Overview of #SMM4H 2024 – Task 2: Cross-Lingual Few-Shot Relation Extraction for Pharmacovigilance in French, German, and Japanese

Lisa Raithel[1,2,3], **Philippe Thomas**[3], **Bhuvanesh Verma**[3], **Roland Roller**[3], **Hui-Syuan Yeh**[4], **Shuntaro Yada**[5], **Cyril Grouin**[4], **Shoko Wakamiya**[5], **Eiji Aramaki**[5], **Sebastian Möller**[2,3], **Pierre Zweigenbaum**[4]
[1]BIFOLD – Berlin Institute for the Foundations of Learning and Data, Germany;
[2]Quality & Usability Lab, Technische Universität Berlin, Germany;
[3]German Research Center for Artificial Intelligence (DFKI), Berlin, Germany;
[4]Université Paris-Saclay, CNRS, LISN, Orsay, France;
[5]Nara Institute of Science and Technology, Nara, Japan;

Abstract

This paper provides an overview of Task 2 from the Social Media Mining for Health 2024 shared task (#SMM4H 2024), which focused on Named Entity Recognition (NER, Subtask 2a) and the joint task of NER and Relation Extraction (RE, Subtask 2b) for detecting adverse drug reactions (ADRs) in German, Japanese, and French texts written by patients. Participants were challenged with a few-shot learning scenario, necessitating models that can effectively generalize from limited annotated examples. Despite the diverse strategies employed by the participants, the overall performance across submissions from three teams highlighted significant challenges. The results underscored the complexity of extracting entities and relations in multi-lingual contexts, especially from user-generated content's noisy and informal nature.

1 Introduction

An adverse drug reaction (ADR) is defined as a "harmful or unpleasant reaction, resulting from an intervention related to the use of a medicinal product" (Edwards and Aronson, 2000). ADRs pose a significant challenge in pharmacovigilance. No medication is devoid of side effects, and despite clinical trials for each drug, the trial populations often fail to represent the entirety of real-world patients in terms of age, gender, health status, or ethnicity (Hazell and Shakir, 2006). Moreover, post-release surveillance efforts may miss patients experiencing issues with the medication (Hazell and Shakir, 2006), emphasizing the need for continuous monitoring of medication usage and effects.

Natural language processing can support this process by extracting potentially *novel* ADRs from text sources. Clinical texts and scientific literature are valuable resources containing information about ADRs. Still, they are either difficult to access for researchers outside a hospital or are published multiple weeks or even months after the occurrence/detection of an adverse effect. Social media, such as X (formerly known as Twitter) or patient forums, in which patients share and discuss their sorrows, concerns, and potential ADRs, instead became an alternative and up-to-date text source (Leaman et al., 2010; Segura-Bedmar et al., 2014; Zolnoori et al., 2019). Not all information from social media is necessarily reliable from a medical point of view. Still, it directly reflects the patient's perspective and at a much faster speed than well-curated scientific text.

To produce robust enough systems to process the vast amount of online data automatically, various datasets have been introduced in this context (Karimi et al., 2015; Klein et al., 2020; Tutubalina et al., 2021; Sboev et al., 2022, inter alia), shared tasks have been conducted (Magge et al., 2021a; Weissenbacher et al., 2022; Klein et al., 2024), and models have been published (Magge et al., 2021b; Portelli et al., 2022). However, like other text processing domains, most datasets exist only for English. To raise interest in this critical topic – particularly for non-English languages, which are not well-represented in this domain (Névéol et al., 2018) – we provide a Shared Task at #SMM4H 2024 (Xu et al., 2024): *Cross-Lingual Few-Shot Relation Extraction for Pharmacovigilance in French, German, and Japanese*. This paper describes the results and findings of this task. It targets joint cross-lingual named entity recognition and relation extraction in a multi-lingual setting. The data consists of French, German, and Japanese texts written by patients on social media such as X and patient fora, and is a subset of the KEEPHA corpus (Raithel et al., 2024).

Proceedings of The 9th Social Media Mining for Health Applications (#SMM4H) Workshop and Shared Tasks, pages 170–182

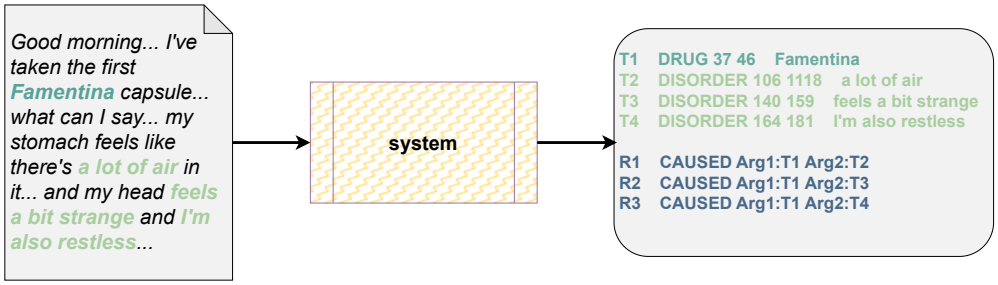

Figure 1: Visualization of input (a document) and expected output for both tasks. The output is a text file with predictions in `brat` format and shows an identifier (e.g., T1), a label (e.g., DRUG), the offsets of the entity, and the actual string (e.g., *Famenita*) in the case of NER. For RE, the annotations/predictions are extended by relations (identified with, e.g., R1), the relation type (e.g., CAUSED), and the head (Arg1) and tail (Arg2) arguments, referring to entity identifiers.

2 Shared Task

In this section, we present the task details and schedule, the data used for the challenge, our baseline system, and the participants' approaches.

2.1 Task

#SMM4H 2024 Task 2 targets extracting drug and disorder/body function mentions (Subtask 2a) and relations between those entities (Subtask 2b). The task is set up in a cross-lingual few-shot scenario: Training data consists mainly of Japanese and German data plus four French documents (see Table 1). The submitted systems are evaluated on Japanese, German, and French data.

Figure 1 visualizes the general process: Given a text document, a system should first predict entities and, subsequently, relations between these. The output format of the predictions is expected to be in `brat` format. The participants were asked to submit multi-lingual systems (FR + DE + JA) for one or all of the following tasks:

a) Named Entity Recognition (NER): Recognize mentions that belong to the classes `DRUG`, `DISORDER`, or `FUNCTION`.

b) Joint Named Entity and Relation Extraction (joint NER+RE): Recognize the entities mentioned above and the relations `TREATMENT_FOR` or `CAUSED` between them without relying on gold entities.

The participants could also submit predictions for only one or two languages.

2.2 Data

The data used for this challenge are a subset (in terms of fewer entities and relations) of the KEEPHA dataset (Raithel et al., 2024). It originates from different (non-parallel) social media sources (online patient fora and X) and is available in German, French, and Japanese. The choice of languages is due to the native languages spoken in the countries in which the labs involved in the data creation are located.[1] To diversify the data, the authors tried to use as many sources as possible, to get different populations of patients and also different types of text, e.g., short texts from X versus longer messages in fora. The German (training and test) data is from an online patient forum, whereas the Japanese documents are from X (training) and a patient forum (test). The French data, finally, is a translation of German documents from the same patient forum as the German data. The translation was necessary because there was no French patient forum (or other resource) that permitted access to the postings of its users. The translated French documents do not overlap with the German originals.

All data were annotated based on the same annotation guidelines[2], with a focus on the detection and extraction of adverse drug reactions, modeled by associating medication mentions (`DRUG`) with disorders (medical signs and symptoms, `DISORDER`) and body function mention (`FUNCTION`) using cause-consequence relations (`CAUSED`) to represent side effects elicited by medication intake, and treatment relations (`TREATMENT_FOR`) to represent medication used to treat medical symptoms. Figure 2 shows an example annotation. The tool used for annotation was `brat` (Stenetorp et al., 2012) (see an

[1] TU Berlin & DFKI in Berlin, Germany; LISN & Université Paris-Saclay in Orsay, France; NAIST in Nara, Japan, and RIKEN in Tokyo, Japan.

[2] `https://shorturl.at/BBHOS`

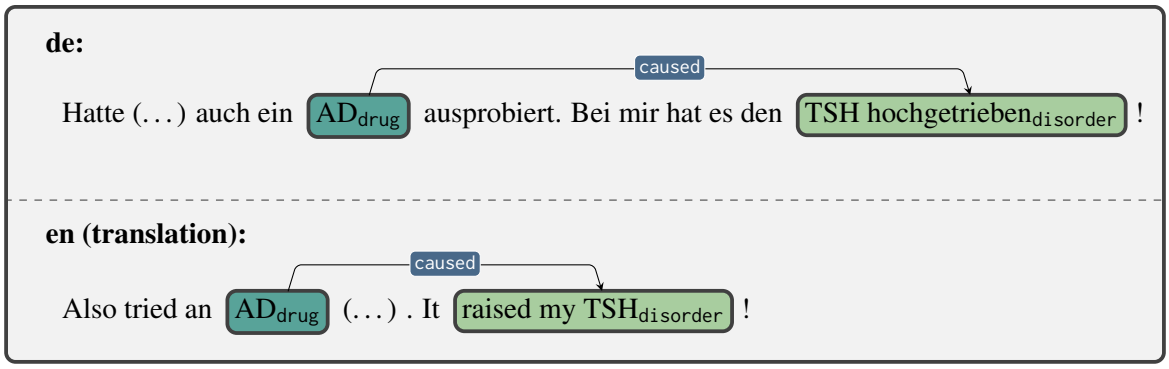

Figure 2: An annotated example. Top: original German text (shortened), bottom: English translation with projected annotations.

example in Figure 1, output of the system). The relation distribution is imbalanced. The number of `treatment_for` relations is much lower than that of `caused` relations, adding to the task's difficulty.

corpus	lang	src	#doc	#ent	#rel
train	de	patient forum	70	1,207	476
	ja	X	392	2,416	619
	fr	patient forum	4	69	32
dev	de	patient forum	23	424	141
	ja	patient forum	168	930	266
	fr	—	0	0	0

Table 1: Number of documents (#doc) with the number of entities (#ent), and relations (#rel) of each type for each language (lang) and source (src) of the training and development data.

2.3 Schedule

Table 2 shows the schedule of #SMM4H 2024 Task 2. The task was announced via several mailing lists (e.g., corpora-list, ML-news) and social media (e.g., X, LinkedIn) in two calls. In the beginning, 12 teams from diverse countries (Switzerland, India, France, China, USA, and Georgia) registered for Task 2.

Training data available	January 10, 2024
CodaLab available	January 17, 2024
Practice predictions due	April 3, 2024
Test data available	April 17, 2024
Evaluation end	April 24, 2024

Table 2: The schedule of Task 2 at #SMM4H 2024.

The CodaLab environment for the task submission was published in mid-January[3], shortly after the training and development data was released. We further provided the opportunity to test the prediction format until the beginning of April, which was, however, not used by many participants. The evaluation period lasted six days in total. Ultimately, only three teams submitted predictions to CodaLab during the evaluation phase.

2.4 Baseline

To provide a meaningful comparison, we developed two baseline systems for the shared task, one for NER and one for joint NER + RE.

2.4.1 NER

The baseline for NER is set up using the PyTorch-IE framework (Binder et al., 2024)[4], which allows to prototype and test information extraction pipelines quickly. For NER, we employed a simple token classification model that encapsulates a (pre-trained) Transformer model from the HuggingFace library (Wolf et al., 2019). We first fine-tuned an NER model for German and Japanese data separately with different hyper-parameters and performed inference on the test data using the corresponding best-performing models. For German, we utilized a German version of BERT (Devlin et al., 2019)[5] as the pre-trained model, while for Japanese, we used multi-lingual XLM-RoBERTa (Conneau et al., 2019)[6]. For French, we used the best-performing Japanese XLM-RoBERTa model. Utilizing pooled output embeddings from the pre-trained models and a classification head, we generate token-level predictions and convert the results back into `brat` format. We then combine the pre-

[3] https://codalab.lisn.upsaclay.fr/competitions/17204

[4] https://github.com/ArneBinder/pytorch-ie

[5] dbmdz/bert-base-german-uncased

[6] FacebookAI/xlm-roberta-base

dictions of all three languages to obtain the overall results.

2.4.2 Joint NER + RE

For joint NER+RE, we combined the NER system from above with a few-shot experiment using an LLM-based approach with Llama-3-8B-UltraMedical[7] (Zhang et al., 2024). We utilized the entities predicted during the NER task as input for the prompt given to the LLM. The specific details of the final prompt used in the experiment are provided in Appendix A.3. Specifically, we include definitions of the relations to be determined. We constructed three prompts, following the format outlined in Appendix A.3. We exclusively used German examples for the first prompt to predict relations within German data. Similarly, Japanese examples were used for the second prompt to predict relations within Japanese data. Finally, the third prompt was created using one German and Japanese example, aiming to predict relations across the entire dataset collectively. This last prompt was used for the French data.

2.5 Submitted Systems

The submitted systems are summarized in Table 3. For the leaderboard on CodaLab and the following tables, we selected the best three runs with more than three distinct submissions.

Yseop (Gupta, 2024) focused on NER for French and Japanese and on RE for Japanese only. For French NER, the authors utilized a combination of advanced language models, including the instruction LLM Mistral-7B (Jiang et al., 2023) and DrBERT-CASM2 (Labrak et al., 2023). For Japanese NER, they employed a multifaceted approach involving rule-based methods (medkit[8]), a Japanese-Multilingual Dictionary (JMdict[9]), and a Japanese medical language model based on RoBERTa (Liu et al., 2019), which was pre-trained on Japanese case reports and fine-tuned for NER using MedTxt-CR (Yada et al., 2022).[10] For the Japanese Relation Extraction, they re-used the RoBERTa model and fine-tuned it with the provided training data. Team Yseop submitted three runs for NER and joint NER + RE.

Team HBUT (Ke et al., 2024) concentrated solely on the named entity recognition task for all three languages. The methodology employed by the team focused on the use of LLMs. They explored three distinct prompting strategies to identify the most effective approach for NER. The team did not explore fine-tuning transformer architectures. The authors initially evaluated two different LLMs and selected GLM-3-Turbo (Zeng et al., 2023) as their preferred model. For the use with LLMs, the task had to be transferred into a generation task (instead of token classification), for which the authors designed specific prompts to get structured output. These outputs were then post-processed to result in `brat` format. Team HBUT submitted two runs to Subtask 2a and did not work on Relation Extraction.

Predictions of a third system were submitted to CodaLab. However, the predictions did not comply with the `brat` format and could not be evaluated.

2.6 Evaluation

The participants' submissions are ranked by non-weighted macro F_1 score ($\mathbf{F}_1$), precision (**P**), and recall (**R**) for both tasks. The evaluation script is a slightly modified version of 'brateval'[11] and can be found online[12]. The modifications were necessary to comply with the required output format for the evaluation platform CodaLab.

For Subtask 2a, we use an exact match of entities to calculate the previously mentioned scores. In the evaluation script, this corresponds to the parameters "-span-match exact".

For Subtask 2b, joint entity and relation extraction, note that both entity boundaries and types and relation types and arguments must match precisely. In the evaluation script, this corresponds to the parameters "-type-match exact -span-match exact". We also provide scores for relaxed (lenient) entity evaluation (for both subtasks) and results per language.

3 Results

In the following, we briefly describe the results for the two subtasks.

3.1 Subtask 2a – Named Entity Recognition

Table 4 presents the overall results for NER across languages using a strict match of entities. The re-

[7] https://huggingface.co/TsinghuaC3I/Llama-3-8B-UltraMedical

[8] https://medkit.readthedocs.io/en/stable/index.html

[9] https://www.edrdg.org/jmdict/j_jmdict.html

[10] https://huggingface.co/daisaku-s/medtxt_ner_roberta

[11] https://github.com/READ-BioMed/brateval

[12] https://github.com/Erechtheus/brateval

Team	Task	Language	P	R	F_1	System Summary
Yseop	NER	fr, ja	58.31	42.14	48.92	fr: Mistral-7B + DrBERT-CASM2; ja: rule-based system + dictionary + RoBERTa
HBUT	NER	de, fr, ja	60.52	26.54	36.90	GLM-3-Turbo + post-processing
Yseop	RE	ja	02.24	01.63	01.89	ja: RoBERTa fine-tuned

Table 3: Summary of the submitted systems with the scores on CodaLab (exact macro F_1, precision, and recall).

Team	Run	Overall			DISORDER			DRUG			FUNCTION		
		P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
baseline	run 1†	47.55	58.83	**52.60**	47.36	58.86	**52.48**	55.08	67.43	60.63	29.49	36.70	**32.70**
Yseop	run 1	23.02	11.10	14.98	49.23	09.09	15.34	15.77	19.09	17.27	00.00	00.00	00.00
	run 2	22.36	09.39	13.22	**57.66**	08.17	14.31	14.01	15.25	14.61	00.00	00.00	00.00
	run 3†	58.31	42.14	48.92	56.65	42.86	48.80	71.03	50.10	58.76	**30.13**	18.35	22.81
HBUT	run 1	55.04	24.21	33.63	47.69	23.60	31.57	67.95	34.75	45.98	00.00	00.00	00.00
	run 2†	**60.52**	26.54	36.90	54.89	27.89	36.98	**71.24**	34.44	46.43	00.00	00.00	00.00
∪	—	42.00	71.10	52.81	41.10	71.66	52.24	50.78	81.12	62.46	25.16	42.82	31.69

Table 4: Named Entity Recognition (NER) results on the test set for all participants using exact match evaluation across languages. † denotes the system with the best overall performance for each team. The overall score is the unweighted micro average across the three languages. ∪ shows the results when combining all predictions of the best (†) systems.

sults indicate the difficulty of our task. The best score on overall NER is an F_1 score of 52.60. While the detection of DRUG appears to be slightly more attainable (F_1 up to 60), the detection of FUNCTION achieves overall the lowest scores. Since team Yseop targeted only Japanese and French, and team HBUT did not find any FUNCTION entities, it is no surprise that the baseline system achieves the best results. Interestingly, the joint NER+RE approach yielded better results for the baseline system than the NER approach described in Section 2.4.1. Therefore, we only show the results of one baseline system for both subtasks.

Note that all systems achieve similar results concerning overall precision but that the submitted systems only produce very low recall, i.e., they fail to catch many of the gold entities. In contrast, precision and recall of the baseline system seem to be relatively balanced, with recall being slightly higher than precision.

A language-specific overview of the NER results is provided in Table 5. The approach of Yseop achieves the best overall results in the few-shot scenario for French. In contrast, HBUT achieves the best performance considering the detection of DRUG mentions in French documents. For Japanese, team Yseop performs similarly as the baseline but substantially drops in performance for FUNCTION.

Finally, Table 6 presents the relaxed scores, i.e., resulting scores for entities that do not exactly match but have some overlap with the gold data. The results highlight (similarly as in Table 4) that while the baseline was optimized for recall, the systems of the team Yseop and the team HBUT both achieve good scores in terms of precision.

3.2 Subtask 2b – Joint Entity and Relation Extraction

Table 7 presents the joint named entity and relation extraction task results, highlighting the task's difficulty. To extract relations correctly, the entities need to be detected accurately in the first place. Here, we differ between exact (where entities are detected correctly according to exact boundaries) and relaxed (where entity boundaries have to overlap at least partially), which allows a more flexible mapping of the corresponding entities. The performance of the NER system directly and strongly

Lang	Team	Overall			DISORDER			DRUG			FUNCTION		
		P	**R**	**F$_1$**	**P**	**R**	**F$_1$**	**P**	**R**	**F$_1$**	**P**	**R**	**F$_1$**
de	baseline	41.26	**55.56**	**47.35**	42.59	**52.09**	**46.86**	51.28	**64.52**	**57.14**	**20.00**	**46.15**	**27.91**
	Yseop	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00
	HBUT	**62.28**	27.51	38.17	**54.63**	27.44	36.53	**76.27**	36.29	49.18	00.00	00.00	00.00
fr	baseline	33.05	**46.43**	38.61	29.97	**36.63**	32.97	45.76	**71.15**	55.70	**08.16**	**16.33**	**10.88**
	Yseop	**60.68**	31.33	**41.32**	49.54	26.26	34.32	**76.52**	48.35	**59.26**	00.00	00.00	00.00
	HBUT	60.57	31.33	41.30	**55.90**	32.25	**40.90**	68.81	38.19	49.12	00.00	00.00	00.00
ja	baseline	**61.57**	**67.79**	**64.53**	**59.76**	**75.38**	**66.67**	67.90	**65.34**	**66.60**	**57.14**	**43.51**	**49.41**
	Yseop	57.52	59.03	58.27	58.98	64.05	61.41	68.22	64.50	66.31	30.13	28.87	29.49
	HBUT	60.00	23.15	33.41	54.12	25.05	34.25	**72.20**	31.09	43.47	00.00	00.00	00.00

Table 5: Language-specific named entity recognition results on the test set for all participants (best system) using exact match evaluation. The overall score is the non-weighted micro average for each language.

Team	Run	Overall			DISORDER			DRUG			FUNCTION		
		P	**R**	**F$_1$**	**P**	**R**	**F$_1$**	**P**	**R**	**F$_1$**	**P**	**R**	**F$_1$**
baseline	run 1	64.61	**79.94**	**71.46**	68.18	**84.74**	**75.57**	69.66	**85.27**	**76.68**	35.26	**43.88**	**39.10**
Yseop	run 1	46.38	22.36	30.17	73.07	13.49	22.77	38.99	47.20	42.70	00.00	00.00	00.00
	run 2	48.65	20.42	28.77	**86.29**	12.23	21.42	39.75	43.26	41.43	00.00	00.00	00.00
	run 3	75.19	54.34	63.08	75.60	57.20	65.13	85.15	60.06	70.44	**43.23**	26.33	32.73
HBUT	run 1	71.74	31.55	43.83	66.86	33.09	44.27	80.32	41.08	54.36	00.00	00.00	00.00
	run 2	**79.85**	35.02	48.68	75.14	38.17	50.63	**88.84**	42.95	57.90	00.00	00.00	00.00
$\cup$	—	51.16	86.60	64.32	51.69	90.11	65.69	58.44	93.36	71.88	31.09	52.93	39.17

Table 6: Named Entity Recognition (NER) results on the test set for all participants using relaxed match evaluation across languages. For each team, the system with the best overall performance is highlighted with †. The overall score is the unweighted micro average across the three languages. $\cup$ shows the results when combining all predictions of the best (†) systems.

influences the overall results. Therefore, as the previous results on NER were low, it is unsurprising that participants and baselines result in an F_1 score below 10. The fact that team Yseop did not target all languages also influenced the results.

Language-specific results are shown in Table 8. Team Yseop predicted only relations for the Japanese dataset.

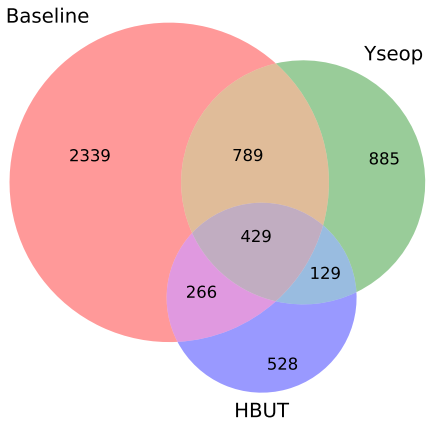

Figure 3: Exact overlap of entity predictions (Subtask 2a) for the best performing submission (†) for each team.

4 Analysis

We provide a brief analysis of the achieved results and the challenges the participants and their systems faced. This section is meant for further versions of the shared task and showcases common pitfalls.

4.1 Common Mistakes and Challenges

Several teams had difficulty providing predictions in the required `brat` format. At first glance, the `brat` format seems quite simple since the predictions are written in a text file, one extracted entity or relation per line, separated by either whitespace or tabular space, as shown in Figure 1. However, it seems that LLMs cannot consistently produce the correct `brat` format. Therefore, LLMs' output must be validated and/or pre-processed to ensure the correct format.

We also noticed that many offsets in the prediction files did not correspond to the actual string (i.e., the extracted entity) and that some spans started with -1, which resulted in invalid entity spans. Finally, we found several relations in the prediction files associated with non-existing entity mentions and, therefore, were ignored during automatic evaluation.

Overall, it seemed that not only was the development of the actual system the challenge in this task, but also, the post-processing of the systems' outputs provided some difficulty, especially when an LLM returned it.

4.2 Overlapping Entities

Figure 3 presents a Venn diagram of detected entities and their overlap between the best-performing submissions for each team. While the baseline system tends to have a high recall, the participants seem to have targeted a high precision. Therefore, it is unsurprising that the baseline detected the most significant number of entities. However, it is interesting to see that each team detected a large number of entities that the others did not detect.

We tested this by building the union of the predictions of the best system for each team. As shown in Table 4 and Table 6, the recall increases substantially, demonstrating that a large proportion of the three entities was indeed found by at least one system.

4.2.1 `FUNCTION` Mentions

`FUNCTION` entities were one of the more difficult mentions to detect. For instance, team HBUT did not find any mention correctly, and the baseline only reached an (exact) F_1 score of 32.7. This might be due to these mentions having a more difficult underlying concept: `FUNCTION` can be simply nouns or verbs (*"... I can sleep too"*), but they can also encompass more complicated phrases (*"I still had a relatively regular cycle."*). Also, the distinction between a `FUNCTION` and an actual `DISORDER` (which might be a negated body function) is often ambiguous. Detecting `FUNCTION` mentions worked much better for the Japanese data than for German and French. This could be because the boundaries of body functions might be easier to detect in the Japanese script than in the Latin script.

5 Discussion & Conclusion

In Task 2 of #SMM4H 2024, the participants had to tackle a difficult task. Starting with a small, multilingual, and layperson dataset and only a few examples for French plus no English data support, their systems had to distinguish three medical entities and two different relations, which are determined by temporal order and medical knowledge: The order of `DRUG` mentions, and `DISORDER`/`FUNCTION` mentions decides if the relation between the expressions is a "cause" or a "treatment" relation.

Team	Run	Exact			Relaxed		
		P	R	F_1	P	R	F_1
baseline	run 1	**04.25**	**06.81**	**05.23**	**07.82**	**12.53**	**09.63**
Yseop	run 3	02.24	01.63	01.89	03.17	02.32	02.68

Table 7: Relation extraction results on the test set for all participants using exact and relaxed match evaluation. Relaxed evaluation allows two entities to match if their boundaries overlap. The overall score is the non-weighted micro average across the three languages.

Lang	Team	Overall		
		P	R	F_1
de	baseline	**08.33**	**10.87**	**09.43**
	Yseop	00.00	00.00	00.00
fr	baseline	**03.88**	**06.38**	**04.83**
	Yseop	00.00	00.00	00.00
ja	baseline	**03.07**	**05.24**	**03.87**
	Yseop	02.26	04.49	03.01

Table 8: Language-specific relation extraction results on the test set for all participants (best system) using exact match evaluation. The overall score is the non-weighted macro average for each language.

Based on our baseline development using an LLM and the participants' submissions, a lot of post-processing seems necessary to be applied to the LLM output. It cannot be taken as is since the desired output format might not be consistently returned. We also noticed that it matters in which language the prompts are given to an LLM and that sometimes, for example, with an English prompt, the returned entities are correct but in English and, therefore, do not match the gold entities. Additionally, even if the LLM approach worked better than a transformer-based approach, the results were still unsatisfactory, especially for relation extraction.

Of course, combining different languages in different scripts and colloquial texts on which the models were not trained is somewhat tricky. However, given the success of LLMs in other domains or other genres of text, e.g., scientific documents, we were surprised that none of the teams beat the baseline. We, therefore, think that there is still a lot of work to be done in the medical domain concerning patient-generated texts, especially for non-English speaking patients, and that even LLMs seem to be only a part of a potential solution.

It is worth looking into the details of the single systems' benefits in future work. Despite the low number of participating teams, there were still many different approaches to Task 2 (rule-, transformer-, and LLM-based). Inspecting the detected entities and relation of each system might yield further insights and lead to a more successful system.

Acknowledgements

We thank the anonymous reviewers for their constructive feedback on this paper. Our work was supported by the Cross-ministerial Strategic Innovation Promotion Program (SIP) on "Integrated Health Care System" Grant Number JPJ012425, by JPMJCR20G9, ANR-20-IADJ-0005-01, and DFG-442445488 under the trilateral ANR-DFG-JST AI Research project KEEPHA, and by the German Federal Ministry of Education and Research under the grant BIFOLD24B.

References

Arne Binder, Leonhard Hennig, and Christoph Alt. 2024. Pytorch-ie: Fast and reproducible prototyping for information extraction. *Preprint*, arXiv:2406.00007.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

I. Ralph Edwards and Jeffrey K. Aronson. 2000. Adverse drug reactions: Definitions, diagnosis, and management. *The Lancet*, 356(9237):1255–1259.

Anubhav Gupta. 2024. "a team at #smm4h 2024: Pharmacovigilance shared task in english, french and japanese". In *Proceedings of The 9th Social Media Mining for Health Research and Applications Workshop and Shared Tasks*, Bangkok, Thailand. Association for Computational Linguistics.

Lorna Hazell and Saad A. W. Shakir. 2006. Under-reporting of adverse drug reactions : A systematic review. *Drug Safety*, 29(5):385–396.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Sarvnaz Karimi, Alejandro Metke-Jimenez, Madonna Kemp, and Chen Wang. 2015. Cadec: A corpus of adverse drug event annotations. *Journal of Biomedical Informatics*, 55:73–81.

Yuanzhi Ke, Zhangju Yin, Xinyun Wu, and Caiquan Xiong. 2024. "hbut at #smm4h 2024 task2: Cross-lingual few-shot medical entity extraction using a large language model". In *Proceedings of The 9th Social Media Mining for Health Research and Applications Workshop and Shared Tasks*, Bangkok, Thailand. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. *Internation Conference on Learning Representations*.

Ari Klein, Ilseyar Alimova, Ivan Flores, Arjun Magge, Zulfat Miftahutdinov, Anne-Lyse Minard, Karen O'Connor, Abeed Sarker, Elena Tutubalina, Davy Weissenbacher, and Graciela Gonzalez-Hernandez. 2020. Overview of the Fifth Social Media Mining for Health Applications (#SMM4H) Shared Tasks at COLING 2020. In *Fifth Social Media Mining for Health Applications(#SMM4H) Shared Tasks at COLING 2020*, page 10.

Ari Z Klein, Juan M Banda, Yuting Guo, Ana Lucia Schmidt, Dongfang Xu, Ivan Flores Amaro, Raul Rodriguez-Esteban, Abeed Sarker, and Graciela Gonzalez-Hernandez. 2024. Overview of the 8th Social Media Mining for Health Applications (#SMM4H) shared tasks at the AMIA 2023 Annual Symposium. *Journal of the American Medical Informatics Association*, 31(4):991–996.

Yanis Labrak, Adrien Bazoge, Richard Dufour, Mickaël Rouvier, Emmanuel Morin, Béatrice Daille, and Pierre-Antoine Gourraud. 2023. Drbert: A robust pre-trained model in french for biomedical and clinical domains. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16207–16221.

Robert Leaman, Laura Wojtulewicz, Ryan Sullivan, Annie Skariah, Jian Yang, and Graciela Gonzalez. 2010. Towards Internet-Age Pharmacovigilance: Extracting Adverse Drug Reactions from User Posts in Health-Related Social Networks. In *Proceedings of the 2010 Workshop on Biomedical Natural Language Processing*, pages 117–125, Uppsala, Sweden. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv:1907.11692 [cs]*.

Arjun Magge, Ari Klein, Antonio Miranda-Escalada, Mohammed Ali Al-garadi, Ilseyar Alimova, Zulfat Miftahutdinov, Eulalia Farre-Maduell, Salvador Lima Lopez, Ivan Flores, Karen O'Connor, Davy Weissenbacher, Elena Tutubalina, Abeed Sarker, Juan M Banda, Martin Krallinger, and Graciela Gonzalez-Hernandez, editors. 2021a. *Proceedings of the Sixth Social Media Mining for Health (#SMM4H) Workshop and Shared Task*. Association for Computational Linguistics, Mexico City, Mexico.

Arjun Magge, Elena Tutubalina, Zulfat Miftahutdinov, Ilseyar Alimova, Anne Dirkson, Suzan Verberne, Davy Weissenbacher, and Graciela Gonzalez-Hernandez. 2021b. DeepADEMiner: A deep learning pharmacovigilance pipeline for extraction and normalization of adverse drug event mentions on Twitter. *Journal of the American Medical Informatics Association*, 28(10):2184–2192.

Aurélie Névéol, Hercules Dalianis, Sumithra Velupillai, Guergana Savova, and Pierre Zweigenbaum. 2018. Clinical Natural Language Processing in languages other than English: Opportunities and challenges. *Journal of Biomedical Semantics*, 9(1):12.

Beatrice Portelli, Simone Scaboro, Emmanuele Chersoni, Enrico Santus, and Giuseppe Serra. 2022. AILAB-Udine@SMM4H'22: Limits of Transformers and BERT Ensembles. In *Proceedings of The Seventh Workshop on Social Media Mining for Health Applications, Workshop & Shared Task*, pages 130–134, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Lisa Raithel, Hui-Syuan Yeh, Shuntaro Yada, Cyril Grouin, Thomas Lavergne, Aurélie Névéol, Patrick Paroubek, Philippe Thomas, Tomohiro Nishiyama, Sebastian Möller, Eiji Aramaki, Yuji Matsumoto, Roland Roller, and Pierre Zweigenbaum. 2024. A Dataset for Pharmacovigilance in German, French, and Japanese: Annotating Adverse Drug Reactions across Languages. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, Turin, Italy.

Alexander Sboev, Sanna Sboeva, Ivan Moloshnikov, Artem Gryaznov, Roman Rybka, Alexander Naumov, Anton Selivanov, Gleb Rylkov, and Vyacheslav Ilyin. 2022. Analysis of the Full-Size Russian Corpus of Internet Drug Reviews with Complex NER Labeling Using Deep Learning Neural Networks and Language Models. *Applied Sciences*, 12(1):491.

Isabel Segura-Bedmar, Ricardo Revert, and Paloma Martínez. 2014. Detecting drugs and adverse events from Spanish social media streams. In *Proceedings of the 5th International Workshop on Health Text Mining and Information Analysis (Louhi)*, pages 106–115, Gothenburg, Sweden. Association for Computational Linguistics.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. Brat: A Web-based Tool for NLP-Assisted Text Annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107, Avignon, France. Association for Computational Linguistics.

Elena Tutubalina, Ilseyar Alimova, Zulfat Miftahutdinov, Andrey Sakhovskiy, Valentin Malykh, and Sergey Nikolenko. 2021. The Russian Drug Reaction Corpus and neural models for drug reactions and effectiveness detection in user reviews. *Bioinformatics (Oxford, England)*, 37(2):243–249.

Davy Weissenbacher, Juan Banda, Vera Davydova, Darryl Estrada Zavala, Luis Gasco Sánchez, Yao Ge, Yuting Guo, Ari Klein, Martin Krallinger, Mathias Leddin, Arjun Magge, Raul Rodriguez-Esteban, Abeed Sarker, Lucia Schmidt, Elena Tutubalina, and Graciela Gonzalez-Hernandez. 2022. Overview of the Seventh Social Media Mining for Health Applications (#SMM4H) Shared Tasks at COLING 2022. In *Proceedings of The Seventh Workshop on Social Media Mining for Health Applications, Workshop & Shared Task*, pages 221–241, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *CoRR*, abs/1910.03771.

Dongfang Xu, Guillermo Lopez Garcia, Lisa Raithel, Roland Roller, Philippe Thomas, Eiji Aramaki, Shuntaro Yada, Pierre Zweigenbaum, Sai Tharuni Samineni, Karen O'Connor, Yao Ge, Sudeshna Das, Abeed Sarker, Ari Klein, Lucia Schmidt, Vishakha Sharma, Raul Rodriguez-Esteban, Juan Banda, Ivan Flores Amaro, Davy Weissenbacher, and Graciela Gonzalez-Hernandez. 2024. Overview of the 9th social media mining for health (#SMM4H) research and applications workshop and shared tasks at ACL 2024. In *Proceedings of The 9th Social Media Mining for Health Research and Applications Workshop and Shared Tasks*, Bangkok, Thailand. Association for Computational Linguistics.

Shuntaro Yada, Shoko Wakamiya, Yuta Nakamura, and Eiji Aramaki. 2022. Real-MedNLP: Overview of REAL document-based MEDical Natural Language Processing Task. In *Proceedings of the 16th NTCIR Conference on Evaluation of Information Access Technologies*, Tokyo, Japan.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, Weng Lam Tam, Zixuan Ma, Yufei Xue, Jidong Zhai, Wenguang Chen, Peng Zhang, Yuxiao Dong, and Jie Tang. 2023. GLM-130B: An Open Bilingual Pre-trained Model. *Preprint*, arXiv:2210.02414.

Kaiyan Zhang, Ning Ding, Biqing Qi, Sihang Zeng, Haoxin Li, Xuekai Zhu, Zhang-Ren Chen, and Bowen Zhou. 2024. Ultramedical: Building specialized generalists in biomedicine. `https://github.com/TsinghuaC3I/UltraMedical`.

Maryam Zolnoori, Kin Wah Fung, Timothy B. Patrick, Paul Fontelo, Hadi Kharrazi, Anthony Faiola, Yi Shuan Shirley Wu, Christina E. Eldredge, Jake Luo, Mike Conway, Jiaxi Zhu, Soo Kyung Park, Kelly Xu, Hamideh Moayyed, and Somaieh Goudarzvand. 2019. A systematic approach for developing a corpus of patient reported adverse drug events: A case study for SSRI and SNRI medications. *Journal of Biomedical Informatics*, 90:103091.

A Baseline Details

A.1 Discontinuous Entities

The dataset contains fragmented spans where a single entity consists of two disjoint spans. To handle these fragmented spans, we split them into two separate spans and establish a temporary relation between the new spans. If a relationship involves a fragmented span, we create new relationships accordingly. For instance, if a CAUSED relation existed between a fragmented span (span 1) and a continuous span (span 2), we create new relations after splitting span 1 into span 11 and span 12: span 11 CAUSED span 2 and span 12 CAUSED span 2. This implies that we train models with simple entities containing only a single span. During prediction, entities linked by the temporary relation are combined to form a fragmented entity with two spans, but only if both entities have the same predicted label; otherwise, the temporary relations are ignored. If any of these entities had a relation with another entity, that relation is maintained after conversion to the fragmented entity.

A.2 Experimental Details

A.2.1 NER

For the NER task on German data, we utilized the dbmdz/bert-base-german-uncased pretrained model. For Japanese and French, we employed the FacebookAI/xlm-roberta-base model. We implemented a BIO encoding scheme for token labeling, resulting in a total of 7 classes. During training, we used a cross-entropy loss function and the Adam optimizer (Kingma and Ba, 2014) with a learning rate of 1e-05. To handle lengthy texts, we split them based on two parameters: max_length (set to 512) and stride (set to 64). Each split contains up to 512 tokens, with an overlap of 64 tokens between consecutive splits.

A.2.2 Joint NER+RE

For joint NER-RE inference, combinations of predicted entities are used for relation classification. However, this approach may introduce entity pairs that are widely separated in the text. To address this, we employed a max_window parameter (set to 512), which specifies the maximum allowed inner distance between entity pairs. Additionally, we included reversed gold relations by swapping the head and tail entities.

A.3 Prompt Format Descriptions

The NER prompt begins by defining the entities identified in the text, including DRUG, DISORDER, and FUNCTION. The prompt then provides examples to illustrate the task. Each example presents an input text followed by a list of identified entities, with explanations for their classification. Each identified entity is presented with its associated information, including the entity text, a True/False label indicating the accuracy of the entity classification, and an explanation justifying the classification. The explanation specifies the entity type (DRUG, DISORDER, or FUNCTION) and provides context for why the entity fits into that category.

The prompt for the joint NER + RE task specifies that we should classify relations between provided entities as CAUSED or TREATMENT_FOR. We then define the two types of relations. Following these definitions, the prompt includes multiple examples to illustrate the process. Each example is structured first to present the input text where the relations must be identified. Then, the identified entities from the text are listed and categorized into DRUG, DISORDER, or FUNCTION. Finally, the output section details the relationships identified between the entities. Each relationship is specified with the first entity, the second entity, a True/False label indicating the presence of the relationship, and an explanation justifying the prediction and identifying the relation type (CAUSED or TREATMENT_FOR).

A.4 Prompt used for NER task

```
Defn: The following are the definitions of the entities
DRUG: any mention of a medication name ("iburpofen"), brand ("Vick"), or agent
("Sorbitan"), including dietary supplements ("magnesium"), even when abbreviated ("AD"
for "anti depressive")
DISORDER: any disease, sign, or symptom related to the patient's health, including
mental issues. Sometimes a disorder may be expressed as a parameter in combination
with a value: high LDL
FUNCTION: all body functions and processes. Body functions are often represented in
biomarkers eg: HDL, WBC. It also includes mental functions
Difference between DISORDER and FUNCTION: We annotate adverse biological processes as
disorder and neutral/positive processes as function

<1>
Example 1: Ich nehme seit zwei Wochen Ibuprofen gegen meine Kopfschmerzen.

Entities:
1. Ibuprofen | True | as it refers to a specific medication (DRUG)
2. Kopfschmerzen | True | as it refers to a type of pain (DISORDER)
</1>

<2>
Example 2: Seitdem ich Magnesium nehme, fühle ich mich weniger müde.

Entities:
1. Magnesium | True | as it refers to a dietary supplement (DRUG)
2. müde | True | as it refers to fatigue, a symptom (DISORDER)
</2>

<3>
Example 3:
Text: {text}

Entities:
```

A.5 Prompt used for RE task

```
# Task: Use ONLY the provided entities to classify relations : [CAUSED, TREATMENT_FOR]

## Definition:
TREATMENT_FOR: This relation connects a DRUG and the targeted DISORDER, describing the
medication that was used to treat the patient's symptoms.
CAUSED: We only annotate a caused relation when the entities DRUG, DISORDER, or
FUNCTION are concerned. Explicit formulation of a <cause>-<consequence> relation,
Lexical semantics of nouns or verbs, e.g., <cause> provokes <consequence>

### Examples
<1>
Example 1:
Input: Ich nehme seit zwei Wochen Ibuprofen gegen meine Kopfschmerzen.
Entities:
DRUG - [Ibuprofen]
DISORDER - [Kopfschmerzen]
Output:
Relations:
Ibuprofen | Kopfschmerzen | True | Ibuprofen is used as a treatment for headaches
(TREATMENT_FOR)
</1>

<2>
Example 2:
Input: Seitdem ich Magnesium nehme, fühle ich mich weniger müde.
Entities:
DRUG - [Magnesium]
DISORDER - [müde]
Output:
Relations:
Magnesium | müde | True | Magnesium is used as a treatment for fatigue (TREATMENT_FOR)
</2>

<3>
Example 3:
Input: {text}
Entities:
{entities}
Output:
Relations:
```