	✓	
FourBytes	IN	1	–	–	–	–	–
GippLab	DE	3		✓			
Hansel&Gretel	IN	3	✓	✓		✓	✓
INSA Lyon	FR	3	✓	✓	✓	✓	
MaxFo-Ajie	CN	3	✓			✓	
MILRIT	FR	3		✓	✓	✓	
ROSTI	FR	3		✓		✓	
Rittik&Souvik	IN	1		✓	✓		
Spinfo	DE	3	✓			✓	
UMUTEAM	ES	3	✓	✓	✓	✓	
VerbaNexAI I	CO	2		✓	✓	✓	
VerbaNexAI II	CO	3	✓	✓	✓		
whereami	EG	2	✓	✓		✓	
HIPE-Random-Baseline	CH	1					
HIPE-Ministral-Baseline	CH	1	✓				

features. The supervised model further incorporated a consistency-aware objective to discourage logically incompatible label assignments. The submitted runs therefore compared knowledge-enhanced prompting with feature-augmented multilingual encoder models.

BIU_NLP. The team approached the task as a zero- and few-shot prompting problem without task-specific fine-tuning [28]. The team evaluated several open-weight instruction-following language models, including Gemma and Mistral variants. For each person–location pair, the models received a structured prompt requiring relation classification together with a short explanatory rationale. Experiments explored prompts written either in English or in the source document language, as well as different numbers of in-context examples. The submitted runs corresponded to different language model backbones and prompting configurations selected on development data.

DS@GT. DS@GT proposed a lightweight relation extraction pipeline based on dependency graphs and engineered features [29]. Documents were first parsed with Stanza and transformed into document-level graphs enriched with cross-sentence connections derived from entity linking, lexical similarity, and geographic containment information. From these structures, the system extracted a set of proximity-based, syntactic, and graph-derived features. Classification was then performed using either traditional machine learning ensembles or graph neural network architectures. Different combinations of graph-based and feature-based classifiers were selected for each language, enabling comparison between neural and non-neural prediction strategies while avoiding the use of pretrained language models during classification.

FI-CODE. FI-CODE investigated three diverse approaches with a focus on computational efficiency and compact model size [30]. One system (run 2) employed a rule-based strategy that inferred relations from the textual distance between person and location mentions. Another system (run 1) adapted the ITER relation extraction architecture to a relation classification setting by providing entity annotations directly as input. The third system (run 3) explored prompt-based inference with large language models, comparing alternative prompting styles, output formats, and model scales. Together, the submissions examined the trade-offs between heuristic methods, encoder-based models, and instruction-following language models.

FourBytes. The team did not provide detailed information about their system. The reported efficiency metadata, including parameter count and disk size, is consistent with the XLM-RoBERTa base model [31], suggesting that the system may have used this model or a closely related architecture.

GippLab. GippLab [32] explored parameter-efficient fine-tuning of multilingual LLMs from the Qwen family [33]. Multiple model sizes were adapted using LoRA, with all linear layers receiving trainable adapters. Rather than training separate systems for individual languages, the team adopted a unified multilingual setup that jointly leveraged English, French, and German training data. The submitted runs corresponded to different Qwen3.5 model sizes selected from a broader set of experiments. This design emphasized scalable multilingual adaptation through lightweight fine-tuning.

Hansel&Gretel. Hansel&Gretel investigated three complementary approaches spanning supervised fine-tuning, multi-agent reasoning, and prompted inference [34]. The first run fine-tuned a Qwen2.5 [35] instruction model using supervised generation of structured JSON outputs, complemented by Wikidata-based temporal consistency rules. The second run introduced a multi-agent framework in which specialized historian and geographer models produced independent assessments that were reconciled by an arbiter model. The third run relied on chain-of-thought prompting with a large reasoning model and dynamically retrieved Wikidata information. Across all runs, external biographical and geographic knowledge was incorporated to support temporal and spatial reasoning.

INSA Lyon. The INSA Lyon system combined handcrafted features, multilingual encoders, and large language models within a unified ensemble framework [36]. Candidate pairs were enriched with contextual and temporal information, including temporal expressions, tense-related cues, and lexical indicators. The resulting representations were processed by several model families, including feature-based classifiers, fine-tuned multilingual Transformer encoders, and zero-shot large language models. A dedicated decision layer aggregated the outputs through voting and stacking strategies. The submitted runs corresponded to the complete heterogeneous ensemble (run 1), a compact encoder-only baseline (run 2), and an ensemble configuration without large language model inference (run 3).

MaxFo-Ajie. MaxFo-Ajie developed a prompt-based multilingual relation extraction pipeline using the GPT-5.5 model accessed through an OpenAI-compatible interface [37]. The system classified person–location pairs with structured prompts and language-matched few-shot examples drawn from the training data. The prompting strategy encouraged conservative predictions in cases of limited evidence and enforced a strict JSON output format. The submitted runs compared inline (run 1) and multi-turn document-level few-shot prompting (run 2), as well as pair-level inference (run 3). Additional components handled normalization, robust output parsing, and schema-compliant prediction generation.

MILRIT. MILRIT submitted three systems representing different approaches to relation extraction [38]. Two runs were based on GLiREL [39], a general-purpose relation extraction framework. One variant fine-tuned GLiREL directly on the shared-task data (run 1), while another used GLiREL-generated relation scores as features for lightweight downstream classifiers (run 2). The third run employed a multilingual DeBERTa encoder augmented with entity markers, publication-date information, and multi-view learning using large-language-model-generated input enrichments. The submissions therefore contrasted direct adaptation of a generalist relation extraction model with task-specific multilingual encoder architectures.

ROSTI. ROSTI pursued a resource-efficient approach centered on logistic regression and sparse textual representations [40]. Sentences containing candidate entities were encoded using TF–IDF features derived from both sentence content and entity mentions. Separate classifiers were trained for the two target relations and for each language. Predictions were generated at the sentence-pair level and subsequently aggregated across all occurrences of a given person–location pair. The resulting

system avoided external resources and large pretrained models while providing a compact baseline for multilingual historical relation extraction.

Rittik&Souvik. Rittik&Souvik addressed the English-language setting using a transfer-learning approach based on BERT [41]. Person–location classification was formulated as a sequence-pair task through a question-answering-inspired input representation. To reduce overfitting in the low-resource setting, the team adopted a parameter-efficient fine-tuning strategy that updated only the upper layers of the encoder and the classification head. Class imbalance was addressed through cost-sensitive learning, with increased emphasis on minority labels during optimization. The system therefore combined lightweight adaptation with task-specific input reformulation.

SPINFO. The team explored prompting strategies for open-weight large language models [42]. After evaluating several model families, the team selected a large reasoning model for all submitted runs. Person–location pairs were processed independently using structured prompts specifying task definitions, label inventories, and output constraints (run 1). Further experiments investigated the effect of temporal information extracted with HeidelTime and conditional prompting strategies that linked predictions across subtasks (run 2), and prompts translated into the document language (run 3). The submitted systems focused on understanding how prompting design influences multilingual historical relation extraction.

UMUTeam. UMUTeam compared supervised multilingual classification and zero-shot large language model reasoning [43]. The supervised approach (run 1) used an XLM-RoBERTa cross-encoder with separate prediction heads for the two target relations. In parallel, two instruction-following language models were evaluated through structured prompting (run 2, run 3). All systems incorporated document metadata, publication dates, entity mentions, and local contextual evidence extracted around candidate pairs. The submitted runs therefore examined the relative strengths of fine-tuned multilingual encoders and prompt-based inference for historical relation extraction.

VerbaNex AI I. VerbaNex AI I proposed a multilingual multitask architecture built on XLM-RoBERTa and enhanced with temporal reasoning components [44]. The system combined contextual representations with named entity recognition features, class balancing strategies, and hyperparameter optimization. During preprocessing, contextual evidence windows were extracted around candidate entities, while a lightweight Temporal Transformer introduced temporal positional information and semantic refinement mechanisms. The submitted approach was designed to capture both immediate and historical person–location associations in multilingual historical documents.

VerbaNex AI II. VerbaNex AI II explored two complementary paradigms: an NLI-based encoder architecture and a self-distilled instruction-following language model [45]. The NLI system reformulated relation extraction as textual inference, pairing document content with hypotheses describing candidate relations. The submitted system relied on Nemotron-3-Nano-4B combined with DSPy-based prompt optimization and self-distillation. Optimized prompting programs were used to generate high-quality training examples, which subsequently served for LoRA-based adaptation of the same model. The final system operated directly on multilingual OCR text and employed a unified inference pipeline across languages.

whereami. The team developed a reasoning-oriented system based on multi-stage knowledge distillation [46]. The approach began with an exploration of prompting strategies across several large language models to identify a suitable teacher model. The teacher was then adapted through supervised fine-tuning and used to generate reasoning traces over the training corpus. These traces were subsequently distilled into a smaller Gemma-based student model through response-level supervision.

To improve prediction consistency, the system incorporated rule-based constraints covering entity compatibility, relation directionality, contextual triggers, and locality constraints.

5.3. Methodological Trends Across Systems

The submitted systems illustrate a broad methodological diversity. Several teams adopted prompting-based approaches using large language models, either through proprietary systems (e.g., Awakened, MaxFo-Ajie) or open-weight models (e.g., BIU_NLP, Spinfo, UMUTeam, or FI-CODE). These approaches generally framed person–place relation extraction as an instruction-following or reasoning task, often relying on few-shot examples, structured output formats, and carefully engineered prompts. A second group focused on adapting open-weight generative models through supervised fine-tuning or parameter-efficient techniques such as LoRA, including GippLab, Hansel&Gretel, whereami, and VerbaNex AI II. These systems explored the extent to which task-specific supervision could be transferred into compact multilingual language models while retaining the flexibility of generative architectures.

Another prominent line of work relied on encoder-based classification models, often built on multilingual Transformer encoders such as XLM-RoBERTa, DeBERTa, or NLI-oriented architectures. Examples include Awakened, INSA Lyon, MILRIT, Rittik&Souvik, UMUTeam, and VerbaNex AI I, which typically formulated the task as supervised pair classification and incorporated contextual, temporal, or entity-specific representations. In parallel, several teams investigated lightweight alternatives that emphasized efficiency and interpretability, including DS@GT's graph-based feature extraction pipeline, ROSTI's TF–IDF and logistic regression framework, and FI-CODE's distance-based heuristics and compact relation classification models.

Many submissions further augmented their core models with external knowledge or rule-based components. Wikidata-derived biographical and geographical information was frequently used to support temporal reasoning and consistency checking, while several systems incorporated post-processing rules, temporal constraints, entity-type validation, or retrieval-based enrichment. Finally, a number of teams explored ensemble and hybrid architectures that combined heterogeneous model families, most notably INSA Lyon, DS@GT, Hansel&Gretel, and MILRIT. Overall, the submissions demonstrate that the task can be addressed using a wide spectrum of techniques, ranging from highly efficient symbolic and statistical methods to large-scale reasoning models, reflecting the dual emphasis of the shared task on predictive quality and computational efficiency.

6. Main Results

6.1. Accuracy Profile on Test A

Table 7 reports the official Test A ranking for complete submissions, including participant runs and organizer baselines. The official ranking is based on the mean of the three language-specific Test A scores.

Spinfo achieved the highest Test A score with run 1 (0.748), followed by MaxFo-Ajie run 1 (0.700) and whereami run 1 (0.688). The leading run improves over the Ministral baseline by 0.166 absolute points. The language columns also show that no single team is uniformly dominant in every setting: Spinfo leads German and French, while MaxFo-Ajie leads English. This split suggests that no single prompting or modeling strategy dominates across all three languages, and that language-specific tuning (e.g. Spinfo's translated-prompt variant, run 3) can matter as much as the underlying architecture.

The per-language winners in Table 8 also suggests that high `isAt` macro recall was often crucial for the best German and French scores, which points to the temporal-scoping judgment — rather than the coarser `at` decision — as the more decisive skill separating strong systems in those two languages.

Table 7
Accuracy Profile ranking on Test A. Only runs submitted for all three Test A languages are included. The language columns report language-specific profile scores, i.e., the mean of `at` and `isAt` macro recall. Avg. is the official Test A score, obtained by averaging these three language-specific scores.

Rank	T-Rank	Team	Run	de	en	fr	Avg. ↑
1	1	Spinfo	1	0.7608	0.7279	**0.7551**	**0.7479**
2		Spinfo	3	**0.7710**	0.6808	0.7349	0.7289
3	2	MaxFo-Ajie	1	0.6720	**0.7493**	0.6790	0.7001
4		Spinfo	2	0.6860	0.6427	0.7383	0.6890
5	3	whereami	1	0.7041	0.7023	0.6578	0.6880
6		whereami	2	0.7072	0.6992	0.6435	0.6833
7		MaxFo-Ajie	2	0.6485	0.7142	0.6444	0.6690
8	4	Awakened	3	0.6746	0.6415	0.6854	0.6671
9		Awakened	1	0.6739	0.5848	0.7163	0.6584
10		MaxFo-Ajie	3	0.6335	0.7003	0.6295	0.6544
11	5	INSA Lyon	1	0.6725	0.5985	0.6459	0.6390
12	6	gipplab	2	0.5941	0.6174	0.6697	0.6271
13	7	Hansel&Gretel	3	0.5994	0.6367	0.6303	0.6221
14		gipplab	1	0.5957	0.6143	0.6322	0.6141
15	8	MILRIT	3	0.5886	0.5881	0.6087	0.5951
16	9	UMUTEAM	2	0.6305	0.5384	0.5880	0.5856
17		Ministral baseline	1	0.5553	0.5551	0.6349	0.5818
18	10	VerbaNexAI II	3	0.5697	0.5812	0.5877	0.5795
19		Hansel&Gretel	2	0.5254	0.5866	0.6243	0.5788
20	11	BIU_NLP	2	0.5050	0.5762	0.6529	0.5781
21		MILRIT	1	0.5802	0.4808	0.6261	0.5623
22		Awakened	2	0.5417	0.4989	0.6076	0.5494
23		Hansel&Gretel	1	0.5210	0.5976	0.5189	0.5458
24		BIU_NLP	3	0.5451	0.5101	0.5617	0.5390
25		VerbaNexAI II	2	0.5022	0.5718	0.4822	0.5187
26	12	DS@GT_HIPE	1	0.5160	0.5103	0.5162	0.5142
27		gipplab	3	0.5181	0.5221	0.4805	0.5069
28		VerbaNexAI II	1	0.4808	0.5607	0.4597	0.5004
29	13	VerbaNexAI I	2	0.5071	0.4808	0.4648	0.4842
30		DS@GT_HIPE	2	0.4937	0.4401	0.5171	0.4836
31		DS@GT_HIPE	3	0.5054	0.4834	0.4424	0.4771
32	14	FI-CODE	2	0.4678	0.4624	0.4902	0.4734
33		INSA Lyon	3	0.4351	0.5107	0.4734	0.4731
34		INSA Lyon	2	0.4435	0.5128	0.4561	0.4708
35		FI-CODE	3	0.4378	0.4466	0.5091	0.4645
36		VerbaNexAI I	1	0.4488	0.4757	0.4640	0.4628
37	15	ROSTI	3	0.5222	0.4541	0.3930	0.4564
38		ROSTI	2	0.5192	0.4431	0.3896	0.4507
39		UMUTEAM	3	0.4527	0.4312	0.4645	0.4495
40		ROSTI	1	0.5196	0.4335	0.3849	0.4460
41		BIU_NLP	1	0.4495	0.4583	0.4208	0.4429
42		UMUTEAM	1	0.4167	0.4702	0.4357	0.4408
43		FI-CODE	1	0.4162	0.4313	0.4333	0.4270
44		MILRIT	2	0.4365	0.4353	0.4073	0.4264
45	16	FourBytes	1	0.3720	0.4187	0.4275	0.4061
46		Random baseline	1	0.4058	0.3733	0.4355	0.4049

Table 8
Best language-specific Test A runs.

Lang.	Best run	Score	`at` MR	`isAt` MR	Runner-up
de	Spinfo run 3	0.771	0.703	0.839	Spinfo run 1, 0.761
en	MaxFo-Ajie run 1	0.749	0.742	0.756	Spinfo run 1, 0.728
fr	Spinfo run 1	0.755	0.679	0.832	Spinfo run 2, 0.738

6.2. Generalization Profile on Test B

Table 9 reports the best runs on the surprise French literary test set. This profile evaluates only the `at` relation and is intentionally separate from Test A, because it changes both domain and time period.

MaxFo-Ajie achieved the strongest generalization result: its three submitted runs occupy the first three ranks in the full ranking, with run 3 reaching 0.816 macro recall. This contrasts with Test A, where Spinfo was strongest overall. Neither team's Test A strategy straightforwardly explains its Test B result: MaxFo-Ajie's few-shot prompting approach (Section 5) appears to transfer particularly well to the literary and historiographical genre of Test B, while Spinfo's strength on Test A does not carry over with the same margin. This suggests that in-domain newspaper performance and robustness to the literary surprise domain are related but partially distinct capabilities, and that a system's ranking on one profile is at best a weak predictor of its ranking on the other.

6.3. Efficiency Profile

Table 10 reports the efficiency profile, which combines each run's accuracy rank, parameter-count rank, and model-size rank by their arithmetic mean.

The efficiency profile changes the interpretation of the leaderboard. MILRIT run 3 ranks first under the efficiency metric, despite not being the highest-accuracy run. FI-CODE and DS@GT_HIPE also rank highly because their resource ranks offset lower accuracy. The result of FI-CODE is particularly informative as a baseline assessment: its heuristic classifies pairs based on the textual distance between person and location mentions (the same proximity measure examined in Section 7.4.3). While this strategy is extremely efficient, its moderate accuracy shows that mention proximity alone is not a strong enough signal to solve the task, especially because the data also contains many long-distance relations between persons and places. Conversely, the most accurate Test A system, Spinfo, ranks lower in the official efficiency profile because of its large reported parameter count and model size.[10] This confirms the usefulness of reporting efficiency separately from accuracy for large-scale historical-text processing.

6.4. Baseline Results

On Test A, the Ministral baseline scored 0.582, clearly above the random baseline score of 0.405, but below the best participant run by a large margin. On Test B, the same baseline scored 0.506, again above random (0.363) but far behind the best surprise-domain run (0.816). These results make the baseline a reasonable reference point for zero-shot LLM prompting, while showing that task-specific system design and model size matter.

7. In-Depth Analyses and Error Diagnostics

Beyond the official rankings, this section examines what drives system behavior in more detail: the added difficulty of the `PROBABLE` label (Section 7.1), qualitative error patterns on the surprise set, how teams used the released silver data, and whether accuracy varies systematically with four data properties external to any single system's design — OCR quality, publication period, person–place

[10] A similar outcome would likely also hold for the first-ranked team in the generalization profile (MaxFo-Ajie) as it is based on a large proprietary LLM (GPT-5.5). However, the team requested to opt out from the efficiency ranking.

Table 9
Generalization Profile ranking on Test B. The official score is `at` macro recall (MR) on the surprise-domain French test set. Accuracy is reported as an additional diagnostic and is not used for ranking.

Rank	T-Rank	Team	Run	MR ↑	Acc.
1	1	MaxFo-Ajie	3	0.8163	0.8729
2		MaxFo-Ajie	1	0.7945	0.8625
3		MaxFo-Ajie	2	0.7712	0.8542
4	2	Spinfo	3	0.6984	0.8104
5		Spinfo	1	0.6910	0.7729
6	3	BIU_NLP	1	0.6837	0.8354
7		Spinfo	2	0.6674	0.7458
8	4	whereami	2	0.6665	0.8333
9	5	gipplab	2	0.6647	0.7458
10	6	Awakened	3	0.6613	0.7292
11	7	Hansel&Gretel	2	0.6349	0.6438
12		Awakened	1	0.6338	0.7250
13		whereami	1	0.6325	0.8167
14		Hansel&Gretel	3	0.6187	0.7708
15		Hansel&Gretel	1	0.6107	0.6896
16		gipplab	1	0.6085	0.7500
17		BIU_NLP	3	0.6000	0.6042
18	8	VerbaNexAI II	3	0.5724	0.7708
19	9	UMUTEAM	2	0.5723	0.5750
20		Awakened	2	0.5509	0.5625
21		gipplab	3	0.5382	0.6375
22		BIU_NLP	2	0.5265	0.4729
23	10	MILRIT	3	0.5152	0.5646
24		VerbaNexAI II	2	0.5076	0.6937
25		Ministral baseline	1	0.5062	0.5583
26	11	INSA Lyon	1	0.4705	0.6604
27		MILRIT	1	0.4679	0.6583
28		VerbaNexAI II	1	0.4419	0.6292
29		INSA Lyon	2	0.4231	0.6438
30		INSA Lyon	3	0.3986	0.6375
31	12	DS@GT_HIPE	3	0.3919	0.5000
32	13	ROSTI	3	0.3840	0.5104
33		ROSTI	1	0.3773	0.5062
34	14	FI-CODE	2	0.3755	0.3729
35		MILRIT	2	0.3742	0.5750
36	15	VerbaNexAI I	2	0.3726	0.4542
37		DS@GT_HIPE	2	0.3721	0.5000
38		ROSTI	2	0.3660	0.4938
39		Random baseline	1	0.3628	0.3604
40		DS@GT_HIPE	1	0.3626	0.3708
41		UMUTEAM	3	0.3620	0.3104
42		FI-CODE	1	0.3580	0.4437
43		FI-CODE	3	0.3546	0.4813
44	16	FourBytes	1	0.3445	0.2667
45		VerbaNexAI I	1	0.3346	0.5375
46		UMUTEAM	1	0.3333	0.6042

proximity, and Wikidata entity-linking status. These are diagnostic analyses, not part of the official ranking; we flag where a pattern is not robust across systems, domains, or label distributions.

7.1. The Role of `PROBABLE`

The official ternary setup evaluates `PROBABLE` as a separate label for `at`. This makes the task more demanding than binary relation detection: systems must distinguish explicit evidence from plausible inference, a fine boundary that can even be slightly fuzzy in some cases. Indeed, we find that `PROBABLE`

Table 10
Efficiency Profile ranking on Test A. The efficiency score is the mean of three component ranks: accuracy-profile score, parameter count, and model size. The final column reports the official Test A accuracy-profile score for reference. The table contains 43 runs rather than the 46 runs in the Accuracy Profile because the three MaxFo-Ajie runs opted out of the efficiency ranking. Tied overall ranks are shown only on their first row and ordered by decreasing accuracy-profile score.

Ranking		Submission		Efficiency	Component ranks			Acc.
Rank	T-Rank	Team Name	Run	Eff. ↓	Acc.	Param.	Size	Score ↑
1	1	MILRIT	3	9.7	12	8	9	0.5951
2	2	FI-CODE	2	10.3	29	1	1	0.4734
3	3	DS@GT_HIPE	1	10.7	23	5	4	0.5142
4		DS@GT_HIPE	2	12.0	27	5	4	0.4836
5		DS@GT_HIPE	3	12.3	28	5	4	0.4771
6		MILRIT	1	13.7	18	12	11	0.5623
	4	ROSTI	3	13.7	34	4	3	0.4564
		ROSTI	2	13.7	35	3	3	0.4507
		ROSTI	1	13.7	37	2	2	0.4460
7		Random baseline	1	15.0	43	1	1	0.4049
8	5	Awakened	2	15.3	19	14	13	0.5494
9		Ministral baseline	1	15.7	14	19	14	0.5818
	6	VerbaNexAI I	2	15.7	26	11	10	0.4842
10	7	INSA Lyon	2	16.0	31	9	8	0.4708
11	8	whereami	1	16.7	4	22	24	0.6880
12		whereami	2	17.0	5	22	24	0.6833
	9	VerbaNexAI II	3	17.0	15	20	16	0.5795
		FI-CODE	1	17.0	40	6	5	0.4270
13	10	UMUTEAM	1	17.3	39	7	6	0.4408
14	11	gipplab	2	17.7	9	21	23	0.6271
15		VerbaNexAI I	1	18.0	33	11	10	0.4628
16		UMUTEAM	2	18.3	13	20	22	0.5856
17		VerbaNexAI II	1	18.7	25	16	15	0.5004
18		gipplab	3	19.7	24	17	18	0.5069
	12	FourBytes	1	19.7	42	10	7	0.4061
19	13	Hansel&Gretel	1	20.0	20	19	21	0.5458
20	14	Spinfo	1	20.3	1	30	30	0.7479
21		Spinfo	3	20.7	2	30	30	0.7289
		gipplab	1	20.7	11	25	26	0.6141
		INSA Lyon	3	20.7	30	15	17	0.4731
22		Spinfo	2	21.0	3	30	30	0.6890
23		Hansel&Gretel	2	21.7	16	24	25	0.5788
		VerbaNexAI II	2	21.7	22	23	20	0.5187
24		MILRIT	2	22.0	41	13	12	0.4264
25		INSA Lyon	1	22.7	8	29	31	0.6390
26		FI-CODE	3	23.0	32	18	19	0.4645
27		Awakened	3	23.7	6	32	33	0.6671
28		Awakened	1	24.0	7	32	33	0.6584
29		Hansel&Gretel	3	24.3	10	31	32	0.6221
30	15	BIU_NLP	2	24.7	17	28	29	0.5781
		BIU_NLP	3	24.7	21	26	27	0.5390
31		UMUTEAM	3	26.0	36	20	22	0.4495
32		BIU_NLP	1	31.0	38	27	28	0.4429

is often the hardest label. For example, on German Test A, Spinfo run 1 obtained high `FALSE` recall (0.952) but a lower `PROBABLE` recall (0.579). The Ministral baseline in this case reached only 0.184 recall for `PROBABLE`.

Table 11 summarizes this diagnostic evaluation. In the collapsed setting, the top Test A score rises from 0.748 to 0.842, and the ordering of the strongest systems changes after the first position. The gain is therefore not merely a uniform score inflation: it also shows that systems differ in how well they

Table 11
Binary sensitivity analysis for Test A. In this diagnostic setting, `PROBABLE` is collapsed into `TRUE` for `at`; `isAt` is unchanged. The table reports the ten highest-ranked complete Test A runs under the binary setting and compares them with their official ternary Test A scores.

Rank	Team	Run	Binary score	Official score	Δ
1	Spinfo	1	0.842	0.748	+0.094
2	Spinfo	3	0.837	0.729	+0.108
3	whereami	2	0.816	0.683	+0.132
4	whereami	1	0.815	0.688	+0.127
5	Spinfo	2	0.800	0.689	+0.111
6	Awakened	3	0.793	0.667	+0.125
7	MaxFo-Ajie	1	0.788	0.700	+0.088
8	Awakened	1	0.786	0.658	+0.128
9	MaxFo-Ajie	2	0.770	0.669	+0.101
10	INSA Lyon	1	0.763	0.639	+0.124
–	Ministral baseline	1	0.682	0.582	+0.100
–	Random baseline	1	0.491	0.405	+0.086

separate direct evidence from plausible but uncertain inference.

While some downstream applications may prefer this collapsed binary setting, the official ternary setup is more informative for evaluating whether systems can distinguish direct evidence from plausible inference. We therefore treat binary `at` results as a useful sensitivity analysis, but not as a replacement for the official evidence-sensitive ternary task.

7.2. System Behavior on the Surprise Set

Figure 5 includes, alongside the gold annotations, the distribution of predictions submitted by all participating systems, offering a window into collective system behavior on the surprise-domain set. The **Tenaud**–*Scythie* pair (gold: `FALSE`) was correctly handled by most systems (79.5%), confirming that absent locative evidence for a named authority is generally well detected. More revealing are the error patterns. The **Gargantua**–*Paris* pair (gold: `FALSE`) attracted 27.3% `PROBABLE` predictions, suggesting that directional movement toward a destination is a systematic source of confusion. The **Grandgousier**–*port de Olone* pair (gold: `PROBABLE`) was assigned `TRUE` by a majority of systems (59.1%), confirming the tendency observed elsewhere to resolve locative uncertainty toward the more confident label. Finally, the **Ponocrates** and **Eudemon**–*Orléans* pairs (gold: `TRUE`) show that inferring physical passage through a location from narrative context is far from trivial, with 38.6% of systems assigning `FALSE`.

These aggregate patterns already hint that the hardest cases are those where locative status must be inferred rather than read off the surface of the text; to understand what drives system errors more precisely, we turn from the distribution of predictions to the individual pairs on which systems' predictions most strongly diverged from the gold standard.

We manually inspect the five cases where all top-5 system runs (best run per team) unanimously disagreed with the gold label, in order to characterize the nature of these residual errors. The disagreement in these cases is not confined to the strongest systems: across all 46 submitted runs, the majority likewise diverged from the gold label, with an average disagreement of 66% per pair. This broad-based divergence suggests that these cases reflect properties of the data and annotation rather than idiosyncratic weaknesses of individual systems. The five cases reduce to three main underlying sources of difficulty.

The first concerns document segmentation and entity resolution errors upstream of the relation itself. In a document drawn from Bussy's *Lettres* (1666), two separate letters had been mistakenly concatenated into a single test document, and two distinct entity mentions—"Chevalier De Rohan" and "Chevalier De G", each belonging to a different letter—were conflated into a single entity. This artefact

confounded human annotators as much as systems: the two annotators disagreed on the gold label (TRUE vs. FALSE) and adjudication was required, ultimately resolving to FALSE.

The second source of difficulty is genuine referential ambiguity in the source text. In the same document, a recurring formula, "Réponse du Comte De Bussy à Madame De S. à Chaseu, ce 1 septembre 1675," was unanimously misread by all systems as indicating that the letter was addressed to a woman residing in Chaseu, whereas "à Chaseu" in fact specifies the location from which the letter's author was writing. The two human annotators also disagreed on this case (BLANK vs. FALSE), and the adjudicator was only able to confirm the gold FALSE label after consulting several additional documents. We note that this ambiguity is plausibly an artifact of transcription: in correspondence of this period, a line break before "à Chaseu" would typically disambiguate the two readings, and its absence in the digitized text may have removed a cue available to a human reader of the original.

The third source of difficulty concerns the PROBABLE label itself, and accounts for the remaining two cases. In Voltaire's *Histoire Générale* (1756), two distinct relations drawn from the same passage were each annotated with disagreement between annotators (PROBABLE vs. FALSE); adjudication settled on PROBABLE. Closer inspection suggests that the source-document evidence in these cases may support the systems' dissenting FALSE predictions better than the adjudicated label, flagging these relations for re-annotation. More broadly, these cases illustrate that the boundary between absence of evidence and plausible inference is one of the most delicate aspects of the annotation scheme.

To test whether these observations generalize beyond the strictest criterion, we relaxed the threshold to entity pairs on which at least half of all submitted runs, and four of the five top-5 runs, disagreed with the gold label. This surfaces nine additional pairs, distributed over three documents (two of which overlap with the high-disagreement cases discussed above). Crucially, these pairs instantiate the same three sources of difficulty rather than introducing new ones: entity over-segmentation, syntactic and referential ambiguity, and edge cases of the probable label. Several involve evidence that is only indirectly recoverable from the text—requiring a chain of inference across several statements—or genuine narrative complexity, and a number were subsequently flagged for re-annotation. The recurrence of the same categories under a relaxed criterion reinforces the conclusion that broad system–gold divergence tracks properties of the data and annotation rather than the capabilities of individual systems.

Taken together, these cases are reassuring for the shared task: the residual unanimous disagreements stem largely from data artifacts (mis-segmented documents, conflated entities, lost transcription cues) and edge cases in the annotation scheme, rather than from systematic failures of the participating systems to reason over locative evidence.

7.3. Utilization and Effects of Silver Data

Teams used the released silver data in markedly different ways. Several high-ranking systems, including MILRIT, MaxFo-Ajie, and Spinfo, do not appear to have relied on the released silver annotations as a central component. Other teams used the silver data primarily for model development, prompt selection, or auxiliary supervision; examples include UMUTEAM, BIU_NLP, and Hansel&Gretel. A smaller number of submissions incorporated silver data more directly during training, such as Awakened's XLM-RoBERTa-large fine-tuning run. In addition, whereami generated its own silver-style data using a Gemma model, rather than relying only on the released silver annotations.

The official results do not allow a clear causal conclusion about the value of the released silver data. Some systems using silver data performed well, but several of the strongest systems did not make substantial use of it. One likely explanation is that the benefit of additional training instances was partly offset by annotation noise or by systematic differences between LLM-generated and human annotations. This interpretation is consistent with Table 5, where LLM annotations assign PROBABLE more often and TRUE less often than human annotations on overlapping documents. Future work should therefore treat silver annotations as useful but noisy supervision, and should investigate filtering, calibration, or confidence-weighting strategies before using them as direct training targets.

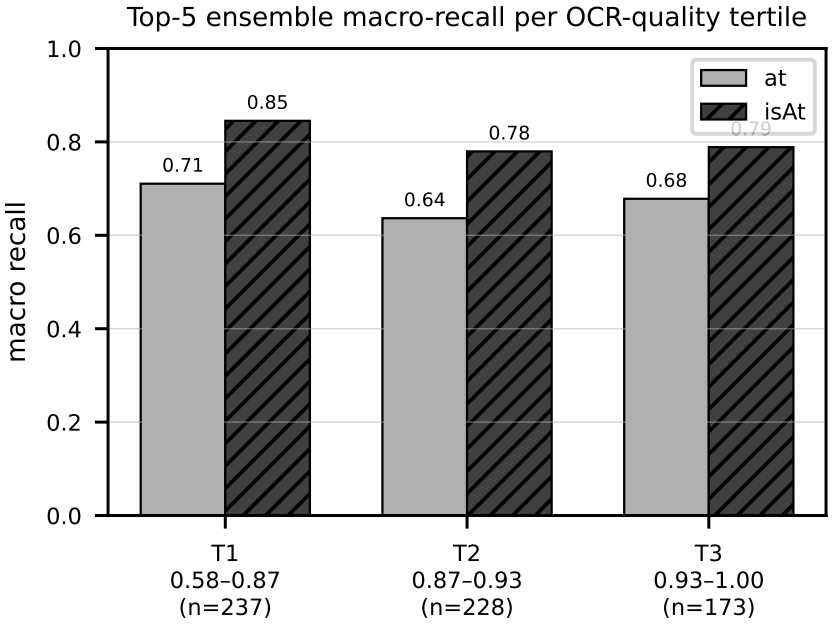

Figure 7: Top-5 ensemble macro recall by OCR-quality tertile (Domain A only). The ordering is not statistically significant and does not replicate across individual systems (see text).

7.4. Understanding the Impact of OCR Quality, Document Release Date, Textual Proximity, and Background Knowledge

For these four analyis perspectives, we are using a *top-5 ensemble* (plurality vote over the five highest-ranked runs on the overall Test A ranking) as a representative and strong system. Given these predictions, we stratify the data in different ways, and compute relative macro Recalls to investigate the role of i) OCR quality, ii) document release date iii) textual proximity of person and place, and iv) the potential impact of background knowledge.

7.4.1. Impact of OCR Quality

OCR quality was quantified per document using the `impresso_pipelines` OCR quality-assessment module,[11] which scores each document's full text against a language-specific lexicon (a proportion of known-word types, in $[0, 1]$, with 1 indicating text free of detectable OCR-induced lexical noise), computed separately per document using its own language field. Documents were then split into equal-sized tertiles (T1: lowest quality, T3: highest) by this score, restricted to Domain A.

We find no significant OCR-quality effect on top-5 ensemble accuracy: the document-level Spearman correlation between OCR score and per-document accuracy is negative and non-significant for both tasks ($\rho = -0.148$, $p = 0.27$ for `at`; $\rho = -0.178$, $p = 0.19$ for `isAt`; $n = 57$ documents) — this should be read as *not detected*, not as evidence against any effect. The tertile-level picture (Figure 7, T1/T2/T3 macro recall) is more cautionary than informative. First, the T1>T3>T2 ordering the ensemble shows (`at`: 0.71/0.64/0.68) is not stable across systems on the identical 638 pairs: the rank-1 individual system (Spinfo) shows a monotonic decrease with no dip (0.71/0.69/0.68), while the Ministral baseline shows an inverted U with T2 as the peak (0.51/0.54/0.51) — three different orderings on the same data. If the tertiles reflected a shared, document-level difficulty signal, the ordering should replicate (at different magnitudes) across systems; instead it points to each system's own idiosyncratic errors on a small set of hard pairs dominating the comparison rather than a common OCR-quality effect. The `isAt` task is messier still: Spinfo reproduces the ensemble's U-shape, but the baseline shows a clean monotonic

[11] https://pypi.org/project/impresso-pipelines/

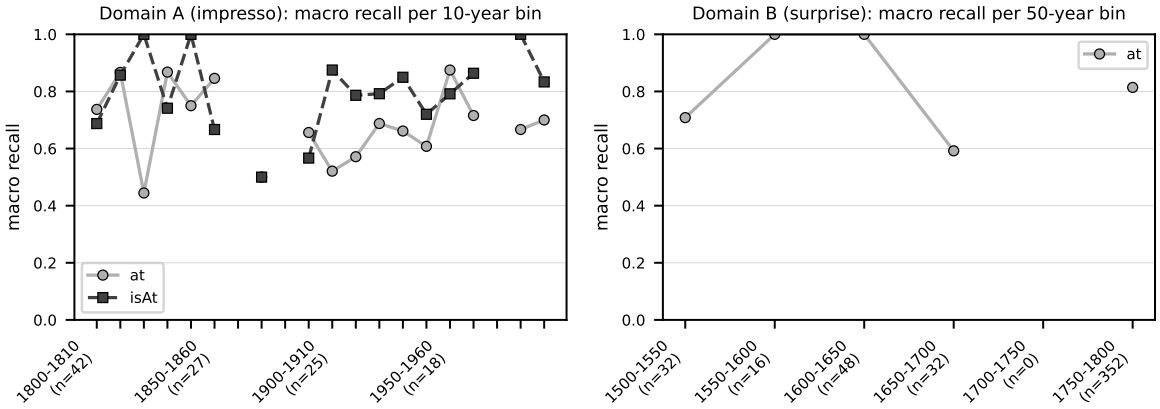

Figure 8: Top-5 ensemble macro recall by 10-year (Domain A) and 50-year (Domain B) bin. Many bins rest on 1–2 documents (Domain A) or 1–2 literary sources (Domain B); see text for the resulting caveats.

increase — the one result in the whole analysis matching the naive "better OCR → better accuracy" expectation — which is itself a reason not to cherry-pick any single system's pattern as confirmation.

Second, and independently, document composition is not comparable across tertiles: German has zero documents in the best-OCR tertile and English has almost none in the worst, so the pooled comparison conflates OCR quality with language/source mix. Restricting to French alone (the only language with real n in all three tertiles) still shows a U-shape (0.70/0.53/0.60), so the pattern is not purely a language-swap artifact — but this also means the pooled tertile comparison does not isolate OCR quality either. Compounding both issues, some tertile-level `PROBABLE` cells are as small as $n = 5$, making single-tertile macro recall highly sensitive to a handful of examples. Overall, the data do not support claiming an OCR-quality effect in either direction: not "no effect" — that would itself overclaim given the confounds — but that any apparent effect is not isolable from language/source composition and not stable across systems at this sample size. The ensemble's T1>T3>T2 ordering should not be presented as a finding; it is neither statistically significant nor system-robust.

7.4.2. Impact of Time Period

Publication year was parsed directly from each document's date metadata and documents were grouped into fixed 10-year (Domain A) and 50-year (Domain B) windows, chosen to keep per-bin document counts as large as the respective domain's size allows.

We find no significant relationship between publication year and top-5 ensemble accuracy in either domain. Document-level Spearman correlations are all non-significant at $p < 0.05$: Domain A $\rho = -0.232$, $p = 0.088$ for `at` and $\rho = -0.173$, $p = 0.206$ for `isAt` ($n = 55$); Domain B $\rho = -0.012$, $p = 0.949$ ($n = 30$, `at` only). Domain A's weak trend is, if anything, in the wrong direction for the naive "older is harder" story — it points toward newer documents being marginally more difficult — and does not survive being one of several correlations examined without correction. The bucketed picture (Figure 8) is not a usable substitute: 8 of Domain A's 17 non-empty bins rest on only 1–2 documents, so a bin's macro recall there reflects that document's idiosyncratic difficulty rather than a period effect. Domain B is worse in a way raw document counts obscure: stripping chunk-level identifiers back to source works, its entire 1500–1800 span reduces to 9 distinct literary/historiographical sources, with most bins covered by 1–2 works and the largest bin (1750–1800, 73% of Domain B's pairs) covered by only two books; any apparent Domain B "period" pattern is statistically indistinguishable from an author/style effect. Recomputing the same bins directly from individual systems' diagnostics (Spinfo, rank-1 overall, and the Ministral baseline, no ensemble voting) shows bin-to-bin direction agreeing across all three systems in only 7 of 16 Domain A transitions (44%), close to chance; Domain B's transitions are similarly unstable and confounded with the book-level fragility above. We do not find

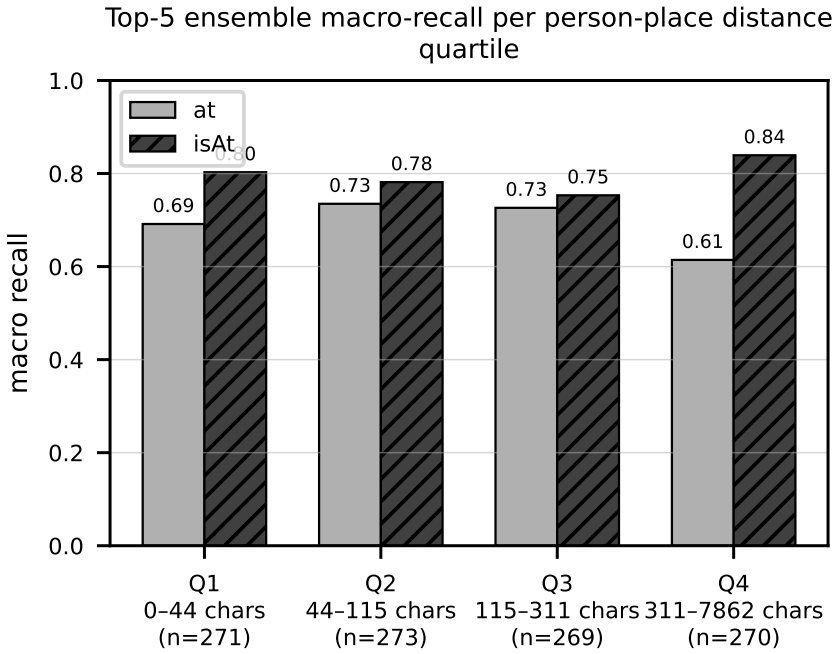

Figure 9: Top-5 ensemble macro recall by person–place text-distance quartile (pooled across Domain A and B). The `isAt` increase at Q4 is a class-imbalance artifact, not a genuine long-range strength (see text).

evidence supporting a time-period effect on accuracy in either domain at this sample size, and flag Domain B in particular as underpowered at the level of independent sources to support any period claim.

7.4.3. Impact of Person–Place Mention Proximity

Since the released schema does not include mention character offsets, person–place distance was recovered per candidate pair by locating each mention string in the document's full text via a word-boundary-aware substring search, and taking the minimum character gap between any occurrence of a person mention and any occurrence of a location mention belonging to that pair; pairs whose mention strings could not be located verbatim (3.1% overall, concentrated in a small number of low-OCR-quality French documents, see below) were excluded rather than imputed. Pairs were then split into distance quartiles, pooled across Domain A and B.

Top-5 ensemble macro recall for `at` declines with person–place text distance (Figure 9, Q1–Q4 quartiles: 0.71/0.73/0.73/0.61), while `isAt` shows an apparent increase in the largest-distance quartile (0.80/0.78/0.75/0.84). Overall, any decreases or increases seem more influenced by outliers due to small bin sizes and are not substantial. This suggests that long-range and short range relationships were about similarly difficult to resolve. More interesting perhaps is the observation that closer pairs are more likely to be in a `at`-relation than longer-range pairs, a fact that has been exploited by FI-Code in building a modestly accurate but maximally efficient baseline system just based on thresholding distances (we refer the reader to their paper [30] for more investigation of this aspect). Still, the only modest accuracy of this baseline suggest that the distance of relationships is not a give-away of the relation, and further reasoning is required in most cases.

7.4.4. Impact of Wikidata Entity-Linking Status

QID-linkage status was read directly from each entity's Wikidata identifier field in the reference data: an entity was treated as linked if this field was non-null, and as NIL otherwise, for the person and location entity in each candidate pair independently.

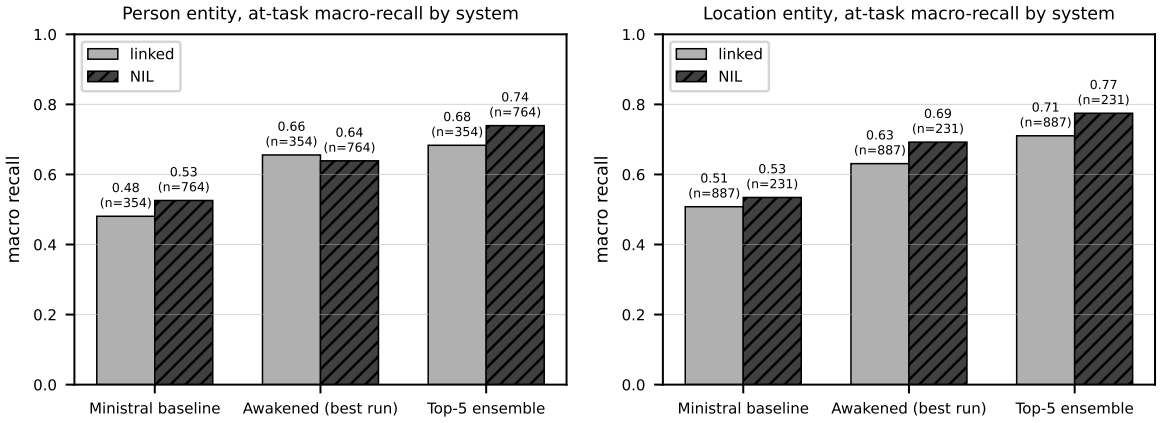

Figure 10: `at`-task macro recall by QID-linkage status (linked vs. NIL), person and location entities, across three systems. The location-entity gap replicates across systems and domains and survives holding gold label constant; the person-entity gap does not replicate for Awakened (see text).

Across all three systems examined (the Ministral baseline, Awakened's best run, and the top-5 ensemble), pairs with a NIL (unlinked) location entity score higher macro recall on `at` than QID-linked ones, by 0.026–0.064 depending on system (Figure 10). Gold-label distribution does not explain this away: while NIL location pairs are modestly more often gold-`FALSE` than linked ones (pooled 64.1% vs. 54.2%), recomputing macro recall *within* the harder `TRUE`/`PROBABLE` subset alone — holding gold label fixed — still shows NIL outperforming linked in 5 of 6 system×entity-type cells, with gap sizes close to the original pooled figures; the effect is not a label-mix artifact. The pattern replicates independently within both domains for location (impresso and surprise separately), and QID-linkage rate for location is stable across domains (79.9% vs. 78.5%), ruling out a domain-composition confound there. A mention-count check further rules out "NIL entities simply have less textual evidence": linked entities carry *more* listed mention strings than NIL ones on average, not fewer. The person-entity result is directionally similar (NIL again outperforming linked for the baseline and the ensemble, gaps of $+0.045$ and $+0.056$) but is not system-robust: Awakened shows the reverse (-0.017), and the person-entity pattern leans on the surprise domain, where linked person pairs form a thin subsample ($n = 80$). Person-entity QID-linkage rate also differs sharply by domain (42.9% impresso vs. 16.7% surprise), unlike location. We therefore treat the location-entity result as the more defensible finding — a genuine, modest reasoning-difficulty pattern in which linked entities are handled somewhat worse than unlinked ones.

The mechanism behind this pattern remains uncertain. One possibility is that linked entities are more frequent or better known, and therefore more likely to activate background knowledge that conflicts with the local evidence in the document. In such cases, systems may rely on general knowledge about an entity rather than on the specific person–place relation expressed, or not expressed, in the source text. This interpretation remains speculative, however, and would require targeted analysis of retrieved or model-internal knowledge to confirm.

7.5. Synthesis: What Drives Difficulty?

Across these analyses, difficulty is not well explained by surface data properties. OCR quality and publication period show no effect that survives scrutiny; person–place proximity has a modest, system-replicable effect on `at`, but distance alone is a poor predictor and leaves the task far from solved; and Wikidata linkage, if anything, makes location relations somewhat *harder* rather than easier. What consistently tracks difficulty instead is the evidentiary distinction between explicit and inferred presence: `PROBABLE` is the hardest label throughout, and the surprise-set error analysis shows the same pattern at the level of individual cases. Notably, the residual disagreements on the surprise set trace largely to data artifacts and annotation edge cases rather than to systems failing to reason over the unfamiliar

literary and historiographical material itself — suggesting that the abductive reasoning skills systems bring to Domain A carry over to Domain B reasonably well, even though overall accuracy on the two domains does not rank the same systems in the same order. We read this as evidence that HIPE-2026's central challenge is abductive reasoning over context, not signal quality, corpus composition, or domain familiarity.

8. Conclusion

HIPE-2026 extends the HIPE evaluation series from named entity recognition and linking toward document-level relation understanding in historical text. The shared task focuses on temporally grounded person–place relation extraction, requiring systems to determine not only whether a person and a location are related, but also whether the evidence supports current presence near the publication date. This formulation moves beyond entity co-occurrence and challenges systems to perform contextual, temporal, and evidence-sensitive reasoning in noisy historical documents.

The results show that the task is challenging but tractable under current modeling approaches. On the historical newspaper benchmark (Domain A), the strongest systems substantially outperformed the organizer baselines, with the best run achieving a Test A score of 0.748 compared to 0.582 for the Ministral baseline and 0.405 for the random baseline. At the same time, the leaderboard remained highly competitive, with multiple teams achieving strong performance through markedly different methodological choices. The submissions ranged from prompted large language models and parameter-efficient fine-tuning to multilingual encoder architectures, graph-based methods, feature-engineered classifiers, and lightweight rule-based systems.

Several observations emerge from the evaluation. First, language-specific performance patterns indicate that the two target relations pose different challenges. High `isAt` recall was particularly important for strong performance in German and French, whereas the best English systems achieved a more balanced trade-off between `at` and `isAt`. Second, the surprise-domain (Domain B) evaluation confirms that in-domain accuracy and cross-domain robustness are somewhat similar but also markedly distinct capabilities. While Spinfo achieved the highest overall score on the historical newspaper benchmark, MaxFo-Ajie dominated the literary surprise set (and vice versa), demonstrating that successful transfer to substantially different genres and historical periods is an intricate challenge.

The ternary formulation of `at` proved to be an important aspect of the task design. Distinguishing between `TRUE` and `PROBABLE` requires systems to separate explicit textual evidence from plausible contextual inference, making the task substantially more demanding than binary relation detection. The binary sensitivity analysis showed that collapsing `PROBABLE` into `TRUE` increases the best Test A score from 0.748 to 0.842 and changes the ordering of several leading systems. This confirms that modeling uncertainty and abductive inference is a central challenge of historical person–place relation extraction rather than a marginal edge case.

The diversity of successful approaches is equally noteworthy. Many of the highest-ranked systems relied on large language models, either through prompting or supervised adaptation, suggesting that modern generative models are effective at integrating dispersed contextual and temporal evidence. However, competitive results were also obtained with encoder-based architectures and feature-driven approaches, indicating that strong performance does not depend on a single modeling paradigm — reasoning-oriented and classification-oriented approaches are both viable routes to good accuracy on this task. The task therefore provides a useful testbed for comparing reasoning-oriented and classification-oriented approaches under a common evaluation framework.

The efficiency profile further highlights the importance of considering computational cost alongside predictive quality. Systems with the highest accuracy often relied on comparatively large models, whereas methods based on compact encoders, engineered features, or lightweight classifiers achieved substantially better efficiency rankings. In particular, the strong efficiency results of MILRIT, FI-CODE, and DS@GT demonstrate that practical large-scale deployment on cultural-heritage collections may favor different solutions than those that maximize accuracy alone. Reporting both dimensions therefore

provides a more complete picture of system suitability for real-world historical-text processing.

The extended diagnostic analyses further suggest that the main sources of difficulty cannot be explained by simple surface properties. OCR quality and publication period did not show robust effects on performance, while person–place proximity had a somewhat modest effect for `at`. The Wikidata-linking analysis further suggests that linked entities are not necessarily easier, especially for locations. Taken together, these findings reinforce the view that the task is driven primarily by contextual, temporal, and abductive interpretation and escapes prediction by simple metadata or shallow surface heuristics.

Overall, HIPE-2026 establishes a benchmark for evidence-sensitive and temporally grounded relation extraction in historical documents. The results show clear progress beyond entity recognition while also revealing persistent challenges involving temporal interpretation, implicit evidence, uncertainty calibration, and domain transfer. Future work should investigate the remaining sources of error in temporal interpretation and plausible inference, extend the benchmark to additional relation types and historical genres, and explore approaches that improve cross-domain robustness without sacrificing computational efficiency.

Returning to the motivating question — *who was where, and when* — HIPE-2026's results suggest a qualified answer. Current systems can already extract many explicit and strongly implied person–place relations from noisy historical text at useful accuracy, making large-scale automatic construction of biographical trajectories a realistic near-term prospect for applications that can tolerate a meaningful error rate or benefit from human review, such as populating draft knowledge graphs or flagging candidate itineraries for further archival research. What remains substantially unsolved is the harder half of the question: reliably telling inferred presence from mere plausibility, and distinguishing a person's past from their present location, are precisely the cases where systems still fail most often. Digital humanities applications that require high-confidence, individual-level trajectories — rather than aggregate or exploratory analysis — should therefore treat current system output as a starting point for scholarly verification, not as a finished answer.

Acknowledgments

The CLEF-HIPE-2026 organizing team thanks the CLEF 2026 Conference and Evaluation Labs Committees for hosting the task. This work is carried out within the framework of the Impresso – Media Monitoring of the Past project, funded by the Swiss National Science Foundation under grant No. CRSII5_213585 and by the Luxembourg National Research Fund under grant No. 17498891.

Declaration on Generative AI

The authors used GenAI tools to assist with writing, polishing, fluency and grammar checks. The authors reviewed and edited all AI-assisted text and take full responsibility for the publication's content.

References

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

[2] D. van Strien, K. Beelen, M. Ardanuy, K. Hosseini, B. McGillivray, G. Colavizza, Assessing the Impact of OCR Quality on Downstream NLP Tasks:, in: Proc. of the 12th International Conference on Agents and Artificial Intelligence, SCITEPRESS, Valletta, Malta, 2020, pp. 484–496. URL: http://www.scitepress.org/DigitalLibrary/Link.aspx?doi=10.5220/0009169004840496.

[3] A. Hamdi, A. Jean-Caurant, N. Sidère, M. Coustaty, A. Doucet, Assessing and Minimizing the Impact of OCR Quality on Named Entity Recognition, in: Digital Libraries for Open Knowledge, Lecture Notes in Computer Science, Springer International Publishing, Cham, 2020, pp. 87–101. doi:10.1007/978-3-030-54956-5_7.

[4] M. Ehrmann, A. Hamdi, E. L. Pontes, M. Romanello, A. Doucet, Named Entity Recognition and Classification in Historical Documents: A Survey, ACM Computing Surveys 56 (2023) 27:1–27:47. URL: https://dl.acm.org/doi/10.1145/3604931.

[5] A. Fokkens, S. Ter Braake, N. Ockeloen, P. Vossen, S. Legêne, G. Schreiber, et al., BiographyNet: Methodological issues when NLP supports historical research., in: LREC, 2014, pp. 3728–3735.

[6] M. Schich, C. Song, Y.-Y. Ahn, A. Mirsky, M. Martino, A.-L. Barabási, D. Helbing, A network framework of cultural history, Science 345 (2014) 558–562. doi:10.1126/science.1240064.

[7] J. Opitz, L. Born, V. Nastase, Y. Pultar, Automatic reconstruction of emperor itineraries from the regesta imperii, in: Proceedings of the 3rd International Conference on Digital Access to Textual Cultural Heritage, 2019, pp. 39–44.

[8] L. Lucchini, R. Sinatra, C. Emery, P. Panzarasa, V. D. P. Servedio, M. Riccaboni, C. Cattuto, Following the footsteps of giants: Modeling the mobility of historically notable individuals, EPJ Data Science 8 (2019). doi:10.1140/epjds/s13688-019-0215-7.

[9] S. D. Cardoso, M. Da Silveira, C. Pruski, Construction and exploitation of an historical knowledge graph to deal with the evolution of ontologies, Knowledge-Based Systems 194 (2020) 105508.

[10] M. Tamper, M. Kettunen, E. Mäkelä, T. Ruotsalo, P. Leskinen, E. Hyvönen, BiographySampo: A linked open data service for prosopographical biography research, in: Proceedings of the Digital Humanities Conference (DH 2023), Graz, Austria, 2023. URL: https://seco.cs.aalto.fi/projects/biographysampo/.

[11] L. Zhong, J. Wu, Q. Li, H. Peng, X. Wu, A comprehensive survey on automatic knowledge graph construction, ACM Computing Surveys 56 (2023) 1–62.

[12] M. Ehrmann, M. Romanello, A. Flückiger, S. Clematide, Overview of CLEF HIPE 2020: Named Entity Recognition and Linking on Historical Newspapers, in: A. Arampatzis, E. Kanoulas, T. Tsikrika, S. Vrochidis, H. Joho, C. Lioma, C. Eickhoff, A. Névéol, L. Cappellato, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction, Lecture Notes in Computer Science, Springer International Publishing, Cham, 2020, pp. 288–310. doi:10.1007/978-3-030-58219-7_21.

[13] M. Ehrmann, M. Romanello, A. Flückiger, S. Clematide, Extended Overview of CLEF HIPE 2020: Named Entity Processing on Historical Newspapers, in: L. Cappellato, C. Eickhoff, N. Ferro, A. Névéol (Eds.), Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum, volume 2696, CEUR-WS, Thessaloniki, Greece, 2020, p. 38. URL: https://infoscience.epfl.ch/record/281054.

[14] M. Ehrmann, M. Romanello, S. Najem-Meyer, A. Doucet, S. Clematide, Overview of HIPE-2022: Named Entity Recognition and Linking in Multilingual Historical Documents, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction: 13th International Conference of the CLEF Association, CLEF 2022, Bologna, Italy, September 5–8, 2022, Proceedings, Springer-Verlag, Berlin, Heidelberg, 2022, pp. 423–446. URL: https://doi.org/10.1007/978-3-031-13643-6_26.

[15] M. Ehrmann, M. Romanello, S. Najem-Meyer, A. Doucet, S. Clematide, G. Faggioli, N. Ferro, A. Hanbury, M. Potthast, Extended Overview of HIPE-2022: Named Entity Recognition and Linking in Multilingual Historical Documents, in: CEUR Workshop Proceedings, 3180, CEUR-WS, 2022, pp. 1038–1063.

[16] J. Opitz, C. Raclé, E. Boros, A. Michail, M. Romanello, M. Ehrmann, S. Clematide, CLEF HIPE-2026: Evaluating Accurate and Efficient Person–Place Relation Extraction from Multilingual Historical Texts, in: R. Campos, A. Jatowt, Y. Lan, M. Aliannejadi, C. Bauer, S. MacAvaney, A. Anand, Z. Ren, S. Verberne, N. Bai, M. Mansoury (Eds.), Advances in Information Retrieval, Springer Nature Switzerland, Cham, 2026, pp. 354–363. doi:10.1007/978-3-032-21321-1_46.

[17] J. Opitz, M. Ehrmann, S. Clematide, R. Corina, E. Boros, A. Michail, M. Romanello, CLEF HIPE-2026 - Shared Task Participation Guidelines, 2026. URL: https://doi.org/10.5281/zenodo.20082076.

doi:10.5281/zenodo.20082076.
[18] J. R. Hobbs, M. E. Stickel, D. E. Appelt, P. Martin, Interpretation as abduction, Artificial intelligence 63 (1993) 69–142.
[19] M. Ehrmann, M. Romanello, A. Doucet, S. Clematide, HIPE 2022 Shared Task Participation Guidelines, Technical Report, Zenodo, 2022. URL: https://zenodo.org/record/6045662. doi:10.5281/zenodo.6045662.
[20] S. Gabay, P. Ortiz Suarez, R. Bawden, A. Bartz, P. Gambette, B. Sagot, Le projet FREEM : ressources, outils et enjeux pour l'étude du français d'ancien régime (the F RE EM project: Resources, tools and challenges for the study of ancien régime French), in: Y. Estève, T. Jiménez, T. Parcollet, M. Zanon Boito (Eds.), Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 1 : conférence principale, ATALA, Avignon, France, 2022, pp. 154–165. URL: https://aclanthology.org/2022.jeptalnrecital-taln.15/.
[21] S. Gabay, T. Clérice, M. Gille Levenson, J.-B. Camps, J.-B. Tanguy, Freem-corpora/freemlpm: Freem lpm (lemma, pos- tags, morphology) corpus, 2022. doi:10.5281/zenodo.6481300.
[22] P. Blumenthal, S. Diwersy, A. Falaise, M.-H. Lay, G. Sourvay, D. Vigier, Presto, un corpus diachronique pour le français des XVIe-XXe siècles, in: atelier " Les corpus annotés du français ", Actes de TALN 2017, Orléans, France, 2017. URL: https://shs.hal.science/halshs-01585010.
[23] J. Opitz, A closer look at classification evaluation metrics and a critical reflection of common evaluation practice, Transactions of the Association for Computational Linguistics 12 (2024) 820–836. URL: https://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00675/2465598/tacl_a_00675.pdf. doi:10.1162/tacl_a_00675.
[24] F. Sebastiani, An axiomatically derived measure for the evaluation of classification algorithms, in: Proceedings of the 2015 International Conference on the Theory of Information Retrieval, 2015, pp. 11–20.
[25] J. Opitz, M. Ehrmann, S. Clematide, C. Raclé, E. Boros, A. Michail, M. Romanello, CLEF HIPE-2026 - Shared Task Participation Guidelines, 2026. URL: https://doi.org/10.5281/zenodo.20082076.
[26] A. H. Liu, K. Khandelwal, S. Subramanian, V. Jouault, A. Rastogi, et al., Ministral 3, 2026. URL: https://arxiv.org/abs/2601.08584. arXiv:2601.08584.
[27] D.-M. Vasile, E.-S. Apostol, C.-O. Truâ, Awakened at hipe-2026: Knowledge-augmented prompting and logic-constrained fine-tuning for person–place relation extraction, 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.
[28] R. Keinan, R. Tsarfaty, Where Were They? Person-Place Relation Extraction in Multilingual Historical Archives, 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.
[29] M.-T. Wesley, Lightweight Person-Place Relation Extraction from Historical Newspapers with Dependency Graphs and Proximity Features, 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.
[30] L. Griesbeck, M. Hennen, F. Babl, M. Geierhos, FI-CODE@HIPE-2026: Efficient Person-Place Relation Classification from Distance Heuristics to Prompted Language Models, 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.
[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: D. Jurafsky, J. Chai, N. Schluter, J. Tetreault (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 8440–8451. URL: https://aclanthology.org/2020.acl-main.747/. doi:10.18653/v1/2020.acl-main.747.
[32] S. Yazdani, J. P. Wahle, B. Gipp, GippLab at HIPE 2026: Unlocking Person–Place Relation Extraction with the Qwen Model Family, 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.
[33] A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv, et al., Qwen3 technical report, arXiv preprint arXiv:2505.09388 (2025).
[34] P. S. Shekhawat, R. Gupta, From Frozen Encoders to Multi-Agent LLM Pipelines: An Iterative Journey through Person–Place Relation Extraction in 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.
[35] A. Yang, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Li, D. Liu, F. Huang, H. Wei, et al., Qwen2.5 technical report, arXiv preprint arXiv:2412.15115 (2024).
[36] S. Nguyen, R. Zheng, D. Nurbakova, K. Barrere, INSA Lyon at HIPE 2026: A Modular Ensemble Pipeline for Person–Place Relation Classification in Historical Newspapers, 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.
[37] H. Cao, Z. Han, X. Duan, Few-Shot Prompting with Large Language Models for Multilingual Person–Place Relation Extraction from Historical Documents, 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.
[38] N. H. Pham, J. G. Moreno, A. Doucet, MILRIT at HIPE 2026: From Generalist Relation Extraction to Multi-View Learning for Historical Person-Place Relations, 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.
[39] J. Boylan, C. Hokamp, D. G. Ghalandari, GLiREL - generalist model for zero-shot relation extraction, in: L. Chiruzzo, A. Ritter, L. Wang (Eds.), Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), Association for Computational Linguistics, Albuquerque, New Mexico, 2025, pp. 8230–8245. URL: https://aclanthology.org/2025.naacl-long.418/. doi:10.18653/v1/2025.naacl-long.418.
[40] T. Checchin, A. Guille, N. Guteherlé, Some LLMs Are Smaller Than Others: A Lightweight Linear Model for Relation Extraction Applied to Historical Documents, 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.
[41] R. Ram, S. Soren, S. Bandyopadhyay, Layer Freezing and Cost-Sensitive Learning for English Historical Relation Classification: Jadavpur University at CLEF-HIPE 2026, 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.
[42] A.-K. Dick, J. Hermes, N. Reiter, Scaffolding Large and Small Reasoning Models for Person-Place Relation Extraction, 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.
[43] J. Gomez-Navalon, T. Bernal-Beltrán, R. Pan, J. A. García-Díaz, R. Valencia-García, UMUTeam at HIPE-CLEF 2026: Person-Place Relation Extraction from Multilingual Historical Documents, 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.
[44] D. Almanza González, J. C. Martinez-Santos, E. Puertas, A Multitask Approach Based on RoBERTa, Temporal Transformer, and Person-Location Relation Extraction in Multilingual Historical Text, 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.
[45] A. Morillo, E. Puertas, J. C. Martinez-Santos, Person-Place Relation Extraction in Historical Texts: A Dual Approach with Fine-Tuned NLI Encoders and SLMs at HIPE-2026, 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.
[46] Y. Aboelwafa, A. Samir, N. Elmakky, M. Torki, DistilledGemma: Balanced Efficiency-Accuracy

for Person-Place Relation Extraction from Multilingual Historical Articles, 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.