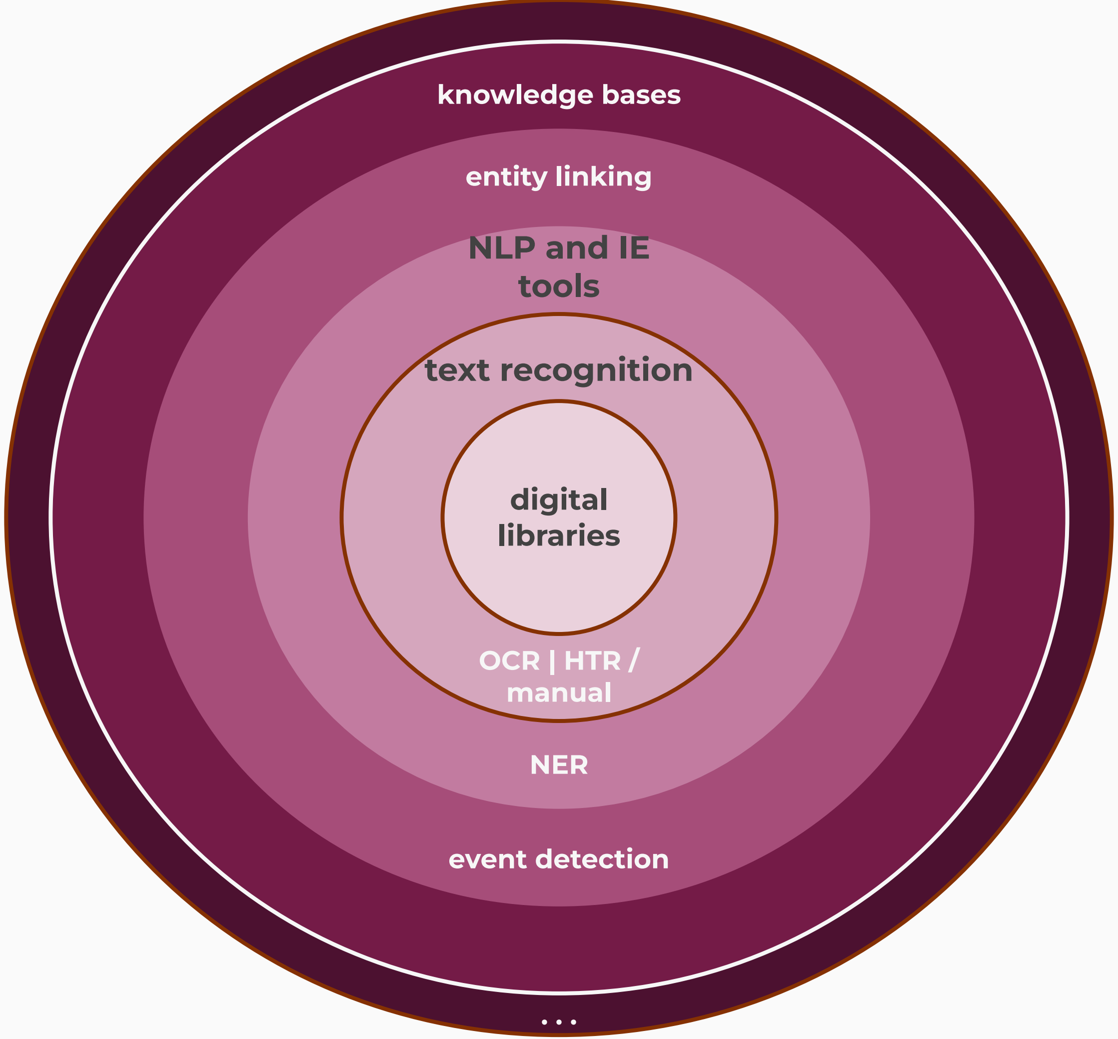
Leveraging Open Large Language Models for Historical Named Entity Recognition

Carlos-Emiliano González-Gallardo [1], Hanh Thi Hong Tran [2,3,4],
Ahmed Hamdi [2] & Antoine Doucet [2]

[1] Université de Tours, France
[2] Université de la Rochelle, France
[3] Jožef Stefan International Postgraduate School, Slovenia
[4] Jožef Stefan Institute, Slovenia

TPDL 26th September 2024

Enriching digital repositories



However... problems and challenges

PRICE: Paris and France, 20c.; Abroad, 30c. EUROPEAN EDITION—PARIS. FRIDAY

REPORT 200 LOST IN IOLAIRE WRECK OFF STORNOWAY

Survivors of Ill-Fated Converted Yacht Have Thrilling Experiences Before Reaching Land.

(SPECIAL TELEGRAM TO THE HERALD.)

LONDON, Thursday.—The wrecking of H.M.S. Iolaire on the rocks at the entrance to Stornoway Harbor took place about two in the morning in the night of Tuesday to Wednesday and the loss of life is reported to exceed 200. The Admiralty has announced, however, that there is no reason to believe the loss of life to be so heavy as this.

The Iolaire, which was the former steam yacht Amalthea, with about 250 naval officers and men on leave, had

NORTHERN PACIFIC, TRANSPORT STRANDED OFF FIRE ISLAND, IS NOW BEING SUCCORED BY A FLEET OF CRUISERS AND LIFEBOATS

Home-Going Transport, with Many Wounded Men Aboard, Goes Aground While Fighting Toward New York Through Gale.

(BY SPECIAL CABLE TO THE HERALD.)

NEW YORK, Thursday.—The Northern Pacific, the large transport which grounded early yesterday morning off Fire Island, has not yet been refloated, but it is hoped that the relief measures now being taken will prove absolutely adequate.

The Northern Pacific, which left Brest on Christmas Eve, transporting 2,451 returning soldiers, (1,671 of whom were

CHART SHOWING SPOT ✠ WHERE TRANSPORT GROUNDED.

Document related

- Documents deteriorated by **time**
- Poor **quality** of printing materials & digitization
- Inaccurate **scanning** processes
- **Language** change & evolution

NER related

- Inconsistency of annotation **guidelines**
- Entity **complexity**
- **Multilingualism**
- code-switching

NER in Historical Corpora

Symbolic systems

- 2000s
- gazetteer lookup
- rule incremental application
- variant matching
- normalization rules

Machine learning models

- annotated training data
- manually selected features
- CRF
- CRF + gazetteer

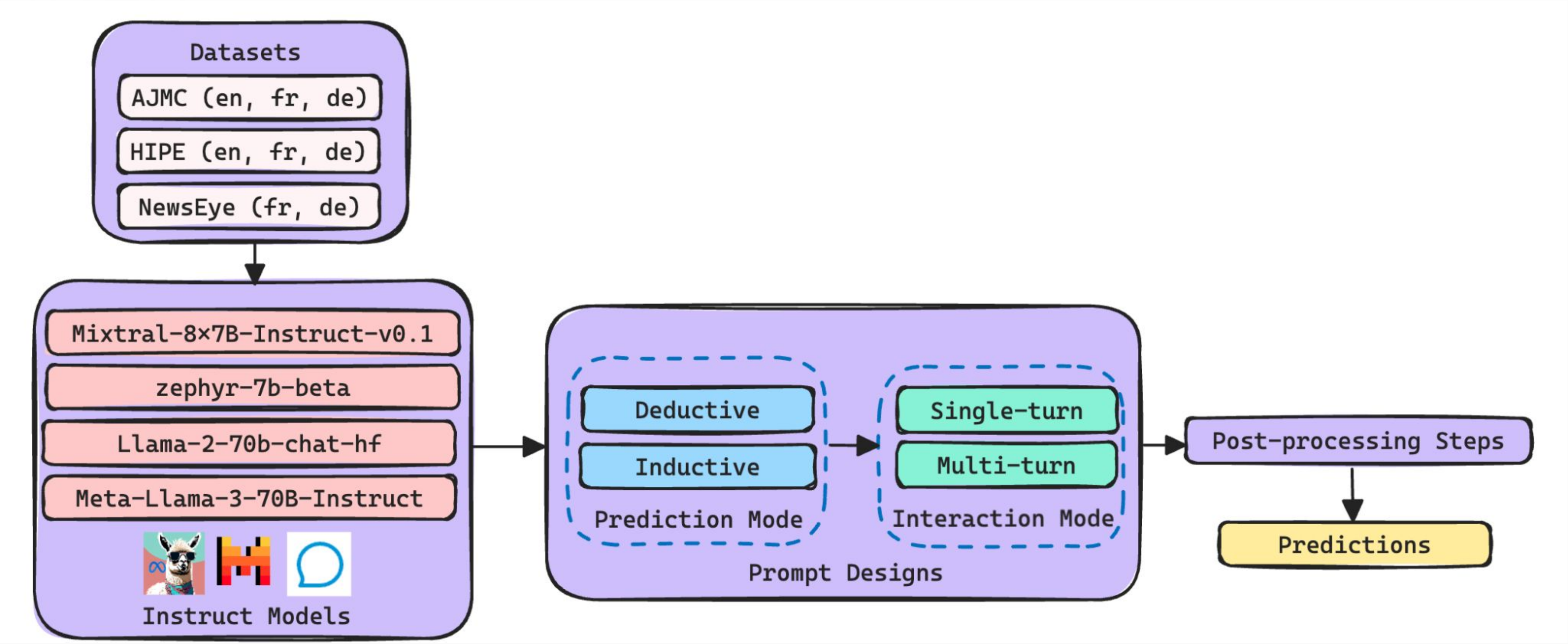
Deep learning models

- static, char-, (sub)word-level contextual, stacked embeddings
- transfer learning
- data augmentation
- context enrichtment

LLMs models

- in-context learning
- API-accesible
- Computer-assisted NER
- Digital libraries processing ad-hoc pipelines

Few-shot Prompting for Historical NER



General workflow for few-shot prompting HNER

Historical document datasets

Greek tragedy written in the 5th century BCE by Sophocles

CLEF-HIPE-2022 evaluation lab corpora

Classic commentaries (19C)

- AJMC (en, fr, de)

Historical newspapers (19C-20C)

- HIPE-2020 (en, fr, de)
- NewsEye (fr, de)

NER annotations

Universal (person, location, organization, time, human production)

Domain-specific (bibliographic references)

Vingt-Deuxième Année. — N° 7848

SIX PAGES — Paris et départements — CINQ CENTIMES

Lundi 21 Août 1905

Le Matin

DERNIERS TÉLÉGRAMMES DE LA NUIT

Stéphane LAUZANNE Rédacteur en chef.

Jules MADELINE Administrateur.

PRICE: Paris and France, 20c.; Abroad, 30c.

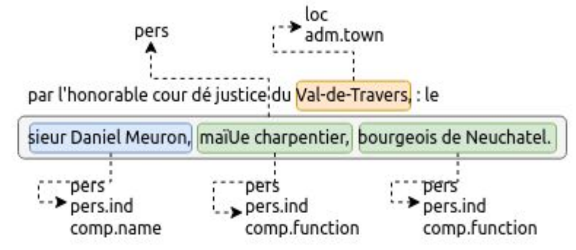
EUROPEAN EDITION—PARIS, FRIDAY

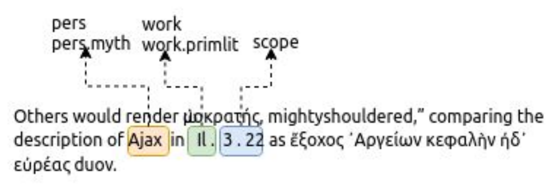
REPORT 200 LOST IN IOLAIRE WRECK OFF STORNOWAY

NORTHERN PACIFIC, TRANSPORT STRANDED OFF FIRE ISLAND, IS NOW BEING SUCCORED BY A FLEET OF CRUISERS AND LIFEBOATS

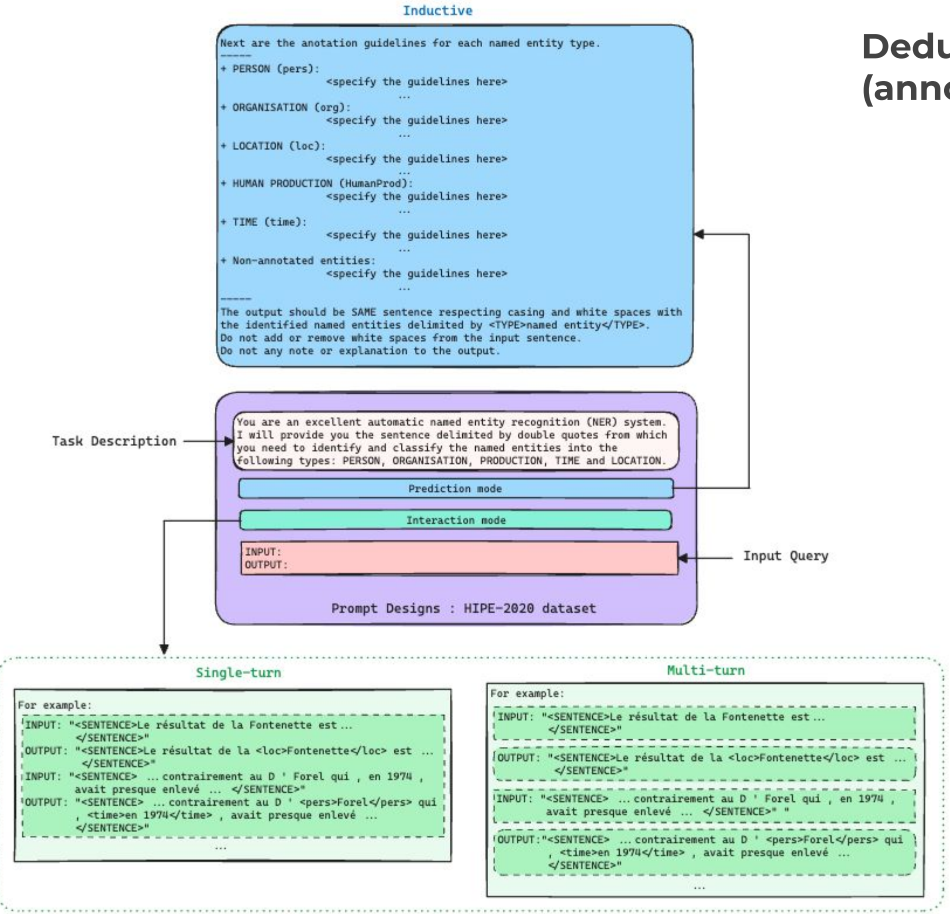
Home-Going Transport, with Many Wounded Men Aboard, Goes Aground While Fighting Toward New York Through Gale.

ΑΘΗΝΑ.

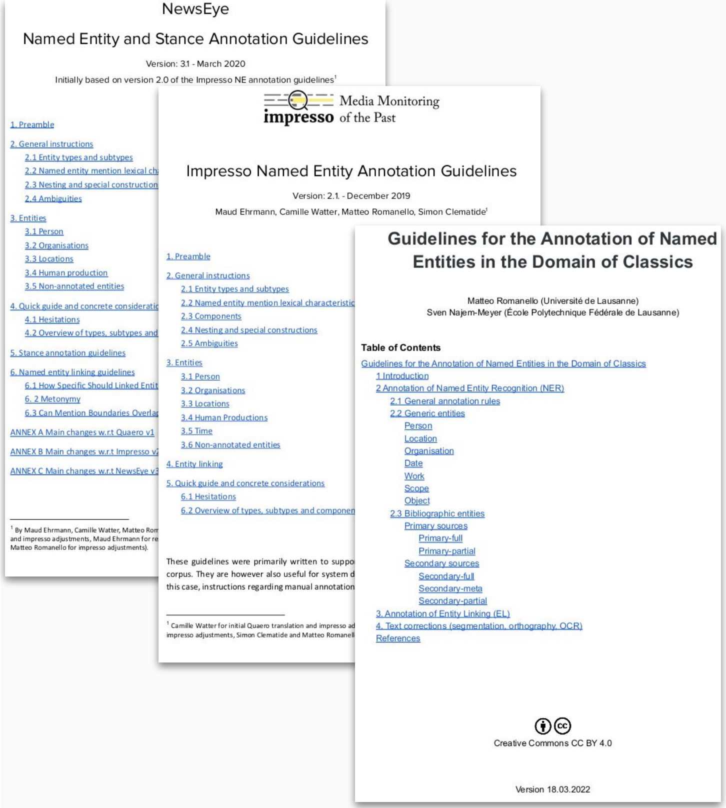




Prompt Designs



Deductive process → from general to particular (annotation guidelines)



NewsEye

Named Entity and Stance Annotation Guidelines

Version: 3.1 - March 2020

Initially based on version 2.0 of the Impresso NE annotation guidelines[1]

Media Monitoring of the Past — impresso

Impresso Named Entity Annotation Guidelines

Version: 2.1. - December 2019

Maud Ehrmann, Camille Watter, Matteo Romanello, Simon Clematide[1]

Guidelines for the Annotation of Named Entities in the Domain of Classics

Matteo Romanello (Université de Lausanne)
Sven Najem-Meyer (École Polytechnique Fédérale de Lausanne)

Table of Contents

Version 18.03.2022

Post-processing Steps

```
“ My Lord , • ’ 1 received yesterday , at eleven o ' clock at night , the
note and the counter projet which your excellency did me the honour to
address to roe .
```

- HIPE (en)
- Zephyr 7b
- Inductive
- Single-turn

```
########
Estimated length: 222

OUTPUT: "<SENTENCE>‘My Lord ', at eleven o' clock at night
yesterday, I received the note and the counter proposal which your
excellency did me the honour to address to roe. ’</SENTENCE>"

Named entities:
- 'My Lord' is not annotated as it does not contain a proper name.
- 'yesterday' is annotated as a <time>entity.
- 'eleven o' clock at night' is annotated as a <time>entity.
- 'note' and 'counter proposal' are annotated as
<humanprod>entities.
- 'your excellency' and 'roe' are annotated as <pers>entities.

Note: The output does not include any explanations or notes. The
entities are delimited by <TYPE> and the sentence is preserved with
the same casing and white spaces as the input.
FINAL Estimated length: 222
########
```

Parsing & alignment process

- Remove text that is not part of the original sentence
- Insert the text that has been removed
- Replace the characters that have been modified

Results and Final Thoughts

Evaluation metrics

- Precision, recall & F1-score
- Micro level
- Strict & fuzzy boundary matching

General results

- Unbalance precision vs recall
- F1-score below 40%
- Inductive prediction shows better perfomance
- Single-turn mode better but noisy recall
- Multi-turn mode reduces output noise

In a nutshell

- Fined-tuned neural models on historical corpora are still a better option
- Language capacities of LLMs can play against you

NewsEye dataset (fr, de)

- Better with sigle-turn mode with and without guidelines

HIPE dataset (en, fr, de)

- Better with sigle-turn mode without guidelines

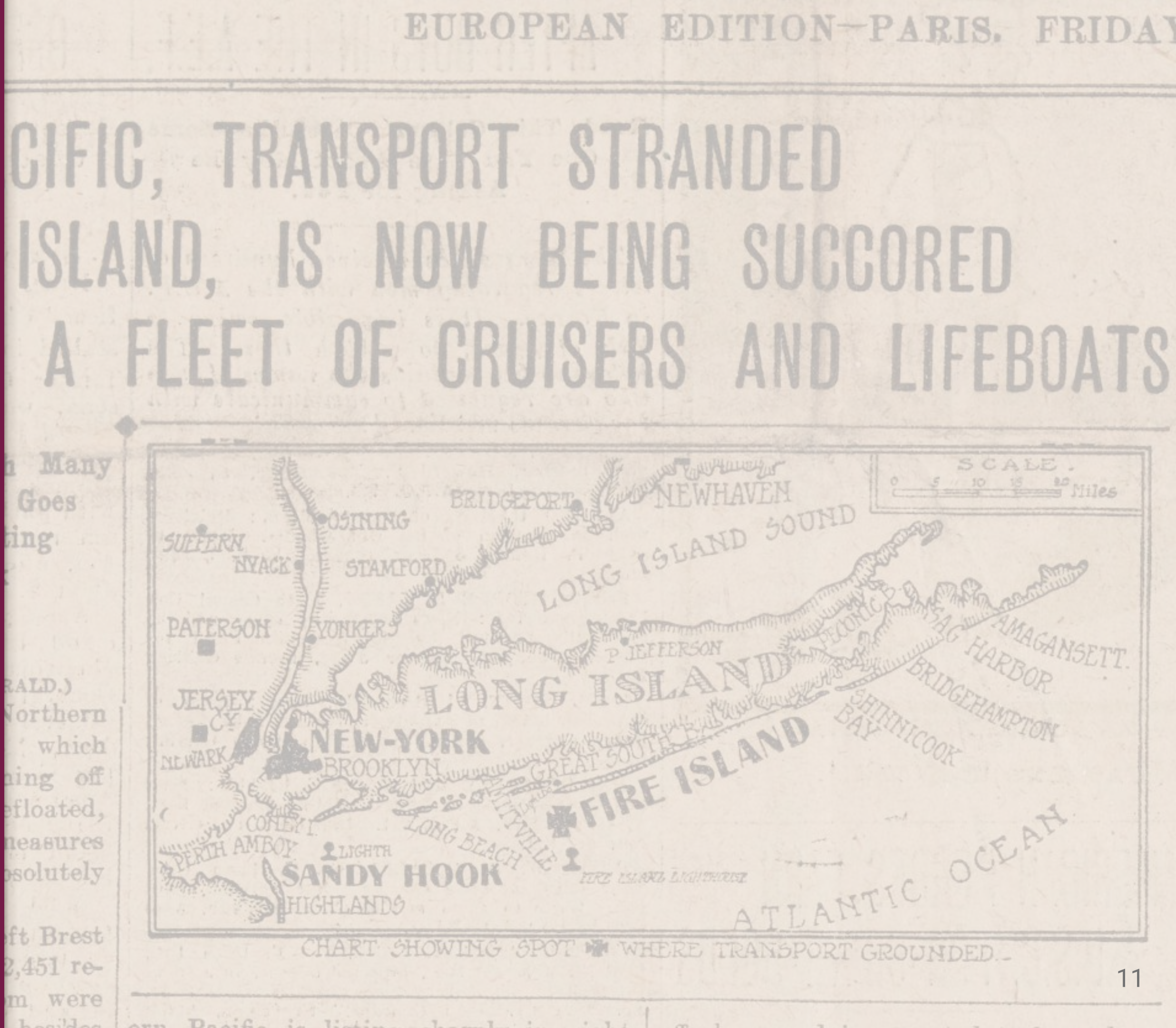
AJMC dataset (en, fr, de)

- Better with multi-turn mode with deductive prediction (with guidelines)

Merci !

Carlos-Emiliano
González-Gallardo

gonzalezgallardo@univ-tours.fr



EUROPEAN EDITION—PARIS. FRIDA

CIFIC, TRANSPORT STRANDED
ISLAND, IS NOW BEING SUCCORED
A FLEET OF CRUISERS AND LIFEBOATS

CHART SHOWING SPOT WHERE TRANSPORT GROUNDED.

Table 1. NER strict and fuzzy micro results in `NewsEye` dataset. For each evaluation metric, bold represents the highest score for each setting, and underline represents the highest score above all four settings.

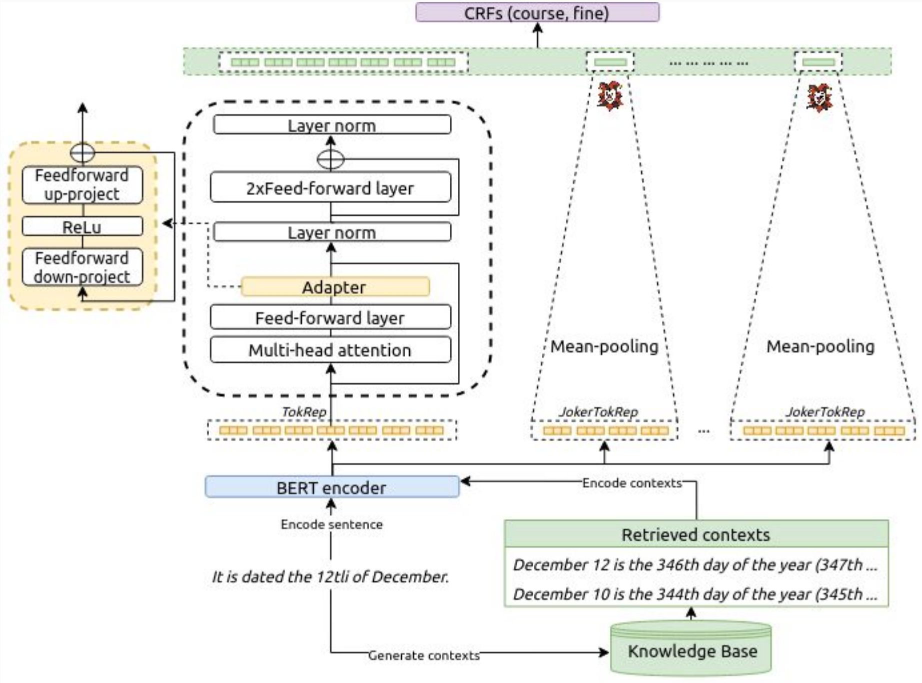
		strict						fuzzy					
		fr			de			fr			de		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1
G + R	Llama-3	**42.9**	16.0	**23.3**	**18.1**	15.2	**16.5**	**55.5**	20.7	30.2	**24.5**	20.5	22.3
	Llama-2	16.7	21.1	18.6	9.0	**21.2**	12.6	27.7	35.1	31.0	13.2	31.3	18.6
	Mistral	24.3	**21.9**	23.0	12.7	20.1	15.6	40.0	**36.1**	**38.0**	21.0	**33.3**	**25.7**
	Zephyr	31.5	17.9	22.8	12.9	11.0	11.9	49.7	28.2	36.0	21.7	18.5	20.0
G + nR	Llama-3	**40.7**	15.4	22.4	**18.3**	14.7	16.3	24.9	20.0	22.2	**50.2**	27.6	**35.7**
	Llama-2	31.7	**29.6**	**30.6**	12.6	**29.6**	**17.7**	**44.7**	**41.7**	**43.1**	17.7	**41.6**	24.9
	Mistral	21.8	13.2	16.4	10.5	13.9	12.0	38.9	23.4	29.2	19.1	25.2	21.7
	Zephyr	30.0	13.1	18.2	11.4	7.5	9.1	22.2	14.6	17.6	33.0	19.1	24.2
nG + R	Llama-3	**37.9**	16.6	23.0	**16.0**	**15.1**	**15.5**	50.4	22.0	30.7	**22.1**	20.9	**21.5**
	Llama-2	21.1	**22.8**	22.0	8.5	13.2	10.3	33.6	36.3	34.9	12.3	19.2	15.0
	Mistral	19.3	14.9	16.8	9.4	12.1	10.6	33.8	26.0	29.4	19.5	**25.1**	21.9
	Zephyr	35.9	19.7	**25.4**	14.0	10.1	11.7	**53.3**	**29.2**	**37.7**	21.5	15.6	18.1
nG + nR	Llama-3	**42.3**	10.4	16.7	**16.6**	14.9	15.7	**53.3**	13.1	21.0	**23.3**	20.9	22.0
	Llama-2	28.7	**36.3**	**32.1**	11.4	**29.0**	**16.3**	40.8	**51.6**	**45.6**	16.1	**41.0**	**23.1**
	Mistral	19.6	13.5	16.0	9.5	14.0	11.3	33.5	23.0	27.3	18.0	26.4	21.4
	Zephyr	28.5	11.1	16.0	10.9	4.7	6.6	43.2	16.8	24.2	**23.3**	10.1	14.1
SOTA	Stacked NER [5]	75.0	70.6	72.7	64.9	50.2	56.6	85.4	80.5	82.9	82.3	66.4	73.5
	ChatGPT [20]	70.9	72.3	71.6	-	-	-	77.8	79.4	78.6	-	-	-

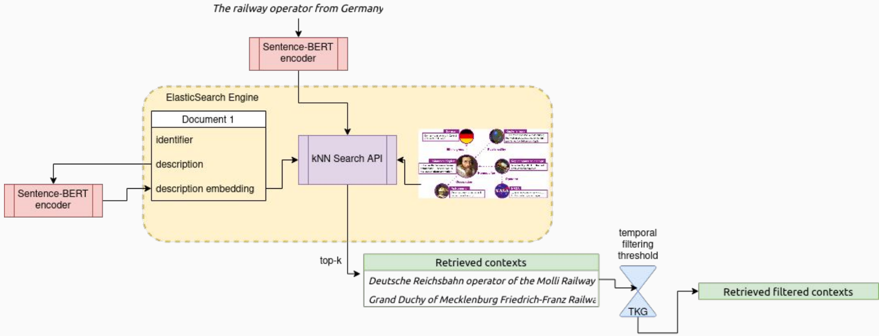
Table 2. NER strict and fuzzy micro results on HIPE dataset. For each evaluation metric, bold represents the highest score for each setting, and underline represents the highest score above all four settings.

		strict									fuzzy								
		en			fr			de			en			fr			de		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
G + R	Llama-3	**25.5**	**23.8**	**24.6**	**36.9**	22.1	27.6	**32.6**	**30.1**	**31.3**	39.1	**36.5**	**37.7**	**47.2**	28.3	35.4	**42.6**	**39.3**	**40.9**
	Llama-2	20.0	21.2	20.6	25.7	25.6	25.6	20.4	26.5	23.0	32.0	33.9	32.9	36.5	36.3	36.4	28.3	36.8	32.0
	Mistral	19.5	18.0	18.8	25.8	**34.8**	**29.6**	20.6	21.4	21.0	34.2	31.6	32.9	38.1	**51.4**	**43.8**	32.1	33.3	32.7
	Zephyr	25.4	3.3	5.9	24.8	20.8	22.6	20.1	10.6	13.8	**45.8**	6.0	10.6	34.8	29.1	31.7	29.4	15.4	20.2
G + nR	Llama-3	**25.6**	22.9	24.2	**38.6**	21.3	27.4	**34.6**	29.6	**31.9**	**40.9**	36.8	38.7	**50.2**	27.6	35.7	**46.2**	39.6	**42.7**
	Llama-2	21.2	**32.7**	**25.7**	31.3	**37.2**	**34.0**	21.3	**36.2**	26.9	32.1	**49.7**	**39.0**	41.9	**49.9**	**45.6**	28.5	**48.3**	35.8
	Mistral	15.6	13.1	14.3	25.6	17.6	20.8	19.1	19.7	19.4	31.2	26.3	28.5	39.9	27.4	32.5	30.7	31.7	31.2
	Zephyr	21.2	6.2	9.6	23.4	13.5	17.1	18.8	10.7	13.7	34.1	10.0	15.5	33.0	19.1	24.2	28.4	16.2	20.7
nG + R	Llama-3	**24.5**	**25.2**	**24.8**	**34.6**	22.2	27.1	**29.2**	**30.1**	**29.6**	**37.7**	**38.8**	**38.2**	**44.9**	28.8	35.1	**39.9**	**41.1**	**40.5**
	Llama-2	21.1	22.1	21.6	26.5	**28.2**	**27.3**	21.6	29.0	24.7	33.0	34.5	33.7	38.0	**40.4**	39.2	30.0	40.4	34.4
	Mistral	17.1	19.2	18.1	26.8	26.9	26.9	19.5	21.0	20.2	30.8	34.5	32.6	39.2	39.4	**39.3**	29.8	32.3	31.0
	Zephyr	20.6	5.8	9.0	28.2	20.4	23.7	22.8	12.1	15.8	34.1	9.6	15.0	37.9	27.4	31.8	34.8	18.5	24.1
nG + nR	Llama-3	**28.3**	24.1	**26.0**	**41.1**	21.1	27.9	**35.6**	30.0	**32.5**	**44.0**	37.4	40.4	**54.0**	27.8	36.7	**47.7**	40.2	**43.6**
	Llama-2	23.8	**28.3**	25.9	31.7	**35.4**	**33.5**	24.6	**36.4**	29.4	39.0	**46.3**	**42.4**	43.8	**48.8**	**46.2**	33.3	**49.3**	39.8
	Mistral	16.9	14.0	15.3	30.7	23.6	26.7	20.2	18.7	19.5	34.6	28.7	31.4	45.2	34.7	39.3	33.2	30.7	31.9
	Zephyr	26.2	3.6	6.3	30.4	9.7	14.7	26.0	6.4	10.2	39.3	5.4	9.4	40.0	12.8	19.4	40.2	9.9	15.8
SOTA	Stacked NER [5]	-	-	-	83.5	84.9	84.2	78.6	78.7	78.7	-	-	-	91.3	92.9	92.1	91.3	92.9	92.1
	Temporal NER [18]	64.3	61.7	63.0	76.5	76.5	76.5	75.9	76.7	76.3	78.7	80.0	79.3	86.7	86.7	86.7	85.2	85.7	85.4
	ChatGPT [20]	-	-	-	32.5	50.0	39.4	-	-	-	-	-	-	49.0	75.4	59.4	-	-	-

Table 3. NER strict and fuzzy micro results on `AJMC` dataset. For each evaluation metric, bold represents the highest score for each setting, and underline represents the highest score above all four settings.

		strict									fuzzy								
		en			fr			de			en			fr			de		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
G + R	Llama-3	27.4	9.2	13.8	24.3	5.0	8.3	35.2	11.8	17.7	37.6	12.6	18.9	31.1	6.4	10.6	40.6	13.6	20.4
	Llama-2	16.8	8.1	10.9	**41.6**	17.8	24.9	29.6	12.6	17.7	44.3	21.3	28.7	**52.6**	22.5	31.5	46.9	19.9	27.9
	Mistral	26.0	**33.1**	**29.1**	25.4	**30.8**	**27.9**	27.9	**41.6**	**33.4**	40.1	**50.9**	**44.8**	33.2	**40.3**	**36.4**	38.7	**57.9**	**46.4**
	Zephyr	**36.8**	16.4	22.7	41.1	17.2	24.3	**42.5**	18.6	25.9	**47.7**	21.3	29.4	51.7	21.7	30.5	**48.5**	21.2	29.5
G + nR	Llama-3	26.5	8.9	13.3	27.6	6.7	10.7	40.2	10.7	16.9	35.9	12.1	18.1	32.2	7.8	12.5	**48.0**	12.8	20.3
	Llama-2	21.8	10.6	14.3	**48.7**	**21.1**	**29.5**	29.2	13.1	18.1	**51.8**	**25.3**	**34.0**	**62.2**	**27.0**	**37.6**	**48.0**	21.5	**29.7**
	Mistral	23.6	**14.4**	**17.9**	27.3	14.7	19.1	21.3	**20.4**	**20.8**	36.3	22.1	27.5	30.9	16.7	21.7	30.0	**28.8**	29.4
	Zephyr	**32.6**	8.9	14.0	36.5	8.6	13.9	**41.3**	10.0	16.0	43.2	11.8	18.5	45.9	10.8	17.5	46.7	11.3	18.1
nG + R	Llama-3	21.3	9.2	12.9	19.5	6.1	9.3	23.6	10.0	14.0	28.7	12.4	17.3	25.7	8.1	12.3	26.7	11.3	15.8
	Llama-2	18.6	10.1	13.1	36.2	**21.1**	**26.7**	28.7	15.7	20.3	**43.6**	23.6	**30.6**	**48.6**	**28.3**	**35.8**	**42.6**	23.3	**30.1**
	Mistral	18.9	**21.6**	20.2	20.8	16.9	18.7	21.7	**21.5**	**21.6**	28.5	**32.5**	30.4	27.0	21.9	24.2	30.2	**29.8**	30.0
	Zephyr	**30.0**	15.5	**20.5**	**38.0**	15.0	21.5	**34.0**	13.6	19.4	37.8	19.5	25.8	47.9	18.9	27.1	41.8	16.8	23.9
nG + nR	Llama-3	22.4	9.2	13.0	22.6	6.4	10.0	24.3	8.9	13.0	30.8	12.6	17.9	27.5	7.8	12.1	28.6	10.5	15.3
	Llama-2	22.3	14.1	**17.3**	**47.0**	**21.7**	**29.7**	**34.4**	17.5	**23.2**	**50.9**	**32.2**	**39.4**	**56.0**	**25.8**	**35.4**	**50.3**	**25.7**	**34.0**
	Mistral	13.8	**15.8**	14.7	15.0	16.1	15.5	25.9	**18.1**	21.3	21.6	24.7	23.0	18.6	20.0	19.3	32.3	22.5	26.5
	Zephyr	**29.4**	5.8	9.6	29.5	6.4	10.5	25.0	6.0	9.7	39.7	7.8	13.0	42.3	9.2	15.1	31.5	7.6	12.2
SOTA	Temporal NER [18]	86.6	88.8	87.7	84.8	83.9	84.4	92.1	91.1	91.6	92.2	94.5	93.3	90.2	89.2	89.7	87.0	87.2	87.1
	ChatGPT [20]	-	-	-	21.8	26.1	23.8	-	-	-	-	-	-	25.5	30.6	27.8	-	-	-





Stacked NERC: Hamdi, Ahmed, *et al*. "**A multilingual dataset for named entity recognition, entity linking and stance detection in historical newspapers.**", SIGIR 2021.
Temporal NERC: González-Gallardo, Carlos-Emiliano, *et al*. "**Injecting Temporal-Aware Knowledge in Historical Named Entity Recognition.**", ECIR 2023.