From Frozen Encoders to Multi-Agent LLM Pipelines: An Iterative Journey through Person–Place Relation Extraction in Multilingual Historical Texts

Team Hansel&Gretel at CLEF HIPE-2026

Pradyuman Singh Shekhawat[1,*], Rohan Gupta[1]

[1]*Indian Institute of Technology Roorkee, Roorkee, Uttarakhand 247667, India*

Abstract

We present the system description and experimental analysis of Team Hansel&Gretel for the CLEF HIPE-2026 Shared Task on Person–Place Relation Extraction from Historical Multilingual Documents. The task requires classifying, for each person–location pair in a historical newspaper article (German, French, or English, spanning the 19th–20th centuries), two temporal relations: `at` (person was ever at the location, 3-class) and `isAt` (person was there within approximately one month of publication, binary). Evaluation uses Macro Recall, equally weighting all label classes. We explore nine progressively sophisticated strategies, from a frozen XLM-RoBERTa encoder with classical ML classifiers (internal dev Global Macro Recall $= 0.6018$) to a Qwen2.5-3B-Instruct Generative Supervised Fine-Tuning pipeline with Wikidata knowledge injection (dev $= 0.7289$) and a three-agent Historian–Geographer–Arbiter multi-agent architecture (dev $= 0.7280$). On the official competition test set, our best run (Run 3: GPT-OSS 120B with Chain-of-Thought and Wikidata injection) achieves an impresso score of **0.6221**, ranking 13th out of 46 ranked entries (44 participant runs covering all three languages plus two organiser baselines; 17 teams and 45 participant runs in total). This paper documents the design decisions, failure modes, and key insights behind each strategy, including an honest analysis of the 17-point dev-to-test gap, providing a comprehensive account of iterative NLP system development under data scarcity, class imbalance, and multilingual historical OCR noise.

Keywords

Person–Place Relation Extraction, Historical NLP, Multilingual Transformers, XLM-RoBERTa, LoRA, Generative Classification, Multi-Agent LLM, Wikidata, HIPE-2026, CLEF

1. Introduction

The digitisation of historical newspapers has created vast corpora of 19th and 20th century text encoding biographical trajectories and cultural history. Automatically recovering *where people were and when* from these documents enables digital humanities scholars to reconstruct spatio-temporal life trajectories at scale. This is the central challenge of **CLEF HIPE-2026** [1, 2, 3]:[1] the Shared Task on Person–Place Relation Extraction from multilingual historical documents, the third edition in the HIPE series following HIPE-2020 [4] and HIPE-2022 [5].

Unlike standard named entity recognition, HIPE-2026 requires *temporal reasoning*: a system must determine whether a text provides evidence for a *historical* presence (`at`) or a *contemporary* one (`isAt`, bounded to $\approx$1 month before publication). This distinction, combined with multilingual OCR-noisy text, severe class imbalance, and a small labelled dataset, makes the task deceptively difficult.

Team Hansel&Gretel approached this as an iterative research problem, developing **nine distinct strategies** that fall into three methodological phases:

1. **Feature extraction** (Phase 1): **S1** frozen XLM-RoBERTa + classical ML classifiers, and **S1.5** multi-task learning (MTL) with Optuna hyperparameter optimisation.

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

*Corresponding author. Also reachable at: shekhawat.pradyuman10@gmail.com

pradyuman_ss@cs.iitr.ac.in (P. S. Shekhawat); rohan_g@cs.iitr.ac.in (R. Gupta)

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

2. **Discriminative fine-tuning** (Phase 2): **S2** full end-to-end XLM-RoBERTa fine-tuning, **S2.5** PEFT/LoRA on XLM-RoBERTa, **S3** monolingual LoRA routing, and **S4/S4.5** Qwen2.5-0.5B MTL with LoRA (and a soft-label variant).
3. **Generative and multi-agent LLMs** (Phase 3): **S5** Qwen2.5-3B-Instruct generative SFT *(best dev score)*, **S6/S7** a three-specialist multi-agent pipeline (Historian/Geographer/Arbiter), and **S8/S9** Chain-of-Thought few-shot via IBM RITS large LLMs with Wikidata injection *(best official test score)*.

The paper is organised as follows. Section 2 defines the task, describes the data and baselines, and states our experimental setup. Section 3 presents the three methodological phases and a cross-strategy comparison. Section 4 reports the official test results and the dev-to-test gap analysis. Section 5 presents the error analysis, and Section 6 discusses limitations, future work, and conclusions.

2. Task and Data

2.1. Task Definition and Evaluation

HIPE-2026 provides historical newspaper documents as input. Each document contains full article text with metadata (title, publication date, language) and a list of sampled (person, location) entity pairs (up to 16 per document). For each pair, systems classify two relations:

$$f(\text{context}, (\text{person}, \text{location})) \mapsto \Big(\underbrace{\texttt{at}}_{\in\{\text{TRUE},\text{PROBABLE},\text{FALSE}\}} , \underbrace{\texttt{isAt}}_{\in\{\text{TRUE},\text{FALSE}\}} \Big) \tag{1}$$

The `at` relation is ternary: TRUE if the text explicitly places the person at the location at any past time; PROBABLE if inferable from implicit cues; FALSE otherwise. The `isAt` relation is binary: TRUE only within $\sim$1 month of publication. A hard *transitivity constraint* applies: $\texttt{isAt} = \text{TRUE} \implies \texttt{at} \in \{\text{TRUE}, \text{PROBABLE}\}$.

The primary metric is **Macro Recall** (also called balanced accuracy [6]), computed independently per sub-task as $\text{MacroRecall} = \frac{1}{|L|} \sum_{\ell \in L} \frac{\#\text{correctly predicted with gold label } \ell}{\#\text{examples with gold label } \ell}$. The **impresso score** averages the two sub-task macro recalls ($\text{MacroRecall}_{\texttt{at}}$ and $\text{MacroRecall}_{\texttt{isAt}}$). Three evaluation *profiles* are assessed: the **Accuracy profile** (Macro Recall on the impresso newspaper Test A set), the **Efficiency profile** (arithmetic mean of three ranks — accuracy, parameter count, and model size — for which *lower is better*), and the **Generalization profile** (Macro Recall on the surprise French literary Test B set, `at` only).

2.2. Data, Splits, and Baselines

The dataset derives from the HIPE-2022 [5] entity-annotated newspaper corpora, re-annotated for person–place relation extraction. **Domain A** (*Test A*, the "impresso" collection) covers historical newspapers in German, French, and English spanning the 19th–20th centuries; **Domain B** (*Test B*, the "surprise" set) is an out-of-domain corpus of *early modern French literary and historiographical texts from the 16th–18th centuries* (annotations derived from the FreEMNER corpus), used to assess cross-domain generalisation on the `at` relation only [3]. All data is distributed under CC-BY 4.0 through the Impresso Media Monitoring of the Past project [7].[2] Domain A training and development labels were produced by majority vote over three runs of GPT-4.1 (*silver*), with a per-language subset subsequently human-annotated (*gold*); candidate pairs were sampled at up to 16 per document with probability proportional to $\exp(-d/T)$ over the person–location text distance d (T=50), and documents exceeding 12,000 characters were removed during construction [3].

Following the organiser terminology, *Gold Train A* (104 docs, 1,251 pairs) is the manually annotated training set and *Silver Train A* (380 docs, 5,209 pairs) the larger auto-annotated set. The official

[2]Competition data (v1.0): https://github.com/hipe-eval/HIPE-2026-data/releases/tag/v1.0. All scores in this paper use the official scorer and evaluation toolkit: https://github.com/hipe-eval/HIPE-2026-eval.

Table 1
Official HIPE-2026 v1.0 data splits, following the organiser overview terminology [3]. *Gold* = human-annotated; *Silver* = GPT-4.1 majority-vote auto-annotated. Test A is the impresso newspaper set; Test B is the surprise French literary set (`at` only).

Split	Lang.	Docs	Pairs	Years
Gold Train A	DE	34	466	1818–1948
	EN	35	307	1790–1960
	FR	35	478	1804–2018
Silver Train A	DE	64	893	1798–1948
	EN	23	206	1790–1960
	FR	293	4,110	1798–2018
Silver Dev A	DE	32	432	1818–1948
	EN	17	151	1790–1920
	FR	107	1,498	1798–1988
Gold Test A	DE	19	238	1808–1948
	EN	19	162	1800–1960
	FR	19	238	1804–1998
Gold Test B (surprise)	FR	30	480	1542–1797

hidden test sets — *Test A* (57 docs, 638 pairs across DE/EN/FR) and *Test B* (30 docs, 480 pairs) — were released separately and scored by the organisers. Two organiser baselines frame the results: a zero-shot `mistralai/Ministral-3-3B-Instruct-2512` [8] (GGUF quantised, greedy decoding) scored an impresso **0.5818** on Test A (rank 17/46), while a random baseline scored **0.4049**; on our internal 254-pair dev set the random baseline yields `at` MR $= 0.3211$, `isAt` MR $= 0.5273$, Global MR $= 0.4242$, establishing the floor.

2.3. Exploratory Data Analysis

We performed detailed EDA on the gold training split available at experiment time (104 documents, 1,251 pairs; Table 2). Three properties shaped every modelling decision: (i) **context length** — the maximum document is ∼12,000 chars (≈3,000 tokens), exceeding standard 512-token encoder limits and motivating long-context generative LLMs; (ii) **class imbalance** — `isAt=FALSE` at 77.7% and PROBABLE at only 9.6% of `at` labels (Figure 1), so Macro Recall heavily penalises naive majority-class predictors; and (iii) **multilinguality by design** — an even DE/EN/FR split, so monolingual models immediately forfeit cross-lingual transfer. These are compounded by OCR noise from historical typesetting (e.g., "Mufschen" for "München"), archaic vocabulary absent from modern pre-training corpora, entity ambiguity, and severe data scarcity (∼100 gold documents across 3 languages and 3 label classes at experiment time).

Table 2
Gold training split EDA statistics (104 documents, early experiment phase).

Metric	Value
Total candidate pairs	1,251
Language split	34 DE / 35 EN / 35 FR
Unique persons	496
Unique locations	469
Average document length	3,465 chars
Maximum document length	11,738 chars
`at`: TRUE / PROBABLE / FALSE	441 / 120 / 690
`isAt`: TRUE / FALSE	279 / 972

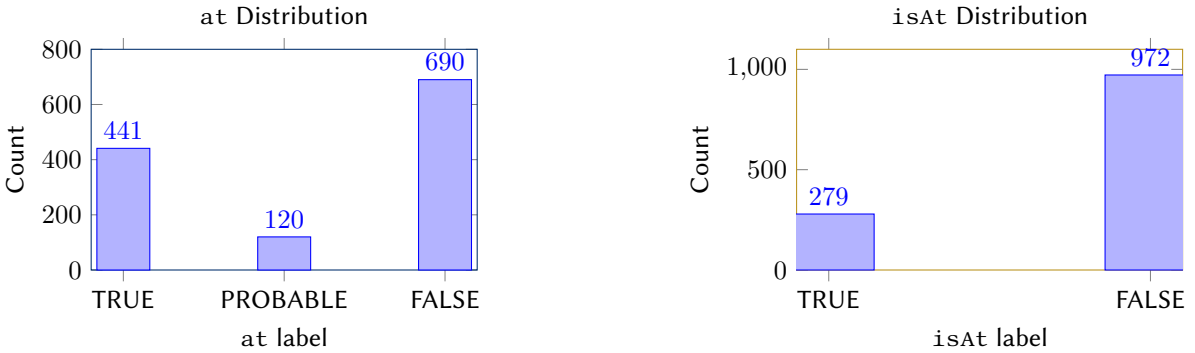

Figure 1: Label distributions in the gold training set. Severe imbalance in `isAt` (78% FALSE) drives the choice of Macro Recall as the evaluation metric.

2.4. Experimental Setup

To keep results comparable across strategies, we fix the following setup once. **Training data:** discriminative and fine-tuned strategies (Phases 1–2 and S5/S6/S7) train on the combined corpus of Gold Train A (104 docs, up-weighted $5\times$ except in S1 which uses $50\times$) and Silver Train A ($1\times$), for a total of up to 15,418 weighted training pairs; the zero-training strategies (S8/S9) use no task-specific training data. **Internal evaluation:** all dev scores reported below are computed with the official scorer on the early-release development subset of **254 pairs (21 documents)** available during the main experimental phase; the full Silver Dev A split became available only near the submission deadline. Because this subset is small, *its scores are optimistic proxies and are not directly comparable to the official held-out Test A scores* (Section 4). **Hardware:** fine-tuning used $1\times$ NVIDIA RTX A5000 (24 GB) for S5 and $3\times$ RTX A5000 for the 7B multi-agent S7; S8/S9 used IBM's RITS inference gateway (no local GPU). Full per-strategy hyperparameters are given in Appendix A.

3. Systems and Results: An Iterative Journey

We now walk through the three phases. Each subsection ends with the best internal-dev Global Macro Recall (GMR) reached in that phase; the consolidated per-strategy scores are collected in Table 4.

3.1. Phase 1 — Frozen Encoder + Classical ML (S1, S1.5)

S1 (feature extraction). With only 104 gold documents, fine-tuning 270M parameters risks catastrophic overfitting, so S1 uses frozen `xlm-roberta-base` [9] as a feature extractor. Person and location mentions are wrapped in `<E1>...</E1>` and `<E2>...</E2>` marker tokens; the contextual hidden states at these markers are concatenated into a 1536-dim relation-aware feature vector. We deliberately pair each sub-task with a classifier matched to its label structure: a `RandomForestClassifier` for the ternary, moderately balanced `at` relation (robust to a three-way split), and a gradient-boosted `LGBMClassifier` [10] for the strongly imbalanced binary `isAt` relation (where boosting with class weighting better recovers the rare `TRUE` class). A post-hoc transitivity rule forces `isAt=FALSE` when `at=FALSE`. Both gold (104 docs) and silver sandbox (617 docs) examples are used, with gold up-weighted $50\times$ so human annotations dominate the decision boundary while silver supplies historical vocabulary coverage. This weighting is decisive: GMR rises from **0.4459** gold-only (80/20 split) to **0.6018** with silver (a $+34.9\%$ relative gain).

S1.5 (MTL + Optuna). Two independent classifiers cannot exploit the `at–isAt` logical correlation, so S1.5 replaces them with a single PyTorch multi-task network over the frozen embeddings (shared dense + GELU + dropout, then two linear heads) trained with a combined loss $\mathcal{L} = \alpha\mathcal{L}_{\text{at}}^{\text{WCE}} + (1-\alpha)\mathcal{L}_{\text{isAt}}^{\text{Focal}}$ [11], with Optuna [12] (50 trials, TPE sampler) tuning α, learning rate, dropout, hidden dimension, and

per-class weights.[3] Jointly modelling the correlation lifts GMR to **0.6249**. *Phase 1 best dev GMR: 0.6249 (S1.5).*

3.2. Phase 2 — Discriminative Fine-Tuning (S2–S4.5)

S2 (full fine-tuning; catastrophic failure). Full fine-tuning of `xlm-roberta-base` at lr$= 2 \times 10^{-5}$ (10 epochs, batch 8) caused **catastrophic mode collapse**: the model predicted TRUE for every `at` and FALSE for every `isAt` (GMR $= 0.4167$, below S1.5 by -0.2082). Three causes compounded: catastrophic forgetting from unfreezing 270M parameters on only a few hundred documents; destructive gradient interference from batches mixing archaic German (1798), Victorian English (1860), and 19th-century French; and randomly initialised classification heads producing large gradients that overwhelmed the pre-trained layers.

S2.5 (PEFT/LoRA). LoRA [13] (rank 8, α 16, 0.11% trainable parameters) prevented complete collapse (GMR $= 0.4541$) and restored partial minority-class recall, but remained far below the frozen baseline because $\sim$230 documents per language are too few to adapt attention matrices without overfitting.

S3 (monolingual LoRA routing; regression). Training three separate LoRA adapters (DE, EN, FR) with document-language-based routing at inference eliminates cross-lingual gradient conflicts but *worsened* performance (GMR $= 0.4232$): partitioning shrinks each adapter's data to $\sim$233 documents, and the shared multilingual volume in S2.5 outweighs the harm from interference. The key lesson is that **data scarcity, not language mixing, is the binding constraint**.

S4 / S4.5 (Qwen2.5-0.5B MTL + LoRA). Pivoting to a decoder-only backbone, `Qwen2.5-0.5B` [14] (500M parameters, 32k-token context) with LoRA (r=16, α=32) and MTL heads matched Phase-1 performance (GMR $= 0.6264$) while handling long documents without truncation. Adding label smoothing ($\epsilon = 0.15$) in S4.5 slightly regressed to **0.6212**, adding noise without compensating signal on the minority PROBABLE class. *Phase 2 best dev GMR: 0.6264 (S4)* — discriminative approaches plateau near 0.63 regardless of backbone, limited by data scarcity and the inherent ambiguity of PROBABLE.

3.3. Phase 3 — Generative and Multi-Agent LLMs (S5–S9)

S5: Generative SFT (best dev score). Temporal reasoning ("is the person at the location *right now*?") requires multi-step inference: locate the event in time, compare it to the publication date, and assess the temporal window. Discriminative classifiers select labels but never reason through *why*. S5 makes a decisive pivot: `Qwen2.5-3B-Instruct` [14] (3.1B parameters) is fine-tuned with LoRA (r=16, α=32, all linear layers) in a Generative Supervised Fine-Tuning regime to *generate* a structured JSON answer (e.g. `{"at": "TRUE", "isAt": "FALSE"}`), with the cross-entropy loss masked to the answer tokens only. We initially targeted Gemma-2-2B-IT but pivoted due to Hugging Face gating; Qwen2.5-3B-Instruct proved superior (non-gated, larger, stronger multilingual instruction following). At inference a **Wikidata death-date hard rule** queries each person's `pers_wikidata_QID`: if the person died >30 days before publication, `isAt` is forced to FALSE (a local JSON cache avoids redundant API calls).[4] We audited this rule on the dev set: of the 254 pairs, 120 carry a person QID, but the cache resolved a valid death date for only 8 distinct persons (spanning 27 pairs). Crucially, in every one of those pairs the person was still alive at the article's publication date, so the condition was satisfied for **0 of 254 dev pairs** and the rule altered no dev prediction. It is therefore best understood as a high-precision safeguard for the hidden test distribution — where clearly deceased historical figures are more likely to be discussed — rather than a contributor to our dev score. Accordingly, S5's **largest single jump in the project**, GMR $0.6212 \rightarrow$ **0.7289** ($+0.1077$), is driven by model scale, the full 1024-token context, and generative reasoning that forces explicit temporal decision-making; the Wikidata rule did not fire on dev.

[3] Code and best-trial hyperparameters: https://github.com/ikshv4ku/HIPE-2026-Team-Hansel-Gretel

[4] Engineering notes: fitting the 3B bf16 model ($\sim$7 GB) on a single 24 GB RTX A5000 required gradient checkpointing and expandable memory segments, and the 16 GB virtual environment was relocated to a mounted volume to work around root-partition disk limits.

S6/S7: Multi-agent pipeline. Human experts reason through specialised lenses, so S6 replaces the single generalist with three agents: a **Historian** (temporal/biographical reasoning), a **Geographer** (spatial/geographic reasoning), and an **Arbiter** that synthesises verdicts. Each specialist receives a dedicated system prompt — the Historian's, for example, reads *"You are a historical biographer. Analyse the text for biographical evidence, life events, official roles, and temporal cues (verb tenses, dates)"* — while the Arbiter is instructed to *"adopt their verdict if they agree; if they disagree, apply the more conservative judgment."* On the 3B backbone, S6 suffered *specialist collapse*: both Historian and Geographer predicted FALSE for everything (hist_TRUE=0, geo_TRUE=0), the Arbiter became the sole decision-maker, agent agreement reached 96.9% (246/254), and GMR slipped to 0.7204. S7 fixes this with three interventions: scaling to a 7B backbone, applying a 3× class-weighted loss on TRUE/PROBABLE answer tokens, and adding three worked few-shot examples per specialist. This restores specialist diversity (Historian TRUE=53, Geographer TRUE=47; agreement drops to 92.9%, so the Arbiter resolves 18 genuine disagreements) and yields the **highest `isAt` MR of any strategy (0.7817)**, though it marginally trails S5 on `at` (0.6742 vs 0.6872), for GMR = 0.7280 (Table 3).

Table 3
Multi-agent progression on the internal dev set: the effect of scale, class weighting, and few-shot prompts on specialist diversity (merges the S5/S6 and S6/S7 comparisons).

Metric	S5 (single 3B)	S6 (3B agents)	S7 (7B, 3× wt.)
`at` Macro Recall	**0.6872**	0.7019	0.6742
`isAt` Macro Recall	0.7706	0.7388	**0.7817**
Global MR	**0.7289**	0.7204	0.7280
Historian TRUE predictions	—	0	53
Geographer TRUE predictions	—	0	47
Agent agreement rate	—	96.9% (246/254)	92.9% (236/254)
Arbiter-resolved cases	—	8/254	18/254

S8/S9: Chain-of-Thought via IBM RITS (best official test score). S8 and S9 attack the task **without any model training**, using Chain-of-Thought (CoT) few-shot prompting [15] through IBM's RITS inference gateway. Each prediction is preceded by three explicit reasoning steps — (1) Biographical & Temporal Analysis, (2) Geographic & Contextual Analysis, (3) Synthesis & Decision — with four worked examples covering all label combinations. S8 was exploratory: it calibrated the prompt template and compared candidate models (Llama 3.3 70B, Granite 3.2 8B, Llama 4 Maverick 17B, and GPT-OSS 120B), identifying **GPT-OSS 120B** as the strongest. S9 extends S8 by dynamically injecting Wikidata facts via SPARQL: for each pair it fetches person (birth date, death date, description) and location (description, country) facts as a `[WIKIDATA FACTS]` block, grounding temporal reasoning in structured knowledge rather than parametric memory. S9 (GPT-OSS 120B) constituted our **Run 3** submission for all four test sets and achieved our best official ranking (13th/46, impresso 0.6221; Section 4). We report S8/S9 as *test-only* to reflect our actual competition workflow: because they require no fine-tuning, we prioritised applying them directly to the official test sets under the deadline. A post-hoc dev-set evaluation could in principle have been added; we note its absence as a reporting limitation rather than a methodological one. *Phase 3 best dev GMR: 0.7289 (S5); S8/S9 evaluated on test only.*

3.4. Cross-Strategy Comparison

Table 4 consolidates all internal-dev scores; the same progression is plotted in Appendix B (Figure 3). Three phases emerge clearly: *feature extraction* (S1–S1.5, best 0.6249) shows frozen encoders excel on tiny datasets; *discriminative fine-tuning* (S2–S4.5, plateau ∼0.63) shows data scarcity limits every fine-tuning paradigm and includes two instructive failures (S2 collapse, S3 regression); and *generative/multi-agent* (S5–S7, best 0.7289) shows reasoning-as-generation breaking the plateau. The single-model S5 and the multi-agent S7 are complementary: S5 leads on `at` (0.6872) while S7 leads on `isAt` (0.7817).

Table 4
Complete cross-strategy performance on the internal dev set (254 pairs, 21 docs). All values are the impresso score (Global Macro Recall). S8 and S9 have no dev scores by design (Section 3).

Strategy	Architecture	at MR	isAt MR	Global MR
Random	Dummy	0.3211	0.5273	0.4242
S1	XLM-R frozen + RF/LGBM	0.5200	0.6835	0.6018
S1.5	XLM-R frozen + MTL	0.5486	0.7013	0.6249
S2	XLM-R full fine-tune	0.3333	0.5000	0.4167
S2.5	XLM-R + LoRA	—	—	0.4541
S3	Monolingual LoRA routing	0.3500	0.5000	0.4232
S4	Qwen2.5-0.5B MTL + LoRA	0.5313	0.7111	0.6264
S4.5	Qwen2.5-0.5B + soft labels	0.5313	0.7111	0.6212
S5	**Qwen2.5-3B Generative SFT**	**0.6872**	0.7706	**0.7289**
S6	Multi-agent 3B	0.7019	0.7388	0.7204
S7	Multi-agent 7B + class weights	0.6742	**0.7817**	0.7280
S8	CoT 120B via RITS (exploratory)	Not evaluated on dev		
S9	Wikidata-aug. CoT 120B (Run 3)	Run directly on test		

4. Official Test Results

We reiterate that the strategy-progression scores in Section 3 are internal-dev estimates on 254 pairs and are **not directly comparable** to the official held-out Test A scores reported here. For context, HIPE-2026 attracted **17 participating teams** submitting **45 runs**; the overall Accuracy-profile ranking comprises 46 complete entries (44 participant runs covering all three languages plus two organiser baselines). Full official rankings are in the overview papers [3, 2].

4.1. Per-Run and Per-Language Performance

Table 5 reports our three official submissions on Test A (impresso score = mean of $\text{MacroRecall}_{\texttt{at}}$ and $\text{MacroRecall}_{\texttt{isAt}}$ across DE/EN/FR). Run 3 (S9) is our best, at 0.6221 (rank 13/46), followed by Run 2 (S7, 0.5788) and Run 1 (S5, 0.5458); all three exceed the organiser Ministral baseline (0.5818 only surpassed by Run 3). Table 6 gives the per-language breakdown for Run 3.

Table 5
Official impresso Test A scores for all three Hansel&Gretel runs. Run 3 (S9) is our best overall submission.

Run	Strategy	Overall	DE	EN	FR	Rank
Run 1	S5 (Qwen2.5-3B SFT)	0.5458	0.521	0.5976	0.5189	23/46
Run 2	S7 (Multi-agent 7B)	0.5788	0.5254	0.5866	0.6243	19/46
Run 3	**S9 (GPT-OSS 120B CoT+Wiki)**	**0.6221**	**0.5994**	**0.6367**	**0.6303**	**13/46**

Table 6
Per-language breakdown of at-MR and isAt-MR for Run 3 (best run, S9) on Test A.

Language	at-MR	isAt-MR	impresso score
German (DE)	0.4877	0.7112	0.5994
English (EN)	0.5582	0.7151	0.6367
French (FR)	0.5268	0.7338	0.6303

4.2. Generalization and Efficiency

On the **Generalization profile** (surprise literary French Test B, `at` only), Run 2 (S7) performed best among our submissions (0.6349, rank 11/46), ahead of Run 3 (S9, 0.6187, rank 14/46) and Run 1 (S5, 0.6107, rank 15/46). That the fine-tuned multi-agent Run 2 leads here is consistent with the multi-agent architecture learning more transferable representations, its specialist decomposition acting as an implicit domain-agnostic inductive bias; it is also expected that S9 does *not* lead on Test B, since its Wikidata facts are tuned to newspaper-era figures and transfer poorly to 16th–18th-century literary texts. On the **Efficiency profile**, Run 1 (S5, 3B, 6,248 MB) ranked 19th, Run 2 (S7, 7B, 15,300 MB) 23rd, and Run 3 (S9, 120B, 240,000 MB) 29th; the low efficiency rank of Run 3 is expected for a 120B cloud-hosted model used without fine-tuning, and Run 1 is our best efficiency–accuracy trade-off.

4.3. Competition Context

Table 7 and Figure 2 situate our best run against the leading systems and baselines. Our Run 3 (0.6221) sits at rank 13/46, well above both organiser baselines.

Table 7
Competition Accuracy-profile context: selected teams ranked by best overall impresso score on Test A.

Rank	Team	Affiliation	Best Score
1	Spinfo	Universität zu Köln	0.7479
2	MaxFo-Ajie	Foshan University	0.7001
3	whereami	Alexandria University	0.6880
	…(ranks 4–12) …		
13	**Hansel&Gretel**	**IIT Roorkee**	**0.6221**
	…(ranks 14–16) …		
17	Baseline (Ministral-3B)	HIPE-2026 Organizers	0.5818
—	Random Baseline	HIPE-2026 Organizers	0.4049

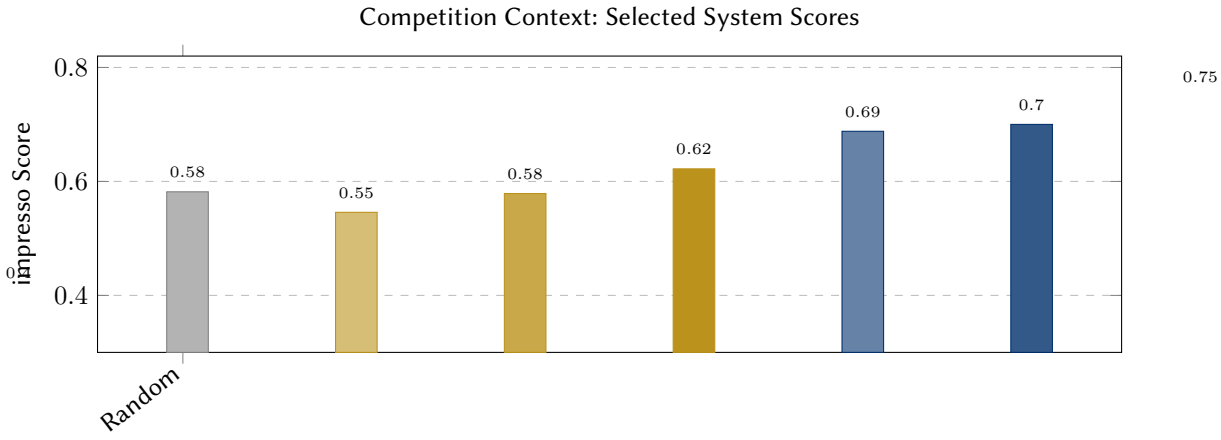

Figure 2: Competition context for the Accuracy profile (impresso Test A). Grey: baselines; gold: Hansel&Gretel runs (R1–R3); blue: top-3 systems. H&G Run 3 (0.6221) substantially outperforms both baselines and ranks 13th of 46 entries.

4.4. Dev-to-Test Gap Analysis

A substantial gap (≈17 points) separates our best internal dev score (S5: 0.7289) from our best official test score (Run 3/S9: 0.6221). We attribute it to four factors. First, **dev-set size**: 254 pairs from 21 documents is a small, potentially non-representative sample, whereas Test A is larger and independently

held out, drawn from a different temporal and geographic distribution within the same collections. Second, **fine-tuned overfitting**: S5 and S7 were trained on the combined data (104 gold Gold-Train-A documents up-weighted $5\times$ plus 380 silver documents); despite this weighting, document-level patterns (source style, annotation conventions) likely fail to replicate on Test A, and fine-tuned models are more sensitive to such shift. Third, **S9's robustness**: Run 3 (0.6221) beats Run 1 (0.5458) and Run 2 (0.5788) on Test A despite S5 having the best dev score — a reversal indicating that the zero-shot GPT-OSS 120B, with no exposure to HIPE training data, is more robust to distribution shift thanks to broad pre-training. Fourth, **Wikidata coverage gaps**: the death-date hard rule fires only when QIDs and biographical data are available. Coverage is sparse — on the dev set the cache resolved death dates for only 8 of 53 distinct person QIDs — and the rule fired on 0/254 dev pairs, so any benefit it provides is confined to (unmeasured) test cases involving clearly deceased figures. The net lesson: dev scores estimated on 254 pairs are optimistic proxies for held-out performance, and robust evaluation requires larger dev sets with explicit distribution control.

5. Error Analysis

Class-level hardness. Across all strategies, the `at=PROBABLE` class was consistently the hardest to classify. In S1 (dev split), `at` recall decomposed as TRUE ≈ 0.72, PROBABLE ≈ 0.18, FALSE ≈ 0.67. PROBABLE (9.6% of labels) requires implicit inference from contextual cues — a person's title, historical role, or narrative proximity — rather than direct textual evidence, so models consistently under-predict it and default to the safer FALSE. This pattern held through S7: even our best strategies could not reliably recall the minority PROBABLE instances, and we believe the small dev sample further inflated apparent PROBABLE performance relative to the more challenging test distribution.

Per-language variation. On the official test (Run 3/S9; Table 6), German has the lowest at-MR (0.4877), plausibly due to heavier OCR noise in digitised German Gothic script (*Fraktur*) and more complex syntax for temporal reasoning; English attains the best impresso score (0.6367), benefiting from richer English historical-newspaper signal in GPT-OSS 120B's pre-training; and French is competitive (0.6303) with a notably strong isAt-MR (0.7338). The fine-tuned multilingual models show distinct cross-language profiles — Run 2 (S7) leads on French (0.6243) while Run 1 (S5) leads on English (0.5976) — indicating that different architectures capture complementary language-specific signals.

OCR noise and its downstream impact. Historical OCR produced two systematic error types: character substitutions (e.g., "Mufschen" for "München" and "Preufien" for "Preußen") and mid-word token splits (e.g., "Kut zeroff" for the garrison commander "Kutzeroff", as in the *Gazette de Lausanne* of 6 May 1928). These fragment entity mentions and had concrete downstream effects on our systems: in S3, mixed-language documents (French text referencing German places) confused the language-conditioned adapter routing; and fragmented names weakened the marker-token embeddings in S1–S1.5 and destabilised sub-word tokenisation for the fine-tuned decoders, contributing to the depressed German at-MR (0.4877). The zero-shot 120B model in S9 proved more robust to such noise via its larger parametric vocabulary, whereas the fine-tuned S5/S7 models partly compensated by training on silver sandbox data that contains similar historical OCR patterns.

6. Limitations, Future Work, and Conclusion

Limitations. Our study is bounded by several factors. The fundamental bottleneck is **data scarcity**: 104 gold documents across three languages and three label classes constrain every fine-tuning approach, and this scarcity — rather than language mixing — was repeatedly the binding constraint (S2, S3). The PROBABLE class remains inherently ambiguous and resists both discriminative and generative treatment. Mixing archaic DE/FR/EN in tiny batches induces destructive gradient interference, as shown empirically in Phase 2. Operationally, the multi-agent S7 is expensive, requiring $\sim$30 minutes for 21

documents on three GPUs, and the Wikidata hard rule depends on sparse coverage (it fired on 0/254 dev pairs), leaving many figures uncovered. Finally, with only 254 dev pairs our per-class recall estimates carry high variance and should be read as directional rather than precise.

Future work. The most promising immediate direction is an **S5 + S7 ensemble** that fuses S5's stronger `at` predictions with S7's stronger `isAt` predictions (via majority voting or confidence-weighted fusion), which we expect to push Global MR past 0.74. Beyond this, we plan curriculum training for the specialists (easy cases before ambiguous PROBABLE cases), self-consistency decoding (multiple reasoning chains per agent with majority vote), retrieval-augmented agents (a biographical knowledge base for the Historian and a historical gazetteer for the Geographer), an iterative-debate protocol in which Historian and Geographer observe each other before the Arbiter synthesises, and domain-adaptive pre-training (continued masked-language modelling on the silver sandbox to adapt vocabulary to historical OCR before SFT).

Conclusion. We presented Team Hansel&Gretel's complete experimental journey through nine strategies for Person–Place Relation Extraction in multilingual historical documents. Our central finding is that **temporal reasoning in historical text cannot be reliably solved by discriminative classifiers alone**: generative models that produce reasoning before their labels substantially outperform classification heads ($+10.77$ points on dev), even at smaller parameter counts, and multi-agent decomposition is especially effective for the temporally-constrained `isAt` sub-task. Our best system (S5, dev Global MR $= 0.7289$) and best official submission (S9/Run 3, impresso $= 0.6221$, rank 13/46) are complementary — one optimised for in-distribution accuracy, the other combining scale with structured knowledge injection for robustness to distribution shift. The 17-point dev-to-test gap is itself an important finding: in low-resource settings, small dev sets yield optimistic proxies, and zero-shot large-model approaches may generalise better than fine-tuned smaller ones.

Acknowledgments

We sincerely thank the **HIPE-2026 organising team** — Juri Opitz, Maud Ehrmann, Simon Clematide, Corina Raclé, Emanuela Boros, Andrianos Michail, and Matteo Romanello — for the exceptional quality of the task design, data, and evaluation infrastructure. HIPE-2026 is organised within the framework of the Impresso – Media Monitoring of the Past project (Swiss NSF grant No. CRSII5_213585; Luxembourg NRF grant No. 17498891). We also thank the **IBM Research RITS team** for API access to Llama 3.3 70B, Llama 4 Maverick, Granite 3.2 8B, and GPT-OSS 120B, which enabled Strategies 8 and 9.

Declaration on Generative AI

During the preparation of this work, the authors used Grammarly in order to: grammar and spelling check. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

A. Per-Strategy Hyperparameters

Table 8 collects the training hyperparameters for the two fine-tuned LLM systems (S5 and the S7 multi-agent pipeline). Phase-1/2 configurations are described inline in Section 3; full best-trial values for the S1.5 Optuna search are in our code repository.

Table 8
Training hyperparameters for the fine-tuned LLM strategies. "—" denotes not applicable / not separately reported.

Parameter	S5 (single model)	S7 (multi-agent)
Base model	Qwen2.5-3B-Instruct	Qwen2.5-7B-Instruct
Precision	bfloat16	bfloat16
LoRA rank / α / dropout	16 / 32 / 0.05	8 / 16 / 0.1
Target modules	all linear layers	q_proj, v_proj
Trainable params	∼50M / 3.1B (1.6%)	—
TRUE/PROBABLE loss weight	—	3.0×
Batch / grad. accum (eff.)	2 / 8 (16)	1 / 16 (16)
Learning rate	2×10^{-4}, cosine + 10% warmup	2×10^{-5}, cosine + warmup
Epochs	5 (peak at epoch 4)	Historian/Geographer 2; Arbiter 3
Max sequence length	1024 train / 8192 inference	2048
Gold / Silver weight	5× / 1×	5× / 1× (15,418 pairs)
Hardware	1× RTX A5000 (24 GB)	3× RTX A5000 (24 GB)
Inference time	—	∼30 min / 21 dev docs

B. Strategy Progression

Figure 3 visualises the Global Macro Recall progression summarised in Table 4. S8/S9 are omitted because they were evaluated on the official test set only (no internal-dev score).

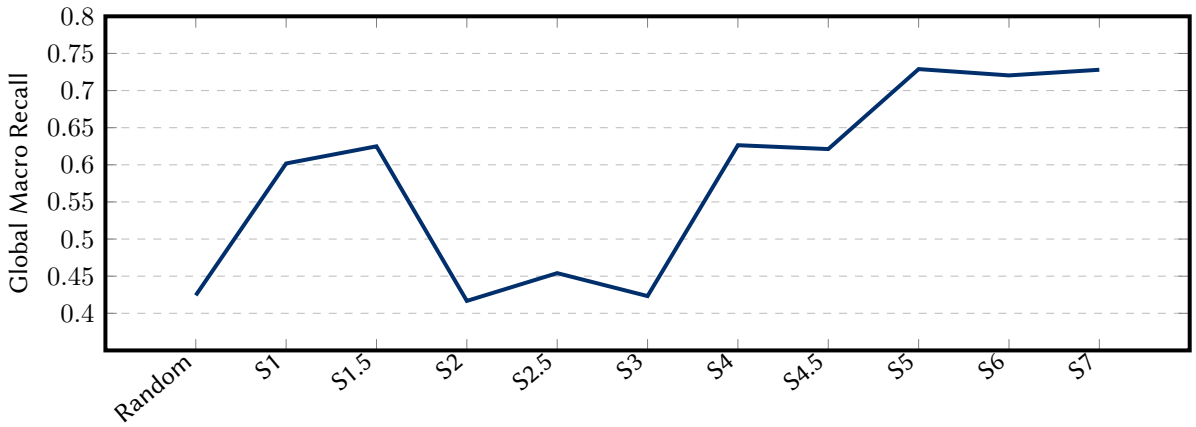

Figure 3: Global Macro Recall progression across strategies on the internal dev set. Note the drops at S2 (full fine-tuning) and S3 (monolingual partitioning) and the decisive jump at S5 (generative SFT). S8/S9 are test-only and not plotted.

References

[1] J. Opitz, M. Ehrmann, S. Clematide, C. Raclé, E. Boros, A. Michail, M. Romanello, CLEF HIPE-2026: Participation guidelines for the shared task on person–place relation extraction from historical documents, Shared task participation guidelines, 2026. URL: https://hipe-eval.github.io/HIPE-2026/. doi:10.5281/zenodo.20082076, v.2025-12-04.

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

[3] 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, pp. 354–363.

[4] M. Ehrmann, M. Romanello, A. Flückiger, S. Clematide, HIPE-2020: Named entity recognition and linking on historical newspapers, in: Advances in Information Retrieval. ECIR 2020, 2020.

[5] M. Ehrmann, A. Hamdi, E. L. Pontes, M. Romanello, A. Doucet, HIPE-2022: Findings and insights from the named entity recognition and linking shared task, in: Proceedings of the Language Resources and Evaluation Conference, 2022.

[6] 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.

[7] Impresso Project, Impresso – media monitoring of the past, Research project, University of Luxembourg, EPFL and University of Zurich, 2023. URL: https://impresso-project.ch, swiss NSF grant No. CRSII5_213585; Luxembourg NRF grant No. 17498891.

[8] J. Opitz, M. Ehrmann, S. Clematide, C. Raclé, E. Boros, A. Michail, M. Romanello, HIPE-2026 LLM baseline, 2026. URL: https://github.com/hipe-eval/HIPE-2026-llm-baseline.

[9] A. Conneau, K. Karthikeyan, G. Lample, B. Oguz, R. Rinott, Y. Wu, S. Guerin, E. Bugliarello, E. Ponti, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2020, pp. 8440–8451.

[10] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, T.-Y. Liu, LightGBM: A highly efficient gradient boosting decision tree, in: Advances in Neural Information Processing Systems, volume 30, 2017.

[11] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, IEEE Transactions on Pattern Analysis and Machine Intelligence 42 (2017) 318–327.

[12] T. Akiba, S. Sano, T. Yanase, T. Ohta, M. Koyama, Optuna: A next-generation hyperparameter optimization framework, in: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2019, pp. 2623–2631.

[13] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, LoRA: Low-rank adaptation of large language models, arXiv preprint arXiv:2106.09685 (2020).

[14] Q. Team, Qwen2.5 technical report, 2024. URL: https://arxiv.org/abs/2412.15115.

[15] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, Chain-of-thought prompting elicits reasoning in large language models, in: Advances in Neural Information Processing Systems, volume 35, 2022, pp. 24824–24837.