Scaffolding Large and Small Reasoning Models for Person-Place Relation Extraction

Anna-Katharina Dick[1,*], Jürgen Hermes[1] and Nils Reiter[1]

[1]University of Cologne, Germany

Abstract

This paper summarizes the participation of the Spinfo team of the University of Cologne in the shared task "Identifying Historical People, Places, and other Entities (HIPE)" of CLEF 2026. We explored three augmentation strategies: supplementing the model with explicit temporal information in the prompt, adding logical constraints to the model's output by conditional prompting, and translating prompts into the document language. Our approach employing the advanced reasoning model GPT-OSS 120B ranked first on the main leaderboard for accuracy and achieved a maximum mean score of 0.748 for the three languages. Our findings suggest that large reasoning models do not generally benefit from explicit temporal reasoning or prompting that enforces logical constraints, but that the performance of smaller models can be improved with these strategies, which is good news for scaling person-place relation extraction to whole corpora.

Keywords

historical documents, reasoning models, language models, prompt engineering, temporal tagging, relation extraction

1. Introduction

Mapping the location and movement patterns of individuals throughout history is an important task for many different research questions. At the same time, extracting these person–place relations in multilingual historical documents is far from trivial. Traditional co-occurrence analysis is usually not sufficient for this task, as multiple entities and places often appear together in the same sentence. Systems that tackle this task must therefore be able to make complex inferences about the text to determine whether a person was actually at a place, and if so, whether this was in the immediate temporal context of the document's publication time, or at some point in the past. This requires handling historical language variation and OCR noise, reasoning about temporal and geographical relationships, and drawing on contextual information that is often sparse or implicit. In this context, CLEF-HIPE-2026 presents the task of extracting person-place relations from historical texts in three different languages [1, 2].

Our submission to the shared task is based on a large language model and employs prompting as a strategy, sometimes also called "in-context learning" [3]. Concretely, this paper makes two contributions: (i) it documents the system we submitted to the shared task. The system achieved first rank based on overall accuracy and second among teams (fourth of 46 runs) on generalization. (ii) it presents additional ablation experiments across two model size classes (24B and 120B parameters), revealing that temporal and logical-constraint augmentation help a smaller model substantially but fail to improve — and sometimes degrade — a large reasoning model.

2. Task description

HIPE-2026 is a shared task on person–place relation extraction from multilingual historical texts. The goal is to assess whether a relation holds between a person and a place mentioned in a document, and

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

*Corresponding author.

adick7@uni-koeln.de (A. Dick); hermesj@uni-koeln.de (J. Hermes); nils.reiter@uni-koeln.de (N. Reiter)

0009-0003-3357-4422 (A. Dick); 0000-0002-8367-8073 (J. Hermes); 0000-0003-3193-6170 (N. Reiter)

Table 1
Overview of the data sets used in this paper

Data set label	Description	Language	Size (documents)
impresso (train)	Public newspaper articles from England, France, Germany, Luxembourg	English	35
		French	35
		German	34
impresso (test)	Public newspaper articles from England, France, Germany, Luxembourg	English	19
		French	19
		German	19
surprise	Literary narratives 16th-18th century	French	30

to classify that relation with respect to its temporal scope.

Participants are asked to build systems that, given a historical newspaper article and a set of candidate `(person, place)` pairs, classify each pair with respect to two relation types, `at` and `isAt`. The `at` relation is assigned one of three labels (TRUE, PROBABLE, FALSE), while `isAt` is binary (TRUE, FALSE). Both relations express whether the person is at the place; they differ in temporal scope: `at` holds if the person was at the place at any point up to the publication date, while `isAt` is restricted to roughly the month before publication. The labels are defined as follows: TRUE – strong textual evidence supports the relation; PROBABLE (`at` only) – the relation can be plausibly inferred from implicit cues; FALSE – no evidence, or the text contradicts the relation.

As an input, the systems received a JSON file with four main entries: (i) the full article text (with possible OCR errors), (ii) a list of person entities mentioned in the article, (iii) a list of place entities mentioned in the article, and (iv) article metadata, such as the document language and its publication date. The expected output is a label for each provided pair and relation-type.

For the evaluation of the shared task, two data sets were employed: "Impresso" is an in-domain test set consisting of historical newspaper articles in English, German and French. "Surprise" is an out-of-domain test set consisting of French literary texts from the 16th-18th century. The latter is used to evaluate generalization across domains. In this paper, we additionally discuss experiments on the training data, which were provided ahead of the submission deadline. The training data (similar to the test data) consists of historical newspaper articles in English, German and French. Table 1 shows an overview of the data sets.

3. System description

In this paper, we discuss experiments with three different open-weight large language models: GPT-OSS[1], Mistral 3.2[2] and Mistral 4[3]. All models are instruction-tuned, offering reasoning capabilities to varying degrees, and are run on the infrastructure of the University of Cologne. All models have been run with a temperature of $t = 1$ and $top_p = 1$.

In our prompting scheme, we treat the two relation types separately, processing `at` first and `isAt` second. For each `(person, location)`-pair in each document, the instantiated prompt template for the relation type is sent to the LLM and its response is stored. The baseline prompt templates are formulated in English and contain a short description of the respective task, explaining the possible labels that can be assigned, and giving examples of the desired output and formatting for each possible label, namely the predicted label followed by a short explanation on a single line. We also included explicit negative constraints prohibiting enumeration and extraneous formatting in addition to the positive formatting examples to ensure reliable parsing of the output. The full prompt templates can be found in Appendices A.1 and A.2.

[1] 120B parameters, 131k context window; [4]
[2] 24B parameters, 128k context window; https://docs.mistral.ai/models/model-cards/mistral-small-3-2-25-06
[3] 119B parameters, 256k context window; https://docs.mistral.ai/models/model-cards/mistral-small-4-0-26-03

For each baseline prompt, we experimented along three independent axes, each discussed in a dedicated subsection:

- a conditional prompting strategy in which the `isAt` prompt is skipped whenever the `at` prediction is not TRUE (§3.1);
- the injection of resolved temporal expressions extracted with HeidelTime, formatted as a *Date mentions* block added to the article (§3.2);
- the use of prompts written in the article's language (German, English, or French) rather than the English template (§3.3).

The two submitted runs combine these axes incrementally. Only GPT-OSS runs were submitted to the shared task: **Run 1** (baseline) uses the English prompt with neither temporal augmentation nor conditional prompting. **Run 2** adds the temporal block and the conditional rule on top of Run 1. We also tested prompt language adaptation as a new contribution in this paper.[4]

3.1. Conditional prompting strategy

The two tasks to be solved in the shared task are not independent. For any given article, the set of TRUE pairs `(person,location)` for the relation `isAt` is a subset of the set for the relation `at`– if a person is presently at a location, they were necessarily also at that location at some point in time.[5] This interdependence is encoded into our prompting regime as follows:

The model is first presented only with the `at` prompt for a given (person, location) pair. If the model replies FALSE or PROBABLE, the pair's `isAt` label is set to FALSE automatically; only pairs for which the model returns `at = TRUE` are subsequently classified with the `isAt` prompt. We hypothesised that enforcing this rule a priori would reduce logically inconsistent predictions and bias the `isAt` classifier away from over-predicting TRUE on pairs that are not even plausibly co-located in the article.

Our implementation goes one step beyond the minimum consistency constraint stipulated by the guidelines: while these forbid only `at=FALSE` together with `isAt=TRUE`, we additionally collapse `at=PROBABLE` cases to `isAt=FALSE`. We adopted this tighter form on the assumption that a PROBABLE verdict indicates evidence too tentative to support the narrower within-publication-horizon claim of `isAt`. This choice accepts a trade-off – it suppresses any `isAt=TRUE` prediction on pairs whose textual support is suggestive but not unambiguous – whose empirical consequences are reflected in the ablation results §4.

3.2. Temporal augmentation with HeidelTime

The `isAt` sub-task is, by design, temporally bounded: a TRUE label requires that the article place the person at the location within roughly one month before publication (HIPE-2026 Guidelines §2.2). Off-the-shelf large language models can read date expressions out of running text, but reasoning about the *distance* from the document creation time (DCT) and filtering candidate evidence to the one-month window is precisely the kind of operation on which even frontier LLMs falter: Fatemi et al. [5] report an average accuracy of 15.4 % across Claude-3-Sonnet, Mistral Large, GPT-4, Gemini 1.5 Pro and GPT-4o on the "Duration" subset of their Test-of-Time benchmark, with common failure modes including miscounting months and confusing the sign of the elapsed interval. This operation corresponds to the lowest level ("time-time") of the temporal reasoning taxonomy of Tan et al. [6]. We hypothesised that offloading temporal normalisation to a deterministic, rule-based component could close this gap and make the temporal constraint of `isAt` explicit in the prompt.

We use **HeidelTime**[6] [7], a rule-based, multilingual temporal tagger that produces TIMEX3 an-

[4]We originally submitted a third run to the shared task, using prompt translations for French and German. Due to missing metadata for the configuration of this specific run we could not reconstruct with full certainty if the model had access to the temporal information at time of inference. We therefore do not discuss these results in our paper and instead refer the reader to our ablation experiments using the same data set in §4.3.

[5]This is made explicit in the task definition: "[i]f `at` is FALSE, then assigning TRUE to `isAt` would be inconsistent" (HIPE-2026 Guidelines §2.2)

[6]Version HeidelTime 2.2.1, integrated via JitPack.

```
Publication date: 1910-09-16
Date mentions within the isAt window (-35 to +1 days around publication):
  - "Thursday"       -> 1910-09-15 (1 day before publication)
  - "Saturday night" -> 1910-09-10 (6 days before publication)
  - "last week"      -> 1910-09-05 (11 days before publication)

If no listed date plausibly links the (person, location) pair to a moment
within the isAt window, isAt is likely FALSE.
```

Figure 1: Concrete example illustrating the added date information. Example is showing the English document `sn88085488-1910-09-16-a-i0002` with DCT `1910-09-16`.

notations in TimeML format and resolves relative expressions ("yesterday", "letzten Monat", "il y a deux jours") against a supplied DCT. HeidelTime ships language-specific pattern resources for German, English and French – the three languages of the HIPE-2026 main test set – and emits ISO-style values that are directly comparable to a document's `date` field.

We run HeidelTime over every training, development, and test document, using each document's `date` field as the DCT. The extracted TIMEX3 spans are written back as a new `time_expressions` array on the same JSONL document; character offsets are preserved, so the existing person and location annotations stay aligned. On the official Test A sets this yielded 131, 60, and 157 TIMEX3 spans in the German, English, and French newspaper subsets respectively (348 in total across 57 newspaper documents).[7] The tagged corpus itself thus serves as the interface between our temporal-tagging component and the downstream LLM classifier: any consumer that reads the JSONL gains the additional field for free, while consumers that ignore the new field see exactly the original document structure.

At classification time we format the `time_expressions` of each document as a compact "Date mentions" block appended to the article text, before the person-place pair is presented. For the `isAt` prompt we keep only spans whose resolved ISO value falls within $[\mathrm{DCT} - 35\,\mathrm{d},\ \mathrm{DCT} + 1\,\mathrm{d}]$. For the broader `at` prompt we keep all resolvable spans but separate them into anchors on or before the DCT and post-DCT references; the latter are surfaced as explicit counter-evidence, since the `at` relation is right-bounded by the publication date. We extend the specification's "about one month" lower bound by a five-day slack to absorb OCR-induced date errors and HeidelTime resolution variance, and tolerate one day past the DCT to cover publication-day mentions and off-by-one cases. When the filtered set is empty, the block is omitted entirely, so prompts on those documents are identical to the baseline. Each line in a non-empty block surfaces the resolved date, the original surface form, and the signed delta to the publication date, so the LLM can verify temporal evidence against the (person, location) pair without performing the normalisation itself.

Figure 1 shows a concrete example to illustrate the structure. The closing instruction makes the negative case explicit: an empty or pair-irrelevant block should bias the model toward `FALSE`.

3.3. Target language prompting

There is some evidence that formulating prompts in the target language can achieve good results in tasks related to natural language understanding and inference (NLU/NLI) [8], given that the model has been sufficiently trained in these languages. For culture-related tasks that need deep language understanding, prompting in the native language has also been shown to be more effective than English as it better captures the nuances of culture and language [9]. Since NLU is an integral part of the given task and it is possible that there are some examples where a more nuanced cultural understanding of the historical source texts would be beneficial, we tested configurations where the `isAt` and `at` prompts are translated to the respective languages of the documents.

[7] Test B (the literary surprise set) evaluates only the `at` sub-task. As our temporal augmentation is primarily motivated by the one-month window of `isAt`, and as the literary documents carry year-only publication dates (e.g., `"1756"`) which our strict DCT parser does not accept, we did not apply temporal augmentation to Test B.

We used DeepL[8] to translate the prompts from English into German and French, and manually corrected them (e.g., changing 'FALSCH' to 'FALSE' in the description of the labels). The few-shot example predictions were not corrected to English, causing the models to give their reasoning in the document language in most cases. Inconsistent labeling occasionally caused some parsing issues with the smaller models, where prediction labels were erroneously given in the document language. These cases were manually corrected to the corresponding English labels.

The appended block with temporal information in the `isAt` prompt was not translated in configurations with temporal augmentation.

3.4. Experimental setup

This paper reports results on three lines of experiments: (i) First, we document and discuss **shared task results** in §4.1. The quantitative results are also documented on the shared task web page and the general paper. (ii) We discuss experiments conducted before the shared task deadline, on **comparing model sizes** using the supplied training data. Specifically, we compare all three models mentioned above: GPT-OSS, Mistral 3.2 and Mistral 4, but only use the base prompt. Results are discussed in §4.2. (iii) We present results on a **full factorial ablation** on Mistral 3.2 and GPT-OSS: We evaluated each of the three axes (temporal augmentation, conditional prompting, target-language prompts) individually, in every pairwise combination, and all three together – eight configurations per model. We ran this $2 \times 2 \times 2$ grid once per model. The goal of this experiment is to isolate the independent contribution of each axis and to gauge the accuracy cost of moving to a less resource-intensive model. We report results of this ablation in §4.3.

4. Results

All results were obtained using the official HIPE-2026 scorer[9] and use the HIPE evaluation metrics specified on the website[10]. We report accuracy (Acc) and macro recall (MR) for both tasks on the impresso data sets, and the same metrics for only the `at`-task for the surprise data set. The main metric we use is global macro recall or impresso score, calculated as the mean of the macro recall for `isAt` and `at` tasks for a specific language and run. The overall accuracy profile score (which, despite its name, is computed from macro-recall scores) is the mean across the three languages.

For the generalization profile, the evaluation score is the macro recall score of the `at`-task on the surprise file. The efficiency profile score is calculated through averaging the accuracy profile score ranking, model parameter count rank, and model size rank. As our system was mainly designed to compete in the accuracy profile, we focus on these results.

4.1. Shared task results

We submitted runs for each of the four test sets (German, English, French, Surprise). The results can be found in table 2 and on the HIPE-2026 website. There were 46 total runs for most configurations, including a random baseline and a baseline run with Ministral-3-3B-Instruct GGUF by the organizers. One team further only submitted results for English, and one team opted their three runs out of the efficiency ranking. Our first run, using the standard method of separate prompting in English with no additional temporal information, achieved a mean impresso profile score of 74.8 % across all languages and ranked first of 46 runs on the leaderboard for the accuracy profile.

On the German test set, run 1 achieved an impresso profile score of 76.1 %, ranking second – just behind our own run 3[11]. Run 1 ranked first on the French test set with a score of 75.5 %. On the English

[8] https://www.deepl.com/en/translator

[9] https://github.com/hipe-eval/HIPE-2026-data/tree/main/scripts

[10] https://hipe-eval.github.io/HIPE-2026/HIPE_2026_evaluation_results/

[11] We only report Runs 1 and 2 in our tables due to the reproducibility issues with run 3. See table 6 for later ablation experiments using translated prompts.

Table 2
Submitted runs and leaderboard results per test set

Test Set	Run and Configuration	Impresso Score	Ranking	at MR	at Acc	isAt MR	isAt Acc
German	Run 1: base model	0.7608	2/46	0.7017	0.8067	0.8199	0.8782
	Run 2: time + conditional	0.6860	5/46	0.6659	0.7647	0.7060	0.8319
English	Run 1: base model	0.7279	2/47	0.6560	0.6667	0.7998	0.8272
	Run 2: time + conditional	0.6427	8/47	0.6215	0.6605	0.664	0.7222
French	Run 1: base model	0.7551	1/46	0.6785	0.7815	0.8318	0.8529
	Run 2: time + conditional	0.7383	2/46	0.6774	0.7773	0.7991	0.8571
Surprise	Run 1: base model	-	5/46	0.691	0.7729	-	-
	Run 2: time + conditional	-	7/46	0.6674	0.7458	-	-

test set, run 1 achieved a score of 72.8 %, placing second. For the generalization profile, the base model results scored 69.1 %, ranking 5th out of 46, the augmented run scored 66.7 %, ranking 7th.

Since our model was among the biggest used in the competition, our base model only placed 20 of 43 runs in the official overall efficiency profile score, but 8th [12] in the balanced efficiency profile ranking, which gives equal weight to ranks based on performance (accuracy) and model power (parameter count, model size) instead of an unweighted average of the three.[13]

4.2. Model size experiments on training set

In our preliminary experiments on the labeled training sets using just the base model configuration (model specifications in table 3, results in table 4), GPT-OSS consistently outperformed the Mistral models by a substantial margin across all languages and metrics. On the German training data set, GPT-OSS achieves a global macro recall of 75.83 %, compared to 59.24 % for Mistral 3.2 and 53.7 % for Mistral 4. The same pattern holds for English and French, where GPT-OSS reaches 73.89 % and 70.37 % global macro recall respectively, while both Mistral variants remain below 62 % in both languages. Notably, the performance gap is most pronounced on the at-task: GPT-OSS outperforms the strongest Mistral model by at least 10 and up to 26 percentage points in all metrics and languages.

Table 3
Model comparison

Model	Parameters	Context Window
Mistral 3.2	24B	128k tokens
Mistral 4	119B	256k tokens
GPT-OSS	120B	131k tokens

4.3. Ablation experiments on test set

After the release of the labeled test set we performed ablation experiments with two models to investigate the effect of each of the components of our system. Table 5 shows these results for the substantially smaller Mistral 3.2, table 6 shows the results using GPT-OSS (including the results from the submitted runs).

As expected, the larger GPT-OSS consistently outperforms Mistral 3.2 by a substantial margin across all languages, with GPT-OSS achieving global macro recall scores of up to 0.777 on German, 0.728 on English, and 0.791 on French, compared to Mistral 3.2's best scores of 0.622, 0.647, and 0.65 respectively.

Again, at proved more difficult than isAt for both models, as both models achieved higher accuracy and macro recall for the latter task in the baseline and most augmented configurations.

[12] Tied with another team's submitted run.
[13] See https://github.com/hipe-eval/hipe-2026-eval for details.

Table 4
Base model results on training set

Language	Model	at MR	at Acc	isAt MR	isAt Acc	Global MR
German	Mistral 3.2	0.502	0.537	0.689	0.549	0.592
	Mistral 4	0.435	0.449	0.64	0.618	0.537
	GPT-OSS	0.701	0.794	0.824	0.893	0.758
English	Mistral 3.2	0.475	0.577	0.707	0.606	0.591
	Mistral 4	0.443	0.469	0.685	0.612	0.564
	GPT-OSS	0.645	0.814	0.833	0.818	0.739
French	Mistral 3.2	0.487	0.609	0.736	0.621	0.612
	Mistral 4	0.459	0.502	0.718	0.701	0.588
	GPT-OSS	0.597	0.799	0.81	0.86	0.704

The results further show that Mistral 3.2 can generally profit from the additional information provided by HeidelTime and the logical constraints from conditional prompting: Both strategies improve the global macro recall score above the base model score in all but one case (french test set with french prompting and added temporal information, which had a negligible negative effect). Combining both strategies did not further improve the results however, and the combination even had a negative effect on the English test set compared to the base model. Translating prompts for non-English languages also had a generally negative effect on performance with this model: While it seemingly improved accuracy scores for the base model predictions, macro recall decreased substantially for the `isAt` prediction task.

Our augmentation methods had a more mixed effect on the performance of GPT-OSS: The addition of temporal information did not show a clear pattern of improvement or decline, and employing the conditional prompting strategy worsened the prediction results compared to the base model predictions. Similarly to Mistral 3.2, mixing both strategies also did not outperform either strategy in isolation.

Our choice to collapse at=PROBABLE to isAt=FALSE is likely a factor in the score degradation for the GPT model using conditional prompting (table 2 and 6). In the German test set for example, there are 25 entity-place pairs with at=PROBABLE/isAt=TRUE that the model is prevented from predicting correctly, regardless of its underlying reasoning quality. The smaller Mistral model, by contrast, is less affected: its overall lower performance corresponds to a much lower propensity to predict PROBABLE at all, so the collapsed label rarely comes into conflict with its predictions.

5. Error analysis

Even though GPT-OSS can use its reasoning skills effectively to resolve ambiguities that smaller models struggle with, it still makes mistakes. We highlight some errors that have been consistently made during our internal testing phase, that hint at systematic weaknesses of the model.

Multi-step implicit reasoning: In one example in the German training data, the article describes Austrian sports players who competed against Czechoslovakia in matches held in Vienna and Bratislava. There are multiple entity pairs that the system predicts incorrectly as `FALSE` for both `isAt` and `at`, since it cannot make the inference that, e.g., "centre half player Gewirk" refers to a member of this team that necessarily was present at the games, and thus is in Czechoslovakia.

In other cases (taken from the English training data), the model gets misled by context clues. *"DIXIELAND JAZZ GETS NATION'S EAR AGAIN NEW YORK.—All over the na tion. Dixieland Jazz is hot again, is the disclosure of Joseph Roddy in the June fl issue of Look Mag azine.[…]"* Since "New York" is only mentioned as the dateline in the article header, there is no indication that Joseph Roddy was actually in New York himself. The model confuses the origin of the article itself with the location of the mentioned entity and places Roddy in New York as well, instead of ignoring this context clue as journalistic convention.

Another mistake that exemplifies the importance of clear prompt instructions, also from the English

Table 5
Mistral 3.2 results on test set by language, prompt language, and variant

Lang	Prompt	Configuration	at MR	at Acc	isAt MR	isAt Acc	Global MR
German	English	baseline	0.364	0.466	0.69	0.639	0.527
		+time	0.532	0.504	0.713	0.685	0.622
		+conditional	0.467	0.466	0.716	0.748	0.592
		+time+conditional	0.440	0.433	0.736	0.777	0.588
	German	baseline	0.445	0.466	0.485	0.706	0.465
		+time	0.456	0.462	0.714	0.706	0.585
		+conditional	0.49	0.513	0.728	0.752	0.609
		+time+conditional	0.513	0.521	0.698	0.748	0.605
English	English	baseline	0.502	0.562	0.686	0.642	0.594
		+time	0.553	0.611	0.709	0.691	0.631
		+conditional	0.51	0.605	0.784	0.778	0.647
		+time+conditional	0.451	0.556	0.679	0.685	0.565
French	English	baseline	0.341	0.571	0.709	0.664	0.525
		+time	0.505	0.597	0.776	0.752	0.64
		+conditional	0.519	0.609	0.78	0.769	0.65
		+time+conditional	0.489	0.576	0.79	0.798	0.64
	French	baseline	0.385	0.622	0.491	0.727	0.438
		+time	0.403	0.635	0.469	0.727	0.436
		+conditional	0.486	0.605	0.713	0.744	0.6
		+time+conditional	0.479	0.605	0.693	0.752	0.586

Table 6
GPT-OSS results on test set by language, prompt language, and variant

Lang	Prompt	Configuration	at MR	at Acc	isAt MR	isAt Acc	Global MR
German	English	baseline	0.702	0.807	0.82	0.878	0.761
		+time	0.724	0.803	0.83	0.887	0.777
		+conditional	0.687	0.782	0.667	0.803	0.677
		+time+conditional	0.666	0.765	0.706	0.832	0.686
	German	baseline	0.688	0.815	0.829	0.878	0.759
		+time	0.683	0.819	0.513	0.845	0.598
		+conditional	0.686	0.807	0.781	0.861	0.734
		+time+conditional	0.689	0.811	0.739	0.84	0.714
English	English	baseline	0.656	0.667	0.8	0.827	0.728
		+time	0.622	0.642	0.702	0.741	0.662
		+conditional	0.576	0.611	0.705	0.759	0.641
		+time+conditional	0.622	0.661	0.664	0.722	0.643
French	English	baseline	0.679	0.782	0.832	0.853	0.755
		+time	0.654	0.765	0.854	0.878	0.754
		+conditional	0.642	0.752	0.737	0.811	0.689
		+time+conditional	0.677	0.777	0.799	0.857	0.738
	French	baseline	0.654	0.79	0.835	0.866	0.745
		+time	0.688	0.819	0.894	0.908	0.791
		+conditional	0.695	0.811	0.771	0.828	0.733
		+time+conditional	0.642	0.773	0.804	0.848	0.723

training data, is *"[…] March 30. Capt. Burger, osthe ship John and Ed- ward, left Lisbon on the 5ih. ult. […]"*. This should be a straightforward case of Capt. Burger being present at Lisbon in the past, but not

at time of publication. Nonetheless, the model claims that the person was present at the place at time of publication, stating "Capt. B. was recorded as leaving Lisbon on the 5th, placing him in Lisbon within the month before the March. 30 report." The system therefore overapplied the instruction it was given where `isAt` should be set as TRUE if the person was present at the location for up to one month prior to publication. Notably, HeidelTime did not recognise the archaic date expression "the 5th ult." (the fifth of the preceding month) as a temporal expression, owing to both the OCR corruption ("5ih. ult.") and the uncommon "ult." abbreviation. The *Date mentions* block for this document therefore contained only the dateline and two bare month names, but not the date the model actually reasoned about. Temporal augmentation could thus not have corrected this error, and the case illustrates a structural limitation of rule-based temporal tagging on noisy historical text: the archaic and OCR-corrupted date forms most characteristic of this domain are also the ones most likely to be missed.

6. Discussion

Our system's first-place finish on the accuracy profile of the impresso newspaper test set, achieved with the simplest configuration we submitted – Run 1, without temporal augmentation or conditional prompting – underscores that a sufficiently large reasoning model can already shoulder most of the inferential burden of the task without further scaffolding. That ranking, however, comes at a measurable efficiency cost: on the efficiency profile the same submission placed 20 of 44 runs. The accuracy gain of using GPT-OSS is therefore not free, and teams prioritising resource frugality may reasonably favour the substantially smaller models we explored in our ablation.

The post-hoc experiments complicate any naive "more is more" reading of the leaderboard along two axes. First, raw parameter count is not a reliable predictor of model quality even within the same family: the 24-billion-parameter Mistral 3.2 consistently outperformed its nominally larger 119-billion-parameter sibling Mistral 4 across all three training-set languages, by 2 to 6 macro-recall points on the global score. Training-data and architectural choices appear to matter at least as much as raw parameter count for downstream performance (cf. 10 for an analogous argument from the training-compute side).

Second, and more surprisingly, the two non-trivial methods we explored – temporal augmentation with HeidelTime and conditional prompting – fail to improve GPT-OSS and, in several configurations, degrade it substantially: applied jointly to the baseline, they cost up to nine points on the global macro recall (English configuration in table 6). The same interventions, applied to Mistral 3.2, produced sizeable gains, in some configurations above fourteen macro-recall points on the global score (German and French with same-language prompts in table 5). The size-dependence of augmentation benefits has been documented in two opposite directions in the literature: retrieval augmentation disproportionately helps smaller models in the 1 B–20 B range [11], while chain-of-thought prompting becomes beneficial only at sufficient scale [12]. Our findings – limited to a two-model comparison between Mistral 3.2 (24 B) and GPT-OSS (120 B) – are consistent with the first pattern at a larger scale: structural scaffolding appears to help the smaller of the two models substantially, while it ceases to help (and in places hurts) the larger one. A broader comparison across more intermediate sizes would be needed to establish whether this is a robust continuation of the smaller-scale trend.

7. Future work

We envision multiple future directions for improving our system. (i) While grid search over all possible parameters will be impossible, covering more parameter settings with experiments will be important. These parameters include temperature, top_k and top_p values, but also employment of system prompts. (ii) Our conditional prompting strategy enforced a stricter rule than the task definition requires, collapsing not only `at=FALSE` but also `at=PROBABLE` predictions to `isAt=FALSE` (§3.1). Testing the spec-compliant variant – propagating only `at=FALSE` to `isAt=FALSE` and leaving PROBABLE pairs to the model – is a low-cost next step that may recover part of the performance lost to the conditional strategy. (iii) The temporal augmentation strategy was essentially to supply the LLM with structured,

symbolic information that a more specialized system extracts from the document. In our experiment, this symbolic information was about dates mentioned in the articles, but we see it as a fruitful line of research to extend this idea to other kinds of information, such as places (potentially geographically resolved to allow more robust geographic reasoning) or names (potentially linked or enriched with additional information taken from databases as Wikidata). Supplying this information may – to a certain extent – counterbalance the performance issues that smaller models have. (iv) The continued use of smaller models is important for scaling up this extraction. Processing 19 documents is feasible with a huge model, but processing an entire corpus of newspapers will take prohibitively more time. For such uses, smaller models are essential. Another way of improving the capabilities of smaller models (and to arrive at efficient corpus-level processing) is to fine-tune them to specific use cases. Our system does not make use of the training data at all, but it would be a natural step to explore.

Acknowledgments

This research was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project number 550309711. We also thank the University of Cologne for provisioning and maintaining a capable technical infrastructure to run such experiments.

Individual contributions

According to the Contributor Role Taxonomy[14].

- Anna Dick: Data curation, formal analysis, software, investigation, validation, writing – original draft, writing – review & editing
- Jürgen Hermes: Data curation, conceptualization, software, methodology, writing – original draft, writing – review & editing
- Nils Reiter: Conceptualization, funding acquisition, supervision, writing – original draft, writing – review & editing

Declaration on Generative AI

During the preparation of this work, we used Claude Opus 4.7 for text translation, paraphrasing and rewording, grammar and spelling checks, and peer review simulations in some parts of the text; as well as formatting tables 4, 5 and 6. Experiments make extensive use of generative AI, as detailed above. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] 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.

[2] 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.

[14] https://credit.niso.org

[3] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, G. Neubig, Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing, ACM Comput. Surv. 55 (2023). doi:10.1145/3560815, place: New York, NY, USA.

[4] OpenAI, S. Agarwal, L. Ahmad, J. Ai, S. Altman, A. Applebaum, E. Arbus, R. K. Arora, Y. Bai, B. Baker, H. Bao, B. Barak, A. Bennett, T. Bertao, N. Brett, E. Brevdo, G. Brockman, S. Bubeck, C. Chang, K. Chen, M. Chen, E. Cheung, A. Clark, D. Cook, M. Dukhan, C. Dvorak, K. Fives, V. Fomenko, T. Garipov, K. Georgiev, M. Glaese, T. Gogineni, A. Goucher, L. Gross, K. G. Guzman, J. Hallman, J. Hehir, J. Heidecke, A. Helyar, H. Hu, R. Huet, J. Huh, S. Jain, Z. Johnson, C. Koch, I. Kofman, D. Kundel, J. Kwon, V. Kyrylov, E. Y. Le, G. Leclerc, J. P. Lennon, S. Lessans, M. Lezcano-Casado, Y. Li, Z. Li, J. Lin, J. Liss, Lily, Liu, J. Liu, K. Lu, C. Lu, Z. Martinovic, L. McCallum, J. McGrath, S. McKinney, A. McLaughlin, S. Mei, S. Mostovoy, T. Mu, G. Myles, A. Neitz, A. Nichol, J. Pachocki, A. Paino, D. Palmie, A. Pantuliano, G. Parascandolo, J. Park, L. Pathak, C. Paz, L. Peran, D. Pimenov, M. Pokrass, E. Proehl, H. Qiu, G. Raila, F. Raso, H. Ren, K. Richardson, D. Robinson, B. Rotsted, H. Salman, S. Sanjeev, M. Schwarzer, D. Sculley, H. Sikchi, K. Simon, K. Singhal, Y. Song, D. Stuckey, Z. Sun, P. Tillet, S. Toizer, F. Tsimpourlas, N. Vyas, E. Wallace, X. Wang, M. Wang, O. Watkins, K. Weil, A. Wendling, K. Whinnery, C. Whitney, H. Wong, L. Yang, Y. Yang, M. Yasunaga, K. Ying, W. Zaremba, W. Zhan, C. Zhang, B. Zhang, E. Zhang, S. Zhao, gpt-oss-120b & gpt-oss-20b model card, 2025. URL: https://arxiv.org/abs/2508.10925. arXiv:2508.10925.

[5] B. Fatemi, S. M. Kazemi, A. Tsitsulin, K. Malkan, J. Yim, J. Palowitch, S. Seo, J. Halcrow, B. Perozzi, Test of Time: A Benchmark for Evaluating LLMs on Temporal Reasoning, International Conference on Learning Representations 2025 (2025) 94426–94447. URL: https://proceedings.iclr.cc/paper_files/paper/2025/hash/eb7295a8bc613b375726659c2ecd6f14-Abstract-Conference.html.

[6] Q. Tan, H. T. Ng, L. Bing, Towards benchmarking and improving the temporal reasoning capability of large language models, in: A. Rogers, J. Boyd-Graber, N. Okazaki (Eds.), Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 14820–14835. doi:10.18653/v1/2023.acl-long.828.

[7] J. Strötgen, M. Gertz, Multilingual and cross-domain temporal tagging, Language Resources and Evaluation 47 (2013) 269–298. doi:10.1007/s10579-012-9179-y.

[8] M. Zhao, H. Schütze, Discrete and soft prompting for multilingual models, in: M.-F. Moens, X. Huang, L. Specia, S. W.-t. Yih (Eds.), Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, pp. 8547–8555. doi:10.18653/v1/2021.emnlp-main.672.

[9] C. Liu, W. Zhang, Y. Zhao, L. A. Tuan, L. Bing, Is translation all you need? a study on solving multilingual tasks with large language models, in: 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), 2025, pp. 9594–9614.

[10] J. Hoffmann, S. Borgeaud, A. Mensch, E. Buchatskaya, T. Cai, E. Rutherford, D. de Las Casas, L. A. Hendricks, J. Welbl, A. Clark, T. Hennigan, E. Noland, K. Millican, G. van den Driessche, B. Damoc, A. Guy, S. Osindero, K. Simonyan, E. Elsen, O. Vinyals, J. Rae, L. Sifre, An empirical analysis of compute-optimal large language model training, in: S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, A. Oh (Eds.), Advances in Neural Information Processing Systems, volume 35, Curran Associates, Inc., 2022, pp. 30016–30030. URL: https://proceedings.neurips.cc/paper_files/paper/2022/file/c1e2faff6f588870935f114ebe04a3e5-Paper-Conference.pdf.

[11] A. Mallen, A. Asai, V. Zhong, R. Das, D. Khashabi, H. Hajishirzi, When Not to Trust Language Models: Investigating Effectiveness of Parametric and Non-Parametric Memories, in: A. Rogers, J. Boyd-Graber, N. Okazaki (Eds.), Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 9802–9822. doi:10.18653/v1/2023.acl-long.546.

[12] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. H. Chi, Q. V. Le, D. Zhou, Chain-of-thought prompting elicits reasoning in large language models, in: Proceedings of the 36th International Conference on Neural Information Processing Systems, NIPS '22, Curran Associates

Inc., Red Hook, NY, USA, 2022, pp. 24824–24837.

A. Prompt templates

A.1. Base prompt for `at`

```
Read the following text.
In the next step I will give you pairs of persons and locations.

EXPLICIT RULES:
Respond for each pair with 'TRUE', 'FALSE' or 'PROBABLE', depending on if there are
    explicit or implicit indications in the text that the mentioned person has been
    present at the specified location at some point in the past.
For each pair respond only with one line: The label and a short explanation.

For example (DO):
FALSE There is no indication that this person was in Rome before.
TRUE There is an indication that this person was in Rome before.
PROBABLE It is likely that this person was in Rome at some point, and this can be
    inferred from implicit cues in the context.

DON'Ts:
1. DO NOT Add numbers to enumerate the text
2. DO NOT add formatting to stray from the above example format.

{Document text}

Pair: {Entity/Location pair}
```

A.2. Base Prompt for `isAt`

```
Read the following text.
In the next step I will give you pairs of persons and locations.

EXPLICIT RULES:
Respond for each pair with 'TRUE' or 'FALSE', depending on if there are indications in
    the text
that the mentioned person is present at the specified location at the time of writing,
    or up to a month before publication. If the person has been at the location before
    this window, answer with 'FALSE'.
For each pair respond only with one line: The label and a short explanation.

For example (DO):
FALSE There is no indication that this person is in Rome.
TRUE There is an indication that this person was in Rome at this time.

DON'Ts:
1. DO NOT Add numbers to enumerate the text
2. DO NOT add formatting to stray from the above example format.

{Document text (with added time block in runs with temporal scaffolding)}

Pair: {Entity/Location pair}
```