# Massively Multilingual Instruction-Following Information Extraction

**Thang Le[1], Huy Huu Nguyen[2], Luu Anh Tuan[3], Thien Huu Nguyen[2]**
[1]Movian AI, Vietnam
[2]Department of Computer Science, University of Oregon, USA
[3]Nanyang Technological University, Singapore
thangld.nlp@gmail.com
{huy,thienn}@uoregon.edu
anhtuan.luu@ntu.edu.sg

## Abstract

The literature on information extraction (IE) has mostly centered around a selected few languages, hindering their applications on multilingual corpora. In this work, we introduce MASSIE - a comprehensive collection for instruction-following multilingual IE that standardizes and unifies 215 manually annotated datasets, covering 96 typologically diverse languages from 18 language families. Based on MASSIE, we conduct empirical studies on few-shot in-context learning and report important factors that either positively or negatively affect LLMs' performance in multilingual IE, covering 21 LLMs sizing from 0.5B to 72B. Additionally, we introduce LF1 - a structure-aware metric that captures partially matched spans, resolving the conservativeness of standard exact matching scheme which overpenalizes LLMs' predictions. Overall, our results signify that multilingual IE remains very challenging for existing LLMs, especially on complex tasks involving relations and events. In addition, performance gap is extremely large among high- and low-performing languages, but the group of similar-performing languages largely overlap between different LLMs, suggesting a shared performance bias in current LLMs.

## 1 Introduction

Multilingual Information Extraction (IE) aims to extract structured insights from unstructured data originating from multiple languages. Despite the existence of approximately 7000 languages worldwide (Joshi et al., 2020), past IE research has predominantly focused on a selected few high-resource languages, such as English or Chinese (Wang et al., 2023a; Jiao et al., 2023; Tang et al., 2024), significantly hindering IE applications on multilingual corpora. This tendency persisted in the era of LLMs, where state-of-the-arts methodologies and evaluations remain centered around these high-resource languages (Gui et al., 2024; Qi et al., 2024), further catalyzing unbalance in technology adoptions. Concretely, we identify 6 challenges of existing research in multilingual IE: (I) **Lack of a standardized benchmark**: Unlike language understanding or text embedding where widely recognized benchmarks such as MMMLU[1](Hendrycks et al., 2021) and MMTEB (Enevoldsen et al., 2025) exist, there is no standardized benchmark for multilingual IE. As a result, past IE works evaluated on separate sets of datasets and languages with different splitting and demonstrations (Gui et al., 2024; Qi et al., 2024; Zuo et al., 2024), making results incomparable; (II) **Lack of diverse evaluation** The set of languages involved in previous IE works, especially LLM-related ones (Gui et al., 2024; Zuo et al., 2024), is often small both in number and language families; (III) **Lack of reliable evaluation** Several IE works report their results on silver-standard datasets (Seganti et al., 2021; Zuo et al., 2024; Parekh et al., 2024), which have often been criticized for low-quality ground truths (Miranda, 2023; Le et al., 2024a); (IV) **Lack of a suitable metric** Past IE works rely on exact matching for evaluation (Lu et al., 2022; Ping et al., 2023; Gui et al., 2024) which overpenalizes LLMs' predictions, especially in few-shot ICL[2] or zero-shot SFT[3] settings where models lack in direct supervisions from the testing domain. In addition, discontinuous metrics such as exact matching have been criticized for obfuscating the measurement of emergent abilities in LLMs (Schaeffer et al., 2023); (V) **Lack of study on prompting techniques** Past IE works rely on manually crafted prompts (Wang et al., 2023a; Gui et al., 2024), whereas the influence of latest prompting techniques such as CHAIN-OF-THOUGHT (Wei et al., 2022) or SELF-IMPROVING PROGRAM (Khattab et al., 2024; Opsahl-Ong et al., 2024) remains un-

[1]https://huggingface.co/datasets/openai/MMMLU
[2]In-Context Learning
[3]Supervised Fine-Tuning



*Findings of the Association for Computational Linguistics: ACL 2025*, pages 3542–3585

derexplored; (VI) **Lack of analysis in large-scale evaluation** Past IE works have limited evaluation scopes (e.g. datasets, languages), raising questions of whether common characteristics of LLMs regarding parameter scaling (Kaplan et al., 2020; Hoffmann et al., 2022) or demonstrations (Min et al., 2022; Zhang et al., 2024) would still hold for multilingual IE at scales. Furthermore, it remains unclear to which extent have modern LLMs made progress in conducting multilingual IE.

In this paper, we seek to improve the status quo regarding these 6 challenges. Particularly for (I, II, III), we introduce the MASSIE collection (Sec. 2) formed through aggregating and standardizing 215 human-annotated IE datasets featuring 96 languages from 18 language families. Based on MASSIE, we construct two benchmarks (Sec. 2.3): M-HEAVY and M-LIGHT, with the prior designed for intensive and sparse evaluation while the latter is reserved for iterative development and more balanced evaluation. For (IV), we implement LF1 (Sec. 2.4) - a structure-aware soft metric that distinguishes between partially correct prediction spans while taking into account task structures. For (VI), we benchmark 21 LLMs sizing from 0.5B to 72B and provide high-level analyses in the few-shot ICL settings (Sec. 3). For (V), we experiment with intermediate reasoning (Sec. 4) and prompt searching (Sec. 5), and delineate important factors that contribute to LLMs' prompting performance for multilingual IE.

## 2 The MASSIE Collection

### 2.1 Task Selection

We select tasks based on availability of datasets i.e. whether a certain task is sufficiently and exclusively annotated in multiple languages. Our selections include five task types: Named Entity Recognition (NER), Relation Extraction (RE), Slot Filling (SF), Event Detection (ED) and Event Extraction (EE). We show examples of these tasks in Figure 1 and further describe them in Appendix A.

### 2.2 Construction

Due to space constraints, we briefly describe the construction of MASSIE and leave further details in Appendix B. For each mentioned task type, we first query search engines and paper corpora to collect corresponding human annotated IE datasets. Afterwards, we conduct preprocessing, de-duplication, de-contamination and convert instances into unified formats. This process was extremely time-consuming as we had to implement individual scripts for each dataset and occasionally needed to consult the dataset's authors on missing guidelines.

**Batched Schema** The schema set of each dataset varies significantly in length, exceeding a hundred for some. Querying the output for all schemas in a single forward pass can be extremely challenging, given the low context utilization problem in LLMs (An et al., 2024). In addition, train-test mismatch in the number of querying schemas have been shown to impair LLMs' performance (Gui et al., 2024). Therefore, we apply *batched schema* (Gui et al., 2024) to partition the schema query of each instance into batched queries $B_i$ of $split\_num$[4] schemas each. For datasets whose total number of schemas $n$ is not divisible by $split\_num$, we allow the last batch $B_n$ if $len(B_n) \geq n\%2$ or merge it with $B_{n-1}$ otherwise.

During training/development batching, as semantically similar schemas could present ambiguity that hinder models' abilities (Gui et al., 2024), we further apply *negative schema mining* to encourage co-occurences of such similar schemas in the same batch. Particularly, we identify $K$ schemas most semantically similar to a schema $S_i$ following a multilingual embedding model[5](Li et al., 2023) and treat these $K$ schemas as *hard negatives* of $S_i$. For each training/development instance with at least one positive schema $S_i$[6], we construct the schema list $L_i = S_i + \bigcup_{j=1}^{K} S_{hard,ij} + N_i$, which include $S_i$, their hard negatives $S_{hard,ij}$ and up-to $split\_num$ sampled non-hard negatives $N_i$, and conduct batching on this schema list. If the instance does not include any positive schema, we construct $L_i$ by sampling negatives from a binomial distribution with $p = 0.5$. For test instances, batching is conducted uniformly without hard negative mining and the schema list $L_i$ features all schemas in the dataset. Simultaneously, we pair batched samples with manually written English instructions of each task. Due to lack of human resources, we adopt automatically translated instructions for other languages and solely use these for selected experiments[7]. Hereafter in this work,

[4]We set $split\_num = 6$

[5]https://huggingface.co/Alibaba-NLP/gte-Qwen2-7B-instruct

[6]Spans related to this schema do exist in the input text

[7]Much of this work focuses on handling multilingual text and schemas through a unified instruction-following interface,

| Task | Input | Schema | Output |
|---|---|---|---|
| NER | 강리나 박사께서 휠체어 타는 씬은 명장면 | ["Person"] | [{"entity": "강리나", "entity_type": "Person"}] |
| RE | 『羅生門』(らしょうもん)は、1950年の日本の映画である | ["AdministrativeLocation"] | [{"head": "らしょうもん", "tail": "日本", "relation": "AdministrativeLocation"}, {"head": "羅生門", "tail": "日本", "relation": "AdministrativeLocation"}] |
| SF | 列出评级最高的中式外卖 | ["food_type","order_type"] | [{"slot": "中式", "slot_type": "food_type"}, {"slot": "外卖", "slot_type": "order_type"}] |
| ED | Nguyên Phó chủ tịch Thanh Hoá ' xin bố trí công việc ' sau khi bị cách chức . | ["End-position"] | [{"event_trigger": "cách chức", "event_type": "End-position"}] |
| EE | The commodities trader who " bought " his appointment to a Top Secret nuclear weapons security advisory board . | [{"transaction.transferownership": {"artifact","giver"}}] | [{"event_trigger": "bought", "event_type": "transaction.transferownership", "arguments": [{"role": "giver", "argument": "commodities trader"}, {"role": "artifact", "argument": "appointment"}]}] |

Figure 1: Examples of IE tasks in 5 languages (Korean, Japanese, Chinese, Vietnamese, English)

we use the terms *instance* and *sample* to refer to pre-batched and after-batched inputs, respectively. Note that although batched schema increases the number of queries, it makes each query shorter and was also shown to improve LLMs' extraction performance (Gui et al., 2024).

**Refine Schema** Through manual inspection, we observe that a portion of ambiguous schemas, whose labels dynamically change in relation to other paired schemas, are not compatible with batched schema. These schemas are often characterized with *Misc*, *Other* or *Other-X*, etc. For example, given the input text *"Mary went to New York City for her grad studies"*, the NER labels for the schema *Other* in the three batches *[Other, Person]*, *[Other, Location]* and *[Other, Person, Location]* would all be different. As resolving these labeling issues on a sample-by-sample basis is beyond our budget, we manually examine the schema set of each dataset to identify and eliminate such schemas beforehand from all instances. In addition, we also observe highly abbreviated schemas in some datasets that are almost impossible to guess. For example, the schema set for the BANGLABIOMED dataset (Sazzed, 2022) is *[AN, CD, DS, P]* which, as noted in the authors' guidelines, translate to *[Anatomy, Chemicals and Drugs, Disease and Symptom, Medical Procedures]*. In some cases, we further observe schemas with potential overlapping label space. For example, the GERMANLERFINE dataset (Leitner et al., 2020) includes the three schemas *[AN, RR, PER]* which translate to *[Lawyer, Judge, Person]*. Here the *Person* schema, as noted by the authors, refers to *person entities that are neither lawyers nor judges*, but it's hard to infer such information implicitly unless the models were trained on such samples, and these cases are also not fully compatible with batched schema. To resolve the two scenarios mentioned, we manually annotate the field *supplementary_description* where we insert necessary descriptions of those schemas for prediction's sakes. **We stress that these problems were neglected in previous works and we are the first to address them for batched schema**.

### 2.3 Dual Versions: HEAVY and LIGHT

| | M-HEAVY | | | M-LIGHT | | |
|---|---|---|---|---|---|---|
| | Train | Dev | Test | Train | Dev | Test |
| RE | 127490 | 28384 | 116471 | 70600 | 2461 | 29068 |
| NER | 2023810 | 336215 | 844693 | 808595 | 13621 | 194588 |
| SF | 6199870 | 1089221 | 1688019 | 780789 | 9880 | 95119 |
| EE | 1510244 | 209193 | 215990 | 89918 | 2250 | 22002 |
| ED | 2049052 | 310980 | 282119 | 197930 | 4333 | 31087 |
| All | 11910466 | 1973993 | 3147292 | 1947832 | 32545 | 371864 |

Table 1: Number of samples in MASSIE: M-HEAVY and M-LIGHT

Ultimately, we obtained the postprocessed ver-

rather than specifically dealing with multilingual instructions

| " Mr Bure left his feet to deliver a forearm blow to Mr Galley as he was about to be checked legally by his opponent , " said NHL discipline chief Brian Burke in handing out the suspension . | | | | | |
|---|---|---|---|---|---|
| {"PERSON": ["Bure", "Galley", "Brian Burke"],"ORGANIZATION":["NHL"]} | | | | | |
| **Type** | **Model Output** | **LF1** | **F1** | **SacreBLEU** | **ROUGE-1** |
| Partially Correct | {"PERSON": ["Mr Bure", "Mr Galley", "Brian Burke"]} | 63.9 ✓ | 28.6 × | 50.2 ✓ | 71.4 ✓ |
| Wrong Structure | {"ORGANIZATION": ["Bure", "Galley"]} | 0.0 ✓ | 0.0 ✓ | 32.8 × | 60.0 × |
| Invalid Structure | {"PERSON": [{"Mr Bure", "Mr Galley","Brian Burke"}]]} | 0.0 ✓ | 0.0 ✓ | 45.0 × | 71.4 × |

Figure 2: Examples of LF1 scores versus other metrics (F1, SACREBLEU, ROUGE-1). Groundtruth labels are colored in yellow.

sion of each data split in MASSIE, which we aggregate to construct M-HEAVY - an intensive benchmark totaling over $17M$ samples. Given its sheer size, M-HEAVY is not suited for iterative development and only designed for intensive evaluations. In addition, we observe that M-HEAVY is highly sparse in some tasks e.g. over $85\%$ samples in slot filling and event tasks do not contain the designated span types to extract from. Thus, solely evaluating on M-HEAVY might create a favored bias towards conservative models i.e. those more likely to produce empty predictions. We therefore additionally construct M-LIGHT - a downsampled, more balanced variant of MASSIE that is better suited for iterative development and benchmarking.

**Source Selection** For each dataset $D$ of task $T$ with data split $D(T, S, L)$, where $S \in [Train, Dev, Test]$ and $L \sim D(L)$ is the set of languages included in $D$, we sample $x_{T,S,L} \sim D(T, S, L)$, thus treating all data splits as sampling targets. This serves to preserve the linguistic coverage of MASSIE and avoid overfitting (or favored attribution) to specific domains/datasets.

**Sparsity-Constrained Sampling** To impose a balance on samples with empty and non-empty labels, we decompose $x_{T,S,L} = x^E_{T,S,L} + x^H_{T,S,L}$ where $x^E_{T,S,L}$ and $x^H_{T,S,L}$ each represents samples with empty and non-empty (i.e. *have*) labels, and apply upperbound constraints $C_{S,E}$, $C_{S,H}$ such that $len(x^E_{T,S,L}) \leq C_{S,E}$ and $len(x^H_{T,S,L}) \leq C_{S,H}$, thus controlling the sparsity ratio while sampling from $D(T, S, L)$. We note that upperbound constraints are applicable to all data splits whereas setting a fixed proportion would not work as different data splits have varying numbers of candidates for $x^E_{T,S,L}$ and $x^H_{T,S,L}$. To construct M-LIGHT, we set $C_{Train,E} = 10000$, $C_{Dev,E} = 100$, $C_{Test,E} = 1000$, $C_{Train,H} = 3000$, $C_{Dev,H} = 30$, $C_{Test,H} = 300$.

### 2.4 Structure-aware soft evaluation

Since the standard F1-Score with exact matching does not reveal the difference between partially correct span outputs, we propose a modified metric based on the Levenshtein distance - which we term LF1.

Denote the list of query schemas as $[S_1, S_2..S_L]$, model predictions as $[Y_1, Y_2..Y_L]$, groundtruth labels as $[G_1, G_2..G_L]$. Here each $Y_I$ and $G_I$ correspondingly refer to the list of predicted and groundtruth spans for each schema $S_I$. LF1 score is then calculated as:

$$LF1^I_P = \frac{1}{|Y_I|}\sum_{i=1}^{|Y_I|}\max_{1\leq j\leq |G_I|} 1 - \frac{Levenshtein(y_i, g_j)}{\max(|y_i|, |g_j|)}$$

$$LF1^I_R = \frac{1}{|G_I|}\sum_{j=1}^{|G_I|}\max_{1\leq i\leq |Y_I|} 1 - \frac{Levenshtein(y_i, g_j)}{\max(|y_i|, |g_j|)}$$

$$LF1^I = \frac{2 * LF1^I_P * LF1^I_R}{LF1^I_P + LF1^I_R}; LF1 = \frac{1}{L}\sum_{I=1}^{L} LF1^I$$

Conceptually, LF1 involves two phases (i) structure routing and (ii) soft matching. The first phase routes prediction spans under the same direction[8] (single- or multi-way, dependent on task structures) and the second phase greedily matches each span with the closest ground truth one that has the smallest normalized Levenshtein distance, and subtracts this distance to obtain the soft score. **Compared to free-form metric such as ROUGE or BLEU which do not take into account the outputs' structures, LF1 is structure-aware while also distinguishes between partially correct spans**. In addition, LF1 is based on character-level Levenshtein which is both efficient and readily usable for multilingual evaluation, eliminating the need for language-specific tokenizers.

[8]For flat structures such as NER, *direction* can be interpreted as *category*. For hierarchical structures such as EE, routing would involve more than a single *category*. Thus, we use the term *direction* for generalizability's sake.

**Case Studies** We consider three evaluation scenarios illustrated in Figure 2: (i) *partially correct* predictions, (ii) *incorrect predictions* (*wrong structures*) and (iii) *invalid structures*. In (i), F1 overpenalizes the model and ignores partially correct predictions. Meanwhile, SACREBLEU and ROUGE-1 overly assign rewards to wrong and invalid structures in (ii) and (iii). In all three cases, LF1 functions as intended: sufficiently giving credits to partially correct predictions while ignoring predictions with incorrect or invalid structures.

Throughout the rest of this paper, we mainly experiment with M-LIGHT and report results in LF1 scores, unless explicitly mentioned otherwise.

## 3 In-Context Learning Evaluation

### 3.1 Settings

In this section, we conduct experiments on instruction-following multilingual LLMs and examine their *few-shot in-context* abilities on multilingual IE using M-LIGHT. As not all languages have training exemplars, for consistency's sake, we only selected English samples as demonstrations for inference in all languages. For each of the 5 tasks, we sample 3 exemplars from the training split of an English dataset[9] and fix this set in few-shot experiments (*default prompt*).

**Benchmarked Models** We select from a diverse range of models: MISTRAL (Jiang et al., 2023), MIXTRAL (Jiang et al., 2024), MINISTRAL (Team, 2024b), QWEN2.5 (Team, 2024a), QWEN2.5-CODER (Hui et al., 2024), LLAMA-3, LLAMA-3.1, LLAMA-3.2, LLAMA-3.3 (Grattafiori et al., 2024), AYA (Üstün et al., 2024), AYA-EXPANSE (Dang et al., 2024), GRANITE (Research, 2023), with variants sizing from 0.5B to 72B.

### 3.2 Results and Analyses

**M-LIGHT is challenging for all LLMs:** Our few-shot evaluation results on M-LIGHT with the default prompts are shown in Figure 3. We group LLMs into 4 categories dependent on each model's size for better comparison. Most LLMs under 3B fail to generate any reasonable output, except for the QWEN2.5-3B-INSTRUCT model. At the 7B-8B scales, LLMs start achieving reasonable results, but overall scores remain quite low (e.g. below 40). For the 14B-35B group, LLMs reach better performance at simple tasks (NER,ED,SF) but still underperform at complex tasks (EE,RE). Above 56B, most LLMs reach around 60 LF1 on the three simple tasks but no model reaches above (or even near) 40 LF1 on the two complex tasks.

**Scaling model size largely improves results:** We group each model family over several parameter scales in Figure 4. We notice a monotonic trend in that, for the same model family, as the number of parameters increase, we obtain significantly better results in each task. For the QWEN2.5 series where multiple scales are released, this trend is clearly visible and improvement is displayed incrementally. When the parameter gap is large, the increases in performance become more substantial (e.g. LLAMA-3.1-70B-INSTRUCT performs several times better than LLAMA-3.1-8B-INSTRUCT). These results show that neural scaling laws (Kaplan et al., 2020) still hold for multilingual IE under large-scale evaluation, which was not confirmed in previous works.

**Scaling exemplars is often harmful:** We experiment with increasing the number of exemplars used in few-shot prompting for 4 languages: German, Thai, Spanish, Japanese (Figure 5[10]). In most cases, adding exemplars either do not yield any improvement or further reduce performance compared to only using 3 exemplars. We also do not observe any consistent trend in performance change, as it usually fluctuates in most cases. Initially, we suspect that this could be due to the usage of English exemplars and that language-specific exemplars might have a different effect. Thus, we repeat this experiment but replacing English exemplars with language-specific exemplars e.g. using German exemplars when evaluating on German datasets. The results (Figure 13), however, display the same phenomena as observed with English exemplars. In fact, these findings also coincide with the conclusions in Asai et al., 2024 where they similarly observed that scaling exemplars was detrimental for understanding and generation tasks.

**Performance varies significantly among demonstration sets:** For each task, we evaluate 5 different sets of English demonstrations in Figure 6. Through observing the interquartile range, we notice large disparity in performance among demonstration sets. The difference is especially big for the Thai and German languages, while smaller (but still significant) for the Spanish and Japanese languages. This suggest that we can further improve

[9] See Appendix I

[10] The tasks shown in each language depend on availability of datasets in that specific language.

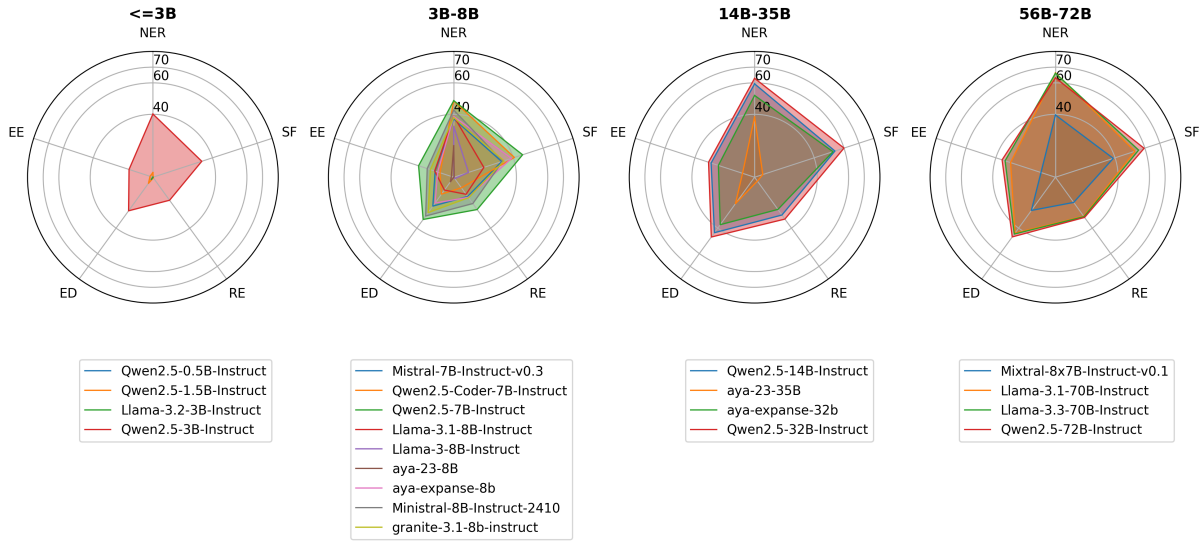


Figure 3: Overall results with few-shot default prompts (M-LIGHT)

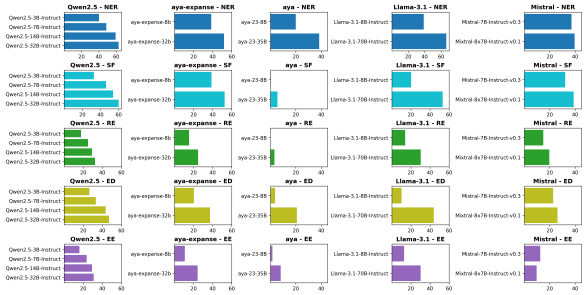

Figure 4: Parameter scaling with few-shot default prompts (M-LIGHT)

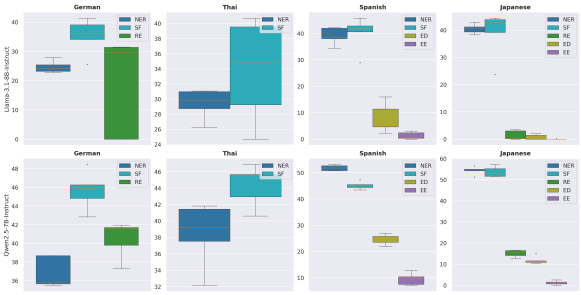

Figure 6: Variance among demonstration sets with few-shot default prompts (M-LIGHT)

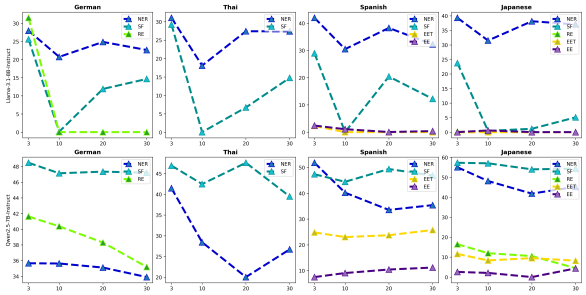

Figure 5: Varying number of demonstrations with few-shot default prompts (M-LIGHT)

performance by crafting a more fitting demonstration set, which we will explore in Sec. 5.

**Performance varies significantly among languages:** We observe significant performance gap between languages for the same model and task in M-LIGHT (Figure 16). Since M-LIGHT is heterogeneous, this phenomenon could have been due to the variance in data difficulty among languages. To clarify, we further examined model performance on the MASSIVE11 dataset (with full samples from M-HEAVY, Figure 17) which is highly parallel as the non-English instances were translated from the English ones. However, we still observe large disparity among different languages for the same model, which suggests that language-wise performance variance is not caused simply by dataset heterogeneity, but rather due to imbalanced allocation of linguistic capacities in LLMs. Interestingly, similar-performing language group of each model often overlaps with that of other models. For example, German and Chinese scores for RE (M-LIGHT) always remain in the top 3 of displayed models (Figure 16). Similarly, Spanish and English scores for MASSIVE11 (M-HEAVY) always rank in the top 3 (Figure 17). In contrast, Amharic or Mongolian scores always remain in the bottom 3 while Bengali and Arabic always place in the middle range (Figure 17).

## 4 Incorporate Reasoning

### 4.1 Settings

In some tasks such as Math (Wei et al., 2023) or Rating (Zheng et al., 2023), prompting the language model to first generate natural language reasoning before producing the output has been shown to improve performance. We are interested in whether this observation holds for IE, particularly with multilingual text and schemas. To execute this, we sample thought $T \sim P_\Theta(T|x, y)$[11], where $\Theta$ is a reasoning teacher, for each demonstration $(x, y)$ in the default prompt and prepend $T$ to its output. This way, the inference model $\theta$ learns in-context to first generate reasoning $T$ before producing the required output.

**Model Choices** We select GPT-4O, LLAMA-3.1-70B-INSTRUCT or the inference model itself as the reasoning teacher $\Theta$. For $\theta$, we either use LLAMA-3.1-8B-INSTRUCT or QWEN-2.5-7B-INSTRUCT.

### 4.2 Results and Analyses

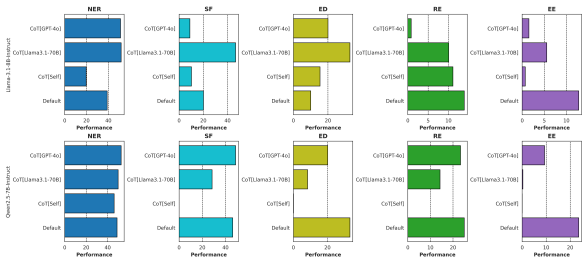


Figure 7: Default prompts with intermediate reasoning (M-LIGHT)

**Reasoning produces mixed results** We observe that incorporating reasoning might damage or improve performance over the default prompts, dependent on specific combinations of task, $\Theta$ and $\theta$. For example, reasoning seems to help on more simple tasks such as NER, SF or ED (considering the most performant teacher), but gives negative results on more complex tasks such as RE and EE, ironically[12]. While the performant boost can be enormous in some cases (e.g. more than $2X$ gains in SF and ED with LLAMA-3.1-INSTRUCT-70B and LLAMA-3.1-INSTRUCT-8B as $\Theta$ and $\theta$), the damages can be rather huge (e.g. over $90\%$ reduction in LF1 scores of RE and EE with GPT-4O and LLAMA-3.1-8B-INSTRUCT as $\Theta$ and $\theta$).

**No universal "best teacher"** As we vary $\Theta$ and $\theta$, we observe that there is no shared best teacher for all inference language models. For LLAMA-3.1-8B-INSTRUCT, LLAMA-3.1-70B-INSTRUCT is often the best teacher (highest scores on $4/5$ tasks). Meanwhile, for QWEN-2.5-7B-INSTRUCT, GPT-4O is always the best teacher (highest scores on $5/5$ tasks).

**Self-generated thoughts (7B-8B) reduce performance** We find that using self-generated thoughts for both LLAMA-3.1-8B-INSTRUCT and QWEN-2.5-7B-INSTRUCT mostly reduce performance compared to the default prompt ($9/10$ cases), and usually underperform the remaining two teachers. This suggests that intermediate thoughts sampled from a more capable teacher (e.g. LLAMA-3.1-70B-INSTRUCT or GPT-4O) yield better results, while thoughts derived from less capable models might even harm performance.

To better mitigate the observed negative impact of incorporating reasoning, we further explore alternative prompting variants in Appendix D, which expresses more stable improvement for reasoning-based IE.

## 5 Automatic Prompt Optimization

### 5.1 Settings

Previous IE works only adopt a set of manual prompts with little variation (Wang et al., 2023b; Qi et al., 2024). Meanwhile, in-context performance of LLMs might vary greatly depending on the prompt being used (Zhuo et al., 2024; Chatterjee et al., 2024). In Section 3, we also witnessed substantial variance depending on the demonstration set being used. In this section, we are interested in exploring whether automatically searching for an optimal prompt in pre-selected datasets can widely benefit multilingual IE. Conceptually, we optimize a prompt $P_{\Phi,\mu}(T, I, S, D)$ for each task $T$ with the instruction and demonstration variables $I$ and $S$, based on an optimizer LM $\Phi$, a metric $\mu$ and a tuning dataset $D$ (whose English training and development splits are used to optimize this prompt).

Concretely, we build on top of the MIPRO framework (Opsahl-Ong et al., 2024). First, we bootstrap $m$ sets of few-shot examples, each with a maximum size of $K$, featuring both (randomly selected) labeled and augmented samples. Each

[11] For simplicity, we use greedy decoding (i.e. temperature 0)

[12] Intermediate thoughts were traditionally introduced to solve complex problems (Wei et al., 2023)

set's size is fixed to $K$[13], and augmented samples $(x, y')$ (whose quantities randomly vary from 1 to $K$) are constructed by replacing the ground truth $y$ in labeled samples $(x, y)$ with $y' = \Phi(x)$. Second, we prompt $\Phi$ to generate candidate instructions $\{I\}$. Inside the prompt, we also provided $\Phi$ with few-shot examples bootstrapped previously, a generated dataset summary, a randomly selected prompting tip [14] and a summary of the program code. We note that providing such auxiliary information has been shown to improve instruction quality in several downstream tasks (Opsahl-Ong et al., 2024). Lastly, we iteratively search among combinations of $I$ and $S$ through Bayesian optimization (Watanabe, 2023; Akiba et al., 2019) and choose the best combination that maximize $\mu(\Phi(D_{Eng,Validation}))$. Since iteratively executing full evaluation over $D_{Eng,Validation}$ would be expensive, we further apply minibatching and approximate the evaluation score using a randomly sampled minibatch $\mathcal{B} \in D_{Eng,Validation}$ in each iteration, and only execute full evaluation on the best found combination[15] every $f$ iterations. Lastly, we integrate the combination with the highest full evaluation score into $P_{\Phi,\mu}(T, I, S, D)$ for test time inference. Note that the language model $\theta$ adopted at inference time might differ from $\Phi$.

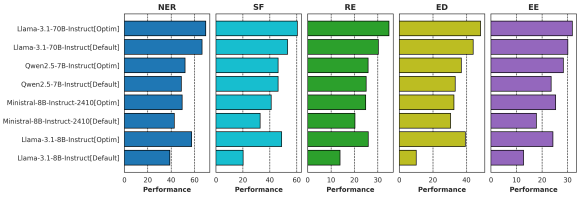


Figure 8: Varying $\theta$ with few-shot optimized prompts (M-LIGHT)

**Hyperparameters** To maintain consistency with the default prompts, we use $K = 3$ and adopt the same choices of $D$ for each task. Besides, we set $m = 12, ||\mathcal{B}|| = 25, f = 10$ and choose LF1 as $\mu$. To sample instruction proposals, we set temperate $T_{init} = 0.5$. For Bayesian optimization, we set the maximum number of iterations as 50.

**Models** For $\Phi$ (prompt teacher), we use LLAMA-3.1-8B-INSTRUCT in most cases (unless explicitly specified otherwise) and choose $\theta$ (inference model) among $\Phi$, LLAMA-3.1-70B-INSTRUCT, QWEN2.5-7B-INSTRUCT and MINISTRAL-8B-INSTRUCT-2410.

[13]Technically, we do consider an empty set as a candidate (zero-shot), though, unsurprisingly, it was never selected as the top candidate

[14]We follow (Opsahl-Ong et al., 2024) for the list of prompting tips

[15]Combination with the highest mini-batch score

### 5.2 Results and Analyses

**Optimized prompts bring significant and consistent improvements** Compared to the default prompts, we see that optimized prompts yield consistent improvements across all tasks (Figure 8). At times, the gains can be as large as over 20 LF1 scores e.g. ED and SF tasks with LLAMA-3.1-8B-INSTRUCT as $\theta$.

**Optimized prompts can transfer across model families and scales** While all models in Figure 8 re-use the same optimized prompts (LLAMA-3.1-8B-INSTRUCT as $\Phi$), we also observe improvements when $\theta$ belongs to a family different from $\Phi$ e.g. both MINISTRAL-8B-INSTRUCT-2410 and QWEN2.5-7B-INSTRUCT (as $\theta$) perform better with optimized prompts than default prompts across all tasks. In addition, prompts optimized by LLAMA-3.1-8B-INSTRUCT also yield improvements when directly applied to LLAMA-3.1-70B-INSTRUCT, suggesting the set of high-performing prompts might be shared among model scales.

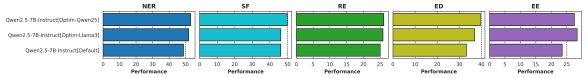


Figure 9: Varying $\Phi$ with few-shot optimized prompts (M-LIGHT)

**Better results if prompt teacher acts as inference model** With QWEN2.5-7B-INSTRUCT as $\theta$, we vary $\Phi$ between QWEN2.5-7B-INSTRUCT and LLAMA-3.1-8B-INSTRUCT, and observed better results with the prior (Figure 9). This suggests that when $\Phi \neq \theta$, there exists an optimization gap and that better performance can be achieved with $\Phi \equiv \theta$.

| | NER | SF | RE | ED | EE | Avg. |
|---|---|---|---|---|---|---|
| OPTIM-F1 | 45.56 | 36.89 | 5.73 | 26.65 | 17.91 | 26.55 |
| OPTIM-LF1 | 45.61 | 36.65 | 8.91 | 29.70 | 17.00 | **27.57** |

Table 2: Prompt optimization results with LLAMA-3.1-8B-INSTRUCT ($\Phi \equiv \theta$). Each row denotes exact scores (F1) with either F1 or LF1 as $\mu$.

**LF1 as guiding metric improves both exact and soft scores** We vary $\mu$ between F1 and LF1, and report according scores in Table 2 and 3. On average, we observe stronger results (both exact and soft scores) with LF1 as the guiding metric,

| | NER | SF | RE | ED | EE | Avg. |
|---|---|---|---|---|---|---|
| OPTIM-F1 | 57.74 | 49.45 | 16.90 | 35.96 | 25.61 | 37.13 |
| OPTIM-LF1 | 57.53 | 48.86 | 25.91 | 39.25 | 24.28 | **39.17** |

Table 3: Prompt optimization results with LLAMA-3.1-8B-INSTRUCT ($\Phi \equiv \theta$). Each row denotes soft scores (LF1) with either F1 or LF1 as $\mu$.

especially on RE and ED tasks. We note, however, that there exist cases with decreasing scores (e.g. EE), which suggests the best choice of $\mu$ for each task might differ.

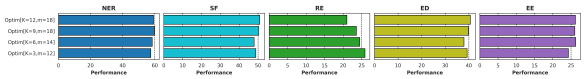


Figure 10: Scaling number of demonstrations in optimized prompts (M-LIGHT)

**Scaling exemplars with prompt optimization yields improvements on most tasks** As we increase the values of $K$ and $m$ in prompt optimization, we observe a rising (though not monotonic) tendency in LF1 scores for $4/5$ tasks. We notice an exception with the RE task where LF1 scores keep decreasing instead, suggesting that scaling exemplars can be harmful even with prompt optimization. Nevertheless, these results show that with proper searching and optimization, a higher number of exemplars can often yield better results, which would go unnoticed otherwise if we only observe the default prompts due to high variance induced by varying demonstration sets and non-optimal instructions.

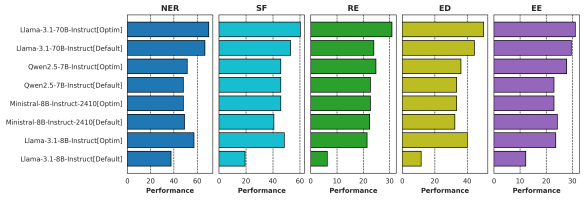


Figure 11: Non-English results with few-shot optimized prompts (M-LIGHT)

**Optimizing prompts on single English datasets improves non-English performance** Earlier in Figure 8, we aggregated LF1 scores of all languages in M-LIGHT. Since each optimized prompt was tuned on an English dataset, we are interested in whether improvements came from the English language only i.e. if these optimized prompts also benefit other languages'. We thus re-aggregated LF1 scores of non-English languages in Figure 11. To our surprise, significant improvements persist in most cases, showing that prompt optimization from one language can transfer to improve IE performance of other languages as well.

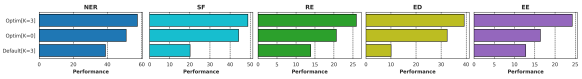


Figure 12: Zero-shot optimized prompts versus few-shot prompts (M-LIGHT)

**Zero-shot optimized prompts outperform few-shot default prompts** Our initial efforts on manually prompting the LLMs with instructions only (zero-shot) often fail to obtain desirable outputs compared to the few-shot alternatives. With prompt optimization, we re-visit this setting and ask whether we could achieve sufficiently good results through solely tuning instruction candidates. Particularly, we re-conduct the optimization process without groundtruth labeled samples and only search for the best instruction $I_{top}$ among generated instruction candidates $\{I\}$. At test time, we directly use $I_{top}$ without any demonstration. To our surprise (Figure 12), zero-shot optimized prompts consistently perform better than few-shot default prompts, often by a substantial margin. However, zero-shot optimized prompts still underperform few-shot optimized prompts in all tasks, as we expected. This suggests that even with carefully optimized instructions, the existence of a demonstration set is still vital to achieving better results.

## 6 Conclusion

In this paper, we identify and work towards 6 challenges of current research for multilingual IE. In particular, we construct the MASSIE collection and the two associated benchmarks: M-HEAVY and M-LIGHT, both featuring samples from 215 human annotated IE datasets in 96 languages. Besides, we introduce LF1 - a structure-aware soft metric for continuous evaluation of multilingual IE. Additionally, we conduct an empirical study involving 21 LLMs on few-shot ICL and re-examine important factors influencing models' performance. Beyond manual default prompts, we also experiment with integrating intermediate reasoning and automatically tuning prompt components, providing high-level analyses in the process. Although our results indicate that existing LLMs still struggle with multilingual IE, we believe the contained resources and findings would help pave the wave for new frontiers in this area.

## Limitations

The MASSIE collection in this work represents significant efforts in unifying language varieties within the research agenda of multilingual IE. That said, among the 500 institutional languages (Bird, 2024), roughly fewer than $1/5$ of them are currently covered in MASSIE, showing much room for improvements. Therefore, we intend for MASSIE to be a continually updating collection and plan to further expand MASSIE through broad community efforts, similar to the development of MMTEB (Enevoldsen et al., 2025).

Additionally, MASSIE's complete public release is not straightforward as $6/215$ datasets in the collection are not publicly accessible, requiring further contacts with the authors or LDC[16]. We plan to first release postprocessed samples of permissively licensed datasets only and provide guidelines to obtain those not yet released. Practitioners who already gained access to those datasets can also contact us to directly obtain postprocessed samples.

Besides, due to lack of budget, we only experiment with open-source LLMs and could not include closed-source LLMs such as CLAUDE 3.5[17], GPT-4 (OpenAI et al., 2024) or GEMINI 1.5 (Team et al., 2024). Given non-trivial API costs, evaluating these closed-source models at large scale is extremely expensive and goes beyond what we could afford.

Lastly, our work focuses on the text modality, but future works can expand to other modalities such as audio, image or video, which have received increasing interest from the NLP community (Perot et al., 2024; Hachmeier and Jäschke, 2024; An et al., 2023; Nguyen et al., 2023b, 2024a,c).

## Acknowledgements

This research has been supported by the NSF grant # 2239570. This research is also supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via the HIATUS Program contract 2022-22072200003. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government.

[16] https://www.ldc.upenn.edu/

[17] https://www.anthropic.com/news/claude-3-5-sonnet

## References

David Adelani, Graham Neubig, Sebastian Ruder, Shruti Rijhwani, Michael Beukman, Chester Palen-Michel, Constantine Lignos, Jesujoba Alabi, Shamsuddeen Muhammad, Peter Nabende, Cheikh M. Bamba Dione, Andiswa Bukula, Rooweither Mabuya, Bonaventure F. P. Dossou, Blessing Sibanda, Happy Buzaaba, Jonathan Mukiibi, Godson Kalipe, Derguene Mbaye, Amelia Taylor, Fatoumata Kabore, Chris Chinenye Emezue, Anuoluwapo Aremu, Perez Ogayo, Catherine Gitau, Edwin Munkoh-Buabeng, Victoire Memdjokam Koagne, Allahsera Auguste Tapo, Tebogo Macucwa, Vukosi Marivate, Mboning Tchiaze Elvis, Tajuddeen Gwadabe, Tosin Adewumi, Orevaoghene Ahia, and Joyce Nakatumba-Nabende. 2022. MasakhaNER 2.0: Africa-centric transfer learning for named entity recognition. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4488–4508, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Noëmi Aepli, Çağrı Çöltekin, Rob Van Der Goot, Tommi Jauhiainen, Mourhaf Kazzaz, Nikola Ljubešić, Kai North, Barbara Plank, Yves Scherrer, and Marcos Zampieri. 2023. Findings of the VarDial evaluation campaign 2023. In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, pages 251–261, Dubrovnik, Croatia. Association for Computational Linguistics.

Rodrigo Agerri, Xavier G'omez Guinovart, German Rigau, and Miguel Anxo Solla Portela. 2018. Developing new linguistic resources and tools for the Galician language. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Gustavo Aguilar, Fahad AlGhamdi, Victor Soto, Mona Diab, Julia Hirschberg, and Thamar Solorio. 2018. Named Entity Recognition on Code-Switched Data: Overview of the CALCS 2018 Shared Task. In *Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching*, pages 138–147, Melbourne, Australia. Association for Computational Linguistics.

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. In *The 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2623–2631.

Abdulmohsen Al-Thubaity, Sakhar Alkhereyf, Wejdan Alzahrani, and Alia Bahanshal. 2022. CAraNER: The COVID-19 Arabic named entity corpus. In *Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 1–10, Abu

Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Jesujoba Alabi, Kwabena Amponsah-Kaakyire, David Adelani, and Cristina España-Bonet. 2020. Massive vs. curated embeddings for low-resourced languages: the case of Yorùb'a and Twi. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2754–2762, Marseille, France. European Language Resources Association.

Iñaki Alegria, Olatz Arregi, Irene Balza, Nerea Ezeiza, Izaskun Fern'andez, and Ruben Urizar. 2004. Design and development of a named entity recognizer for an agglutinative language.

Wazir Ali, Junyu Lu, and Zenglin Xu. 2020. SiNER: A large dataset for Sindhi named entity recognition. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2953–2961, Marseille, France. European Language Resources Association.

Shengnan An, Zexiong Ma, Zeqi Lin, Nanning Zheng, Jian-Guang Lou, and Weizhu Chen. 2024. Make your LLM fully utilize the context. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Siyu An, Ye Liu, Haoyuan Peng, and Di Yin. 2023. VKIE: The application of key information extraction on video text. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 532–540, Singapore. Association for Computational Linguistics.

Spela Arhar Holdt, Simon Krek, Kaja Dobrovoljc, Tomaz Erjavec, Polona Gantar, Jaka Cibej, Eva Pori, Luka Tercon, Tina Munda, Slavko Zitnik, Nejc Robida, Neli Blagus, Sara Moze, Nina Ledinek, Nanika Holz, Katja Zupan, Taja Kuzman, Teja Kavcic, Iza Skrjanec, Dafne Marko, Lucija Jezersek, and Anja Zajc. 2024. Training corpus SUK 1.1. Slovenian language resource repository CLARIN.SI.

Jordi Armengol-Estap'e, Casimiro Pio Carrino, Carlos Rodriguez-Penagos, Ona de Gibert Bonet, Carme Armentano-Oller, Aitor Gonzalez-Agirre, Maite Melero, and Marta Villegas. 2021. Are multilingual models the best choice for moderately under-resourced languages? A comprehensive assessment for Catalan. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4933–4946, Online. Association for Computational Linguistics.

Ekaterina Artemova, Tatiana Batura, Anna Golenkovskaya, Vitaly Ivanin, Vladimir Ivanov, Veronika Sarkisyan, Ivan Smurov, and Elena Tutubalina. 2020. So what's the plan? mining strategic planning documents. In *Digital Transformation and Global Society: Proceedings of the 5th International Conference (DTGS 2020)*, St. Petersburg, Russia.

Masayuki Asahara, Hiroshi Kanayama, Takaaki Tanaka, Yusuke Miyao, Sumire Uematsu, Shinsuke Mori, Yuji Matsumoto, Mai Omura, and Yugo Murawaki. 2018. Universal Dependencies version 2 for Japanese. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Akari Asai, Sneha Kudugunta, Xinyan Yu, Terra Blevins, Hila Gonen, Machel Reid, Yulia Tsvetkov, Sebastian Ruder, and Hannaneh Hajishirzi. 2024. BUFFET: Benchmarking large language models for few-shot cross-lingual transfer. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1771–1800, Mexico City, Mexico. Association for Computational Linguistics.

Majid Asgari-Bidhendi, Mehrdad Nasser, Behrooz Janfada, and Behrouz Minaei-Bidgoli. 2020. Perlex: A bilingual persian-english gold dataset for relation extraction. *ArXiv*, abs/2005.06588.

Dan Bareket and Reut Tsarfaty. 2021. Neural Modeling for Named Entities and Morphology (NEMO2). *Transactions of the Association for Computational Linguistics*, 9:909–928.

Nikos Bartziokas, Thanassis Mavropoulos, and Constantine Kotropoulos. 2020. Datasets and Performance Metrics for Greek Named Entity Recognition. In *11th Hellenic Conference on Artificial Intelligence (SETN 2020)*, SETN 2020, pages 160–167, New York, NY, USA. Association for Computing Machinery.

Elisa Bassignana and Barbara Plank. 2022. CrossRE: A cross-domain dataset for relation extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3592–3604, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Vuk Batanovi'c, Nikola Ljubesi'c, Tanja Samardzi'c, and Tomaz Erjavec. 2023. Serbian linguistic training corpus SETimes.SR 2.0. Slovenian language resource repository CLARIN.SI.

Darina Benikova, Chris Biemann, Max Kisselew, and Sebastian Pad'o. 2014. Germeval 2014 named entity recognition shared task.

Steven Bird. 2024. Must NLP be extractive? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14915–14929, Bangkok, Thailand. Association for Computational Linguistics.

Baptiste Blouin, C'ecile Armand, and Christian Henriot. 2024. A dataset for named entity recognition and entity linking in Chinese historical newspapers. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 385–394, Torino, Italia. ELRA and ICCL.

Prachya Boonkwan, Vorapon Luantangsrisuk, Sitthaa Phaholphinyo, Kanyanat Kriengket, Dhanon Leenoi, Charun Phrombut, Monthika Boriboon, Krit Kosawat, and Thepchai Supnithi. 2020. The annotation guideline of lst20 corpus. *arXiv preprint arXiv:2008.05055*.

Savong Bou, Naoki Suzuki, Makoto Miwa, and Yutaka Sasaki. 2020. Ontology-style relation annotation: A case study. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4867–4876, Marseille, France. European Language Resources Association.

Elena Bruches, Alexey Pauls, Tatiana Batura, and Vladimir Isachenko. 2020. Entity recognition and relation extraction from scientific and technical texts in russian. In *2020 Science and Artificial Intelligence conference (SAI ence)*, pages 41–45.

Weerayut Buaphet, Can Udomcharoenchaikit, Peerat Limkonchotiwat, Attapol Rutherford, and Sarana Nutanong. 2022. Thai nested named entity recognition corpus. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1473–1486, Dublin, Ireland. Association for Computational Linguistics.

MohanRaj Chanthran, Lay-Ki Soon, Huey Fang Ong, and Bhawani Selvaretnam. 2024. Malaysian English news decoded: A linguistic resource for named entity and relation extraction. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10999–11022, Torino, Italia. ELRA and ICCL.

Dmytro Chaplynskyi and Mariana Romanyshyn. 2024. Introducing NER-UK 2.0: A rich corpus of named entities for Ukrainian. In *Proceedings of the Third Ukrainian Natural Language Processing Workshop (UNLP) @ LREC-COLING 2024*, pages 23–29, Torino, Italia. ELRA and ICCL.

Anwoy Chatterjee, H S V N S Kowndinya Renduchintala, Sumit Bhatia, and Tanmoy Chakraborty. 2024. POSIX: A prompt sensitivity index for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14550–14565, Miami, Florida, USA. Association for Computational Linguistics.

Harsh Vijay Chaudhari, Anuja Dinesh Patil, Dhanashree Lavekar, Pranav Khairnar, and Raviraj Joshi. 2024. L3cube-mahasocialner: A social media based marathi named entity recognition dataset and bert models. In *Proceedings of the 15th Annual Meeting of the Forum for Information Retrieval Evaluation*, FIRE '23, page 93–100, New York, NY, USA. Association for Computing Machinery.

Liang Chen, Shuo Xu, Lijun Zhu, Jing Zhang, Xiaoping Lei, and Guancan Yang. 2020. A deep learning based method for extracting semantic information from patent documents. *Scientometrics*, 125(1):289–312.

Pei Chen, Haotian Xu, Cheng Zhang, and Ruihong Huang. 2022. Crossroads, buildings and neighborhoods: A dataset for fine-grained location recognition. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3329–3339, Seattle, United States. Association for Computational Linguistics.

Yi-Pei Chen, An-Zi Yen, Hen-Hsen Huang, Hideki Nakayama, and Hsin-Hsi Chen. 2023. LED: A dataset for life event extraction from dialogs. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 384–398, Dubrovnik, Croatia. Association for Computational Linguistics.

Chiamaka Chukwuneke, Ignatius Ezeani, Paul Rayson, and Mahmoud El-Haj. 2022. IgboBERT models: Building and training transformer models for the Igbo language. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5114–5122, Marseille, France. European Language Resources Association.

Sandra Collovini, Gabriel Machado, and Renata Vieira. 2016. A sequence model approach to relation extraction in Portuguese. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1908–1912, Portoroz, Slovenia. European Language Resources Association (ELRA).

Camiel Colruyt, Orph'ee De Clercq, Thierry Desot, and Veronique Hoste. 2022. Eventdna: a dataset for dutch news event extraction as a basis for news diversification. *Language Resources and Evaluation*, 57:189 – 221.

Wei Congcong, Feng Zhenbing, Huang Shutan, Li Wei, and Shao Yanqiu. 2023. CHED: A cross-historical dataset with a logical event schema for classical Chinese event detection. In *Proceedings of the 22nd Chinese National Conference on Computational Linguistics*, pages 875–888, Harbin, China. Chinese Information Processing Society of China.

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, Sandra Kublik, Meor Amer, Viraat Aryabumi, Jon Ander Campos, Yi-Chern Tan, Tom Kocmi, Florian Strub, Nathan Grinsztajn, Yannis Flet-Berliac, Acyr Locatelli, Hangyu Lin, Dwarak Talupuru, Bharat Venkitesh, David Cairuz, Bowen Yang, Tim Chung, Wei-Yin Ko, Sylvie Shang Shi, Amir Shukayev, Sammie Bae, Aleksandra Piktus, Roman Castagné, Felipe Cruz-Salinas, Eddie Kim, Lucas Crawhall-Stein, Adrien Morisot, Sudip Roy, Phil Blunsom, Ivan Zhang, Aidan Gomez, Nick Frosst, Marzieh Fadaee, Beyza Ermis, Ahmet Üstün, and Sara Hooker. 2024. Aya expanse: Combining research breakthroughs for a new multilingual frontier. *Preprint*, arXiv:2412.04261.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning

of quantized llms. In *Advances in Neural Information Processing Systems*, volume 36, pages 10088–10115. Curran Associates, Inc.

George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, and Ralph Weischedel. 2004. The automatic content extraction (ACE) program – tasks, data, and evaluation. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Rezarta Islamaj Doğan, Robert Leaman, and Zhiyong Lu. 2014. Ncbi disease corpus: a resource for disease name recognition and concept normalization. *Journal of biomedical informatics*, 47:1–10.

Stefan Daniel Dumitrescu and Andrei-Marius Avram. 2020. Introducing RONEC - the Romanian named entity corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4436–4443, Marseille, France. European Language Resources Association.

Seth Ebner, Patrick Xia, Ryan Culkin, Kyle Rawlins, and Benjamin Van Durme. 2020. Multi-sentence argument linking. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Maud Ehrmann, Matteo Romanello, Sven Najem-Meyer, Antoine Doucet, and Simon Clematide. 2022. Extended overview of HIPE-2022: Named Entity Recognition and Linking in Multilingual Historical Documents. In *Proceedings of the Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum*, volume 3180. CEUR-WS.

Roald Eiselen. 2016. Government domain named entity recognition for South African languages. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3344–3348, Portoroz, Slovenia. European Language Resources Association (ELRA).

Kenneth Enevoldsen, Isaac Chung, Imene Kerboua, Márton Kardos, Ashwin Mathur, David Stap, Jay Gala, Wissam Siblini, Dominik Krzemiński, Genta Indra Winata, Saba Sturua, Saiteja Utpala, Mathieu Ciancone, Marion Schaeffer, Diganta Misra, Shreeya Dhakal, Jonathan Rystrøm, Roman Solomatin, Ömer Veysel Çağatan, Akash Kundu, Martin Bernstorff, Shitao Xiao, Akshita Sukhlecha, Bhavish Pahwa, Rafał Poświata, Kranthi Kiran GV, Shawon Ashraf, Daniel Auras, Björn Plüster, Jan Philipp Harries, Loïc Magne, Isabelle Mohr, Dawei Zhu, Hippolyte Gisserot-Boukhlef, Tom Aarsen, Jan Kostkan, Konrad Wojtasik, Taemin Lee, Marek Suppa, Crystina Zhang, Roberta Rocca, Mohammed Hamdy, Andrianos Michail, John Yang, Manuel Faysse, Aleksei Vatolin, Nandan Thakur, Manan Dey, Dipam Vasani, Pranjal A Chitale, Simone Tedeschi, Nguyen Tai, Artem Snegirev, Mariya Hendriksen, Michael Günther, Mengzhou Xia, Weijia Shi, Xing Han Lù, Jordan Clive, Gayatri K, Maksimova Anna, Silvan Wehrli, Maria Tikhonova, Henil Shalin Panchal, Aleksandr Abramov, Malte Ostendorff, Zheng Liu, Simon Clematide, Lester James Validad Miranda, Alena Fenogenova, Guangyu Song, Ruqiya Bin Safi, Wen-Ding Li, Alessia Borghini, Federico Cassano, Lasse Hansen, Sara Hooker, Chenghao Xiao, Vaibhav Adlakha, Orion Weller, Siva Reddy, and Niklas Muennighoff. 2025. MMTEB: Massive multilingual text embedding benchmark. In *The Thirteenth International Conference on Learning Representations*.

Elena Epure and Romain Hennequin. 2023. A human subject study of named entity recognition (ner) in conversational music recommendation queries. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*.

Alexander Erdmann, David Joseph Wrisley, Benjamin Allen, Christopher Brown, Sophie Cohen-Bod'enès, Micha Elsner, Yukun Feng, Brian Joseph, B'eatrice Joyeux-Prunel, and Marie-Catherine de Marneffe. 2019. Practical, efficient, and customizable active learning for named entity recognition in the digital humanities. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2223–2234, Minneapolis, Minnesota. Association for Computational Linguistics.

Jack FitzGerald, Christopher Hench, Charith Peris, Scott Mackie, Kay Rottmann, Ana Sanchez, Aaron Nash, Liam Urbach, Vishesh Kakarala, Richa Singh, Swetha Ranganath, Laurie Crist, Misha Britan, Wouter Leeuwis, Gokhan Tur, and Prem Natarajan. 2023. MASSIVE: A 1M-example multilingual natural language understanding dataset with 51 typologically-diverse languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4277–4302, Toronto, Canada. Association for Computational Linguistics.

Neda Foroutan, Markus Schr"oder, and Andreas Dengel. 2024. CO-fun: A German dataset on company outsourcing in fund prospectuses for named entity recognition and relation extraction. In *Proceedings of the 20th Conference on Natural Language Processing (KONVENS 2024)*, pages 117–122, Vienna, Austria. Association for Computational Linguistics.

Kata G'abor, Davide Buscaldi, Anne-Kathrin Schumann, Behrang QasemiZadeh, Ha"ıfa Zargayouna, and Thierry Charnois. 2018. SemEval-2018 task 7: Semantic relation extraction and classification in scientific papers. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 679–688, New Orleans, Louisiana. Association for Computational Linguistics.

Ziqi Gao, Qichao Wang, Aochuan Chen, Zijing Liu, Bingzhe Wu, Liang Chen, and Jia Li. 2024.

Parameter-efficient fine-tuning with discrete fourier transform. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*.

Marcos Garcia. 2016. Incorporating lexico-semantic heuristics into coreference resolution sieves for named entity recognition at document-level. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3357–3361, Portoroz, Slovenia. European Language Resources Association (ELRA).

Tsolak Ghukasyan, Garnik Davtyan, K. Avetisyan, and Ivan Andrianov. 2018. pioner: Datasets and baselines for armenian named entity recognition. *2018 Ivannikov Ispras Open Conference (ISPRAS)*, pages 56–61.

Bastien Giordano, Maxime Prieur, Nakanyseth Vuth, Sylvain Verdy, Kévin Cousot, Gilles Sérasset, Guillaume Gadek, Didier Schwab, and Cédric Lopez. 2024. Popcorn: Fictional and synthetic intelligence reports for named entity recognition and relation extraction tasks. *Procedia Computer Science*, 246:1170–1180. 28th International Conference on Knowledge Based and Intelligent information and Engineering Systems (KES 2024).

Denis Gordeev, Moscow Russia Ranepa, Adis Davletov, Alexey Rey, G. R. Akzhigitova, and G. A. Geymbukh. 2020. Relation extraction dataset for the russian. *Computational Linguistics and Intellectual Technologies*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty,

Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Felix Grezes, Sergi Blanco-Cuaresma, Thomas Allen, and Tirthankar Ghosal. 2022. Overview of the first shared task on detecting entities in the astrophysics literature (DEAL). In *Proceedings of the first Workshop on Information Extraction from Scientific Publications*, pages 1–7, Online. Association for Computational Linguistics.

Runwei Guan, Ka Lok Man, Feifan Chen, Shanliang Yao, Rongsheng Hu, Xiaohui Zhu, Jeremy Smith, Eng Gee Lim, and Yutao Yue. 2024. Findvehicle and vehiclefinder: a ner dataset for natural language-based vehicle retrieval and a keyword-based cross-modal vehicle retrieval system. *Multimedia Tools and Applications*, 83(8):24841–24874.

Honghao Gui, Shuofei Qiao, Jintian Zhang, Hongbin Ye, Mengshu Sun, Lei Liang, Huajun Chen, and Ningyu Zhang. 2023. Instructie: A bilingual instruction-based information extraction dataset. *CoRR*, abs/2305.11527.

Honghao Gui, Lin Yuan, Hongbin Ye, Ningyu Zhang, Mengshu Sun, Lei Liang, and Huajun Chen. 2024. IEPile: Unearthing large scale schema-conditioned information extraction corpus. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 127–146, Bangkok, Thailand. Association for Computational Linguistics.

Harsha Gurulingappa, Abdul Mateen Rajput, Angus Roberts, Juliane Fluck, Martin Hofmann-Apitius, and Luca Toldo. 2012. Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports. *Journal of Biomedical Informatics*, 45(5):885 – 892. Text Mining and Natural Language Processing in Pharmacogenomics.

Simon Hachmeier and Robert Jäschke. 2024. Information extraction of music entities in conversational music queries. In *Proceedings of the 3rd Workshop on NLP for Music and Audio (NLP4MusA)*, pages 37–42, Oakland, USA. Association for Computational Lingustics.

Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid 'O S'eaghdha, Sebastian

Pad'o, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. 2010. SemEval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 33–38, Uppsala, Sweden. Association for Computational Linguistics.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Devin Hoesen and Ayu Purwarianti. 2018. Investigating bi-lstm and crf with pos tag embedding for indonesian named entity tagger. *2018 International Conference on Asian Language Processing (IALP)*, pages 35–38.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. 2022. Training compute-optimal large language models. *Preprint*, arXiv:2203.15556.

Jiwoo Hong, Noah Lee, and James Thorne. 2024. ORPO: Monolithic preference optimization without reference model. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11170–11189, Miami, Florida, USA. Association for Computational Linguistics.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Helena Hubkov'a, Pavel Kral, and Eva Pettersson. 2020. Czech historical named entity corpus v 1.0. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4458–4465, Marseille, France. European Language Resources Association.

Binyuan Hui, Jian Yang, Zeyu Cui, Jiaxi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang, Bowen Yu, Keming Lu, Kai Dang, Yang Fan, Yichang Zhang, An Yang, Rui Men, Fei Huang, Bo Zheng, Yibo Miao, Shanghaoran Quan, Yunlong Feng, Xingzhang Ren, Xuancheng Ren, Jingren Zhou, and Junyang Lin. 2024. Qwen2.5-coder technical report. *Preprint*, arXiv:2409.12186.

Rasmus Hvingelby, Amalie Brogaard Pauli, Maria Barrett, Christina Rosted, Lasse Malm Lidegaard, and Anders Søgaard. 2020. DaNE: A named entity resource for Danish. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4597–4604, Marseille, France. European Language Resources Association.

Stockmark Inc. 2021. ner-wikipedia-dataset. https://github.com/stockmarkteam/ner-wikipedia-dataset.

Svanhv'ıt Lilja Ing'olfsd'ottir, 'Asmundur Alma Guðj'onsson, and Hrafn Loftsson. 2020. MIM-GOLD-NER 2.0 – named entity recognition corpus (22.06) (2022-06-10). CLARIN-IS.

Vitaly Ivanin, Ekaterina Artemova, Tatiana Batura, Vladimir Ivanov, Veronika Sarkisyan, Elena Tutubalina, and Ivan Smurov. 2020. Rurebus-2020 shared task: Russian relation extraction for business. In *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog" [Komp'iuternaia Lingvistika i Intellektual'nye Tehnologii: Trudy Mezhdunarodnoj Konferentsii "Dialog"]*, Moscow, Russia.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts. *Preprint*, arXiv:2401.04088.

Yizhu Jiao, Ming Zhong, Sha Li, Ruining Zhao, Siru Ouyang, Heng Ji, and Jiawei Han. 2023. Instruct and extract: Instruction tuning for on-demand information extraction. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10030–10051, Singapore. Association for Computational Linguistics.

Ebrahim Chekol Jibril and A. Cüneyd Tantuğ. 2023. Anec: An amharic named entity corpus and transformer based recognizer. *IEEE Access*, 11:15799–15815.

Ragger Jonkers. 2016. Named entity recognition on dutch parliamentary documents using frog. Bachelor's thesis, University of Amsterdam.

Fredrik Jørgensen, Tobias Aasmoe, Anne-Stine Ruud Husevag, Lilja Øvrelid, and Erik Velldal. 2020. NorNE: Annotating named entities for Norwegian. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4547–4556, Marseille, France. European Language Resources Association.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *Preprint*, arXiv:2001.08361.

Siti Oryza Khairunnisa, Aizhan Imankulova, and Mamoru Komachi. 2020. Towards a standardized dataset on indonesian named entity recognition. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing: Student Research Workshop*.

Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan, Saiful Haq, Ashutosh Sharma, Thomas T. Joshi, Hanna Moazam, Heather Miller, Matei Zaharia, and Christopher Potts. 2024. Dspy: Compiling declarative language model calls into self-improving pipelines.

Lingxing Kong, Yougang Chu, Zheng Ma, Jianbing Zhang, Liang He, and Jiajun Chen. 2024. MixRED: A mix-lingual relation extraction dataset. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11361–11370, Torino, Italia. ELRA and ICCL.

Dawid Jan Kopiczko, Tijmen Blankevoort, and Yuki M Asano. 2024. VeRA: Vector-based random matrix adaptation. In *The Twelfth International Conference on Learning Representations*.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. *Preprint*, arXiv:2309.06180.

Duong Minh Le, Yang Chen, Alan Ritter, and Wei Xu. 2024a. Constrained decoding for cross-lingual label projection. In *The Twelfth International Conference on Learning Representations*.

Khoi M. Le, Trinh Pham, Tho Quan, and Anh Tuan Luu. 2024b. Lampat: Low-rank adaption for multilingual paraphrasing using adversarial training. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16):18435–18443.

Thang Le. 2024. Cross-lingual summarization with pseudo-label regularization. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 4644–4677, Mexico City, Mexico. Association for Computational Linguistics.

Thang Le and Anh Tuan Luu. 2023. A parallel corpus for Vietnamese central-northern dialect text transfer. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13839–13855, Singapore. Association for Computational Linguistics.

Elena Leitner, Georg Rehm, and Julián Moreno-Schneider. 2020. A dataset of german legal documents for named entity recognition. *arXiv preprint*.

Jakob Lenardi c, Jaka Cibej, Spela Arhar Holdt, Toma z Erjavec, Darja Fi ser, Nikola Ljube si'c, Katja Zupan, and Kaja Dobrovoljc. 2022. CMC training corpus janes-tag 3.0. Slovenian language resource repository CLARIN.SI.

Haoran Li, Abhinav Arora, Shuohui Chen, Anchit Gupta, Sonal Gupta, and Yashar Mehdad. 2021a. MTOP: A comprehensive multilingual task-oriented semantic parsing benchmark. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2950–2962, Online. Association for Computational Linguistics.

Jiao Li, Yueping Sun, Robin J. Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J. Mattingly, Thomas C. Wiegers, and Zhiyong Lu. 2016. Biocreative V CDR task corpus: a resource for chemical disease relation extraction. *Database J. Biol. Databases Curation*, 2016.

Sha Li, Heng Ji, and Jiawei Han. 2021b. Document-level event argument extraction by conditional generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 894–908, Online. Association for Computational Linguistics.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*.

Onkar Litake, Maithili Ravindra Sabane, Parth Sachin Patil, Aparna Abhijeet Ranade, and Raviraj Joshi. 2022. L3Cube-MahaNER: A Marathi named entity recognition dataset and BERT models. In *Proceedings of the WILDRE-6 Workshop within the 13th Language Resources and Evaluation Conference*, pages 29–34, Marseille, France. European Language Resources Association.

Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. 2024a. Dora: Weight-decomposed low-rank adaptation. *Preprint*, arXiv:2402.09353.

Weiyang Liu, Zeju Qiu, Yao Feng, Yuliang Xiu, Yuxuan Xue, Longhui Yu, Haiwen Feng, Zhen Liu, Juyeon Heo, Songyou Peng, Yandong Wen, Michael J. Black, Adrian Weller, and Bernhard Schölkopf. 2024b. Parameter-efficient orthogonal finetuning via butterfly factorization. In *The Twelfth International Conference on Learning Representations*.

Zihan Liu, Yan Xu, Tiezheng Yu, Wenliang Dai, Ziwei Ji, Samuel Cahyawijaya, Andrea Madotto, and Pascale Fung. 2020. Crossner: Evaluating cross-domain named entity recognition. In *AAAI Conference on Artificial Intelligence*.

Nikola Ljube si'c, Toma z Erjavec, Vuk Batanovi'c, Maja Mili cevi'c, and Tanja Samard zi'c. 2023. Croatian twitter training corpus ReLDI-NormTagNER-hr 3.0. Slovenian language resource repository CLARIN.SI.

Nikola Ljube si'c and Tanja Samard zi'c. 2023. Croatian linguistic training corpus hr500k 2.0. Slovenian language resource repository CLARIN.SI.

Nikola Ljubesi'c, Toma z Erjavec, Vuk Batanovi'c, Maja Milicevi'c, and Tanja Samardzi'c. 2023. Serbian twitter training corpus ReLDI-NormTagNER-sr 3.0. Slovenian language resource repository CLARIN.SI.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Natalia Loukachevitch, Ekaterina Artemova, Tatiana Batura, Pavel Braslavski, Ilia Denisov, Vladimir Ivanov, Suresh Manandhar, Alexander Pugachev, and Elena Tutubalina. 2021. NEREL: A Russian dataset with nested named entities, relations and events. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 876–885, Held Online. INCOMA Ltd.

Yaojie Lu, Qing Liu, Dai Dai, Xinyan Xiao, Hongyu Lin, Xianpei Han, Le Sun, and Hua Wu. 2022. Unified structure generation for universal information extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5755–5772, Dublin, Ireland. Association for Computational Linguistics.

Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. 2018. Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3219–3232, Brussels, Belgium. Association for Computational Linguistics.

Hanjun Luo, Yibing Jin, Xuecheng Liu, Tong Shang, Ruizhe Chen, and Zuozhu Liu. 2024. Geic: Universal and multilingual named entity recognition with large language models. *ArXiv*, abs/2409.11022.

Ling Luo, Po-Ting Lai, Chih-Hsuan Wei, Cecilia N Arighi, and Zhiyong Lu. 2022. Biored: A rich biomedical relation extraction dataset. *Briefing in Bioinformatics*.

Jouni Luoma, Miika Oinonen, Maria Pyyk"onen, Veronika Laippala, and Sampo Pyysalo. 2020. A broad-coverage corpus for Finnish named entity recognition. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4615–4624, Marseille, France. European Language Resources Association.

Pedro Henrique Luz de Araujo, Teofilo de Campos, Renato Oliveira, Matheus Stauffer, Samuel Couto, and Paulo De Souza Bermejo. 2018. *LeNER-Br: A Dataset for Named Entity Recognition in Brazilian Legal Text: 13th International Conference, PROPOR 2018, Canela, Brazil, September 24–26, 2018, Proceedings*, pages 313–323.

Youmi Ma, An Wang, and Naoaki Okazaki. 2024. Building a Japanese document-level relation extraction dataset assisted by cross-lingual transfer. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2567–2579, Torino, Italia. ELRA and ICCL.

Bernardo Magnini, Begoña Altuna, Alberto Lavelli, Manuela Speranza, and Roberto Zanoli. 2020. The e3c project: Collection and annotation of a multilingual corpus of clinical cases. *Proceedings of the Seventh Italian Conference on Computational Linguistics CLiC-it 2020*.

Stephen Mayhew, Terra Blevins, Shuheng Liu, Marek Suppa, Hila Gonen, Joseph Marvin Imperial, Börje Karlsson, Peiqin Lin, Nikola Ljubešić, Lester James Miranda, Barbara Plank, Arij Riabi, and Yuval Pinter. 2024. Universal NER: A gold-standard multilingual named entity recognition benchmark. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4322–4337, Mexico City, Mexico. Association for Computational Linguistics.

Rendani Mbuvha, David Ifeoluwa Adelani, Tendani Mutavhatsindi, Tshimangadzo Rakhuhu, Aluwani Mauda, Tshifhiwa Joshua Maumela, Andisani Masindi, Seani Rananga, Vukosi Marivate, and Tshilidzi Marwala. 2023. MphayaNER: Named entity recognition for tshivenda. In *4th Workshop on African Natural Language Processing*.

Yu Meng, Mengzhou Xia, and Danqi Chen. 2024. Simpo: Simple preference optimization with a reference-free reward. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Arnav Mhaske, Harshit Kedia, Sumanth Doddapaneni, Mitesh M. Khapra, Pratyush Kumar, Rudra Murthy, and Anoop Kunchukuttan. 2023. Naamapadam: A large-scale named entity annotated data for Indic languages. In *Proceedings of the 61st Annual Meeting*

*of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10441–10456, Toronto, Canada. Association for Computational Linguistics.

Alice Millour, Yoann Dupont, Karen Fort, and Liam Duignan. 2024. Unveiling strengths and weaknesses of NLP systems based on a rich evaluation corpus: The case of NER in French. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17217–17224, Torino, Italia. ELRA and ICCL.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Lester James Miranda. 2023. Developing a named entity recognition dataset for Tagalog. In *Proceedings of the First Workshop in South East Asian Language Processing*, pages 13–20, Nusa Dua, Bali, Indonesia. Association for Computational Linguistics.

A Miranda-Escalada, E Farr'e, and M Krallinger. 2020. Named entity recognition, concept normalization and clinical coding: Overview of the cantemist track for cancer text mining in spanish, corpus, guidelines, methods and results. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020), CEUR Workshop Proceedings*.

Maria Mitrofan and Verginica Barbu Mititelu. 2020. The romanian medical treebank - simonero. In *Proceedings of the The 15th Edition of the International Conference on Linguistic Resources and Tools for Natural Language Processing*.

Behrang Mohit, Nathan Schneider, Rishav Bhowmick, Kemal Oflazer, and Noah A. Smith. 2012. Recall-oriented learning of named entities in Arabic Wikipedia. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 162–173, Avignon, France. Association for Computational Linguistics.

Hanane Nour Moussa and Asmaa Mourhir. 2023. Darnercorp: An annotated named entity recognition dataset in the moroccan dialect. *Data in Brief*, 48:109234.

Mr.Wannaphong. Thai ner 2.2. `https://zenodo.org/records/10795907`.

Rudra Murthy, Pallab Bhattacharjee, Rahul Sharnagat, Jyotsana Khatri, Diptesh Kanojia, and Pushpak Bhattacharyya. 2022. HiNER: A large Hindi named entity recognition dataset. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4467–4476, Marseille, France. European Language Resources Association.

Chien Nguyen, Huy Nguyen, Franck Dernoncourt, and Thien Nguyen. 2023a. Transitioning representations between languages for cross-lingual event detection via langevin dynamics. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14085–14093, Singapore. Association for Computational Linguistics.

Cong-Duy Nguyen, Thong Nguyen, Duc Vu, and Anh Tuan Luu. 2023b. Improving multimodal sentiment analysis: Supervised angular margin-based contrastive learning for enhanced fusion representation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14714–14724, Singapore. Association for Computational Linguistics.

Cong-Duy Nguyen, Thong Nguyen, Xiaobao Wu, and Anh Tuan Luu. 2024a. KDMCSE: Knowledge distillation multimodal sentence embeddings with adaptive angular margin contrastive learning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 733–749, Mexico City, Mexico. Association for Computational Linguistics.

Huyen T M. Nguyen, Xuan-Son Vu, and Chi Mai Luong, editors. 2020. *Proceedings of the 7th International Workshop on Vietnamese Language and Speech Processing*. Association for Computational Lingustics, Hanoi, Vietnam.

Thi-Nhung Nguyen, Bang Tien Tran, Trong-Nghia Luu, Thien Huu Nguyen, and Kiem-Hieu Nguyen. 2024b. BKEE: Pioneering event extraction in the Vietnamese language. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2421–2427, Torino, Italia. ELRA and ICCL.

Thong Nguyen, Yi Bin, Xiaobao Wu, Xinshuai Dong, Zhiyuan Hu, Khoi Le, Cong-Duy Nguyen, See-Kiong Ng, and Luu Anh Tuan. 2024c. Meta-optimized angular margin contrastive framework for video-language representation learning. *ECCV*.

Thuat Nguyen and Hieu Man. 2020. Vietnamese relation extraction with BERT-based models at VLSP 2020. In *Proceedings of the 7th International Workshop on Vietnamese Language and Speech Processing*, pages 30–34, Hanoi, Vietnam. Association for Computational Lingustics.

Nobal Niraula and Jeevan Chapagain. 2022. Named entity recognition for nepali: Data sets and algorithms. In *The International FLAIRS Conference Proceedings*, volume 35.

Mai Omura and Masayuki Asahara. 2018. UD-Japanese BCCWJ: Universal Dependencies annotation for the Balanced Corpus of Contemporary Written Japanese. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 117–125, Brussels, Belgium. Association for Computational Linguistics.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Krista Opsahl-Ong, Michael J Ryan, Josh Purtell, David Broman, Christopher Potts, Matei Zaharia, and Omar Khattab. 2024. Optimizing instructions and demonstrations for multi-stage language model programs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9340–9366, Miami, Florida, USA. Association for Computational Linguistics.

Siim Orasmaa, Kadri Muischnek, Kristjan Poska, and Anna Edela. 2022. Named entity recognition in Estonian 19th century parish court records. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5304–5313, Marseille, France. European Language Resources Association.

Teresa Paccosi and Alessio Palmero Aprosio. 2022. KIND: an Italian multi-domain dataset for named entity recognition. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 501–507, Marseille, France. European Language Resources Association.

Rui Pan, Xiang Liu, Shizhe Diao, Renjie Pi, Jipeng Zhang, Chi Han, and Tong Zhang. 2024. LISA: Layerwise importance sampling for memory-efficient large language model fine-tuning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Tanmay Parekh, I-Hung Hsu, Kuan-Hao Huang, Kai-Wei Chang, and Nanyun Peng. 2024. Contextual label projection for cross-lingual structured prediction. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5738–5757, Mexico City, Mexico. Association for Computational Linguistics.

Sungjoon Park, Jihyung Moon, Sungdong Kim, Won Ik Cho, Ji Yoon Han, Jangwon Park, Chisung Song, Junseong Kim, Youngsook Song, Taehwan Oh, Joohong Lee, Juhyun Oh, Sungwon Lyu, Younghoon Jeong, Inkwon Lee, Sangwoo Seo, Dongjun Lee, Hyunwoo Kim, Myeonghwa Lee, Seongbo Jang, Seungwon Do, Sunkyoung Kim, Kyungtae Lim, Jongwon Lee, Kyumin Park, Jamin Shin, Seonghyun Kim, Lucy Park, Alice Oh, Jung-Woo Ha, and Kyunghyun Cho. 2021. KLUE: Korean language understanding evaluation. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. *Preprint*, arXiv:1912.01703.

Nanyun Peng and Mark Dredze. 2015. Named entity recognition for Chinese social media with jointly trained embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 548–554, Lisbon, Portugal. Association for Computational Linguistics.

Siyao Peng, Zihang Sun, Huangyan Shan, Marie Kolm, Verena Blaschke, Ekaterina Artemova, and Barbara Plank. 2024. Sebastian, Basti, Wastl?! recognizing named entities in Bavarian dialectal data. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 14478–14493, Torino, Italia. ELRA and ICCL.

Vincent Perot, Kai Kang, Florian Luisier, Guolong Su, Xiaoyu Sun, Ramya Sree Boppana, Zilong Wang, Zifeng Wang, Jiaqi Mu, Hao Zhang, Chen-Yu Lee, and Nan Hua. 2024. LMDX: Language model-based document information extraction and localization. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15140–15168, Bangkok, Thailand. Association for Computational Linguistics.

Trinh Pham, Khoi Le, and Anh Tuan Luu. 2024. Unibridge: A unified approach to cross-lingual transfer learning for low-resource languages. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3168–3184, Bangkok, Thailand. Association for Computational Linguistics.

Yang Ping, JunYu Lu, Ruyi Gan, Junjie Wang, Yuxiang Zhang, Pingjian Zhang, and Jiaxing Zhang. 2023. UniEX: An effective and efficient framework for unified information extraction via a span-extractive perspective. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16424–16440, Toronto, Canada. Association for Computational Linguistics.

Jakub Piskorski, Michał Marci'nczuk, and Roman Yangarber. 2024. Cross-lingual named entity corpus for Slavic languages. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4143–4157, Torino, Italia. ELRA and ICCL.

Hanieh Poostchi, Ehsan Zare Borzeshi, Mohammad Abdous, and Massimo Piccardi. 2016. PersoNER: Persian named-entity recognition. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3381–3389, Osaka, Japan. The COLING 2016 Organizing Committee.

Hanieh Poostchi, Ehsan Zare Borzeshi, and Massimo Piccardi. 2018. BiLSTM-CRF for Persian named-entity recognition ArmanPersoNERCorpus: the first entity-annotated Persian dataset. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Amir Pouran Ben Veyseh, Javid Ebrahimi, Franck Dernoncourt, and Thien Nguyen. 2022a. MEE: A novel multilingual event extraction dataset. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9603–9613, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Amir Pouran Ben Veyseh, Minh Van Nguyen, Franck Dernoncourt, and Thien Nguyen. 2022b. MINION: a large-scale and diverse dataset for multilingual event detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2286–2299, Seattle, United States. Association for Computational Linguistics.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Hwee Tou Ng, Anders Bj"orkelund, Olga Uryupina, Yuchen Zhang, and Zhi Zhong. 2013. Towards robust linguistic analysis using OntoNotes. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 143–152, Sofia, Bulgaria. Association for Computational Linguistics.

Sampo Pyysalo and Sophia Ananiadou. 2014. Anatomical entity mention recognition at literature scale. *Bioinformatics*, 30(6):868–875.

Vasile Păiș, Maria Mitrofan, Carol Luca Gasan, Alexandru Ianov, Corvin Ghiță, Vlad Silviu Coneschi, and Andrei Onuț. 2022. Romanian named entity recognition in the legal domain (legalnero).

Yunjia Qi, Hao Peng, Xiaozhi Wang, Bin Xu, Lei Hou, and Juanzi Li. 2024. Adelie: Aligning large language models on information extraction. *Preprint*, arXiv:2405.05008.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn.

2023. Direct preference optimization: Your language model is secretly a reward model. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Md Jamiur Rahman Rifat, Sheikh Abujar, Sheak Rashed Haider Noori, and Syed Akhter Hossain. 2019. Bengali named entity recognition: A survey with deep learning benchmark. In *2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, pages 1–5.

Aniketh Janardhan Reddy, Monica Adusumilli, Sai Kiranmai Gorla, Lalita Bhanu Murthy Neti, and Aruna Malapati. 2018. Named entity recognition for telugu using lstm-crf. In *WILDRE4–4th Workshop on Indian Language Data: Resources and Evaluation*, page 6.

Yubing Ren, Yanan Cao, Hao Li, Yingjie Li, Zixuan ZM Ma, Fang Fang, Ping Guo, and Wei Ma. 2024. DEIE: Benchmarking document-level event information extraction with a large-scale Chinese news dataset. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4592–4604, Torino, Italia. ELRA and ICCL.

Ibm Research. 2023. Granite foundation models.

Dan Roth and Wen-tau Yih. 2002. Probabilistic reasoning for entity & relation recognition. In *COLING 2002: The 19th International Conference on Computational Linguistics*.

Marco Rovera. 2024. EventNet-ITA: Italian frame parsing for events. In *Proceedings of the 8th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2024)*, pages 77–90, St. Julians, Malta. Association for Computational Linguistics.

Teemu Ruokolainen, Pekka Kauppinen, Miikka Silfverberg, and Krister Lind'en. 2019. A finnish news corpus for named entity recognition. *Language Resources and Evaluation*, pages 1–26.

Sovan Kumar Sahoo, Saumajit Saha, Asif Ekbal, and Pushpak Bhattacharyya. 2020. A platform for event extraction in hindi. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2241–2250.

Taneeya Satyapanich, Francis Ferraro, and Timothy W. Finin. 2020. Casie: Extracting cybersecurity event information from text. In *AAAI Conference on Artificial Intelligence*.

Salim Sazzed. 2022. BanglaBioMed: A biomedical named-entity annotated corpus for Bangla (Bengali). In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 323–329, Dublin, Ireland. Association for Computational Linguistics.

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. 2023. Are emergent abilities of large language models a mirage? In *Thirty-seventh Conference on Neural Information Processing Systems*.

Sebastian Schuster, Sonal Gupta, Rushin Shah, and Mike Lewis. 2019. Cross-lingual transfer learning for multilingual task oriented dialog. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3795–3805, Minneapolis, Minnesota. Association for Computational Linguistics.

Alessandro Seganti, Klaudia Firlkag, Helena Skowronska, Michal Satlawa, and Piotr Andruszkiewicz. 2021. Multilingual entity and relation extraction dataset and model. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1946–1955, Online. Association for Computational Linguistics.

Emiko Shinohara, Daisaku Shibata, and Yoshimasa Kawazoe. 2022. Development of comprehensive annotation criteria for patients' states from clinical texts. *Journal of Biomedical Informatics*, 134:104200.

Mariana O. Silva and Mirella M. Moro. 2024. PPORTAL_ner: An annotated corpus of Portuguese literary entities. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12927–12937, Torino, Italia. ELRA and ICCL.

Eszter Simon and No'emi Vad'asz. 2021. Introducing nytk-nerkor, A gold standard hungarian named entity annotated corpus. In *Text, Speech, and Dialogue - 24th International Conference, TSD 2021, Olomouc, Czech Republic, September 6-9, 2021, Proceedings*, volume 12848 of *Lecture Notes in Computer Science*, pages 222–234. Springer.

Olof Mogren Simon Almgren, Sean Pavlov. 2016. Named entity recognition in swedish medical journals with deep bidirectional character-based lstms. In *Proceedings of the Fifth Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM 2016)*, page 1.

Kushagra Singh, Indira Sen, and Ponnurangam Kumaraguru. 2018a. Language identification and named entity recognition in Hinglish code mixed tweets. In *Proceedings of ACL 2018, Student Research Workshop*, pages 52–58, Melbourne, Australia. Association for Computational Linguistics.

Vinay Singh, Deepanshu Vijay, Syed Sarfaraz Akhtar, and Manish Shrivastava. 2018b. Named entity recognition for Hindi-English code-mixed social media text. In *Proceedings of the Seventh Named Entities Workshop*, pages 27–35, Melbourne, Australia. Association for Computational Linguistics.

Kairit Sirts. 2023. Estonian named entity recognition: New datasets and models. In *Proceedings of the*

*24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 752–761, T'orshavn, Faroe Islands. University of Tartu Library.

Larry L. Smith, Lorraine K. Tanabe, Rie Ando, Cheng-Ju Kuo, I-Fang Chung, Chun-Nan Hsu, Yu-Shi Lin, Roman Klinger, C. Friedrich, Kuzman Ganchev, Manabu Torii, Hongfang Liu, Barry Haddow, Craig A. Struble, Richard J. Povinelli, Andreas Vlachos, William A. Baumgartner, Lawrence E. Hunter, Bob Carpenter, Richard Tzong-Han Tsai, Hong-Jie Dai, Feng Liu, Yifei Chen, Chengjie Sun, Sophia Katrenko, Pieter W. Adriaans, Christian Blaschke, Rafael Torres, Mariana L. Neves, Preslav Nakov, Anna Divoli, Manuel Maña-L'opez, Jacinto Mata, and W. John Wilbur. 2008. Overview of biocreative ii gene mention recognition. *Genome Biology*, 9:S2 – S2.

David Suba, Marek Suppa, Jozef Kubik, Endre Hamerlik, and Martin Takac. 2023. WikiGoldSK: Annotated dataset, baselines and few-shot learning experiments for Slovak named entity recognition. In *Proceedings of the 9th Workshop on Slavic Natural Language Processing 2023 (SlavicNLP 2023)*, pages 138–145, Dubrovnik, Croatia. Association for Computational Linguistics.

Zhaoyue Sun, Jiazheng Li, Gabriele Pergola, Byron Wallace, Bino John, Nigel Greene, Joseph Kim, and Yulan He. 2022. PHEE: A dataset for pharmacovigilance event extraction from text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5571–5587, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Zhaoyue Sun, Gabriele Pergola, Byron Wallace, and Yulan He. 2024. Leveraging ChatGPT in pharmacovigilance event extraction: An empirical study. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 344–357, St. Julian's, Malta. Association for Computational Linguistics.

Qingyu Tan, Lu Xu, Lidong Bing, Hwee Tou Ng, and Sharifah Mahani Aljunied. 2022. Revisiting DocRED - addressing the false negative problem in relation extraction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8472–8487, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Xuemei Tang, Qi Su, Jun Wang, and Zekun Deng. 2024. CHisIEC: An information extraction corpus for Ancient Chinese history. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3192–3202, Torino, Italia. ELRA and ICCL.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, Soroosh Mariooryad, Yifan Ding, Xinyang Geng, Fred Alcober, Roy Frostig, Mark Omernick, Lexi Walker, Cosmin Paduraru, Christina Sorokin, Andrea Tacchetti, Colin Gaffney, Samira Daruki, Olcan Sercinoglu, Zach Gleicher, Juliette Love, Paul Voigtlaender, Rohan Jain, Gabriela Surita, Kareem Mohamed, Rory Blevins, Junwhan Ahn, Tao Zhu, Kornraphop Kawintiranon, Orhan Firat, Yiming Gu, Yujing Zhang, Matthew Rahtz, Manaal Faruqui, Natalie Clay, Justin Gilmer, JD Co-Reyes, Ivo Penchev, Rui Zhu, Nobuyuki Morioka, Kevin Hui, Krishna Haridasan, Victor Campos, Mahdis Mahdieh, Mandy Guo, Samer Hassan, Kevin Kilgour, Arpi Vezer, Heng-Tze Cheng, Raoul de Liedekerke, Siddharth Goyal, Paul Barham, DJ Strouse, Seb Noury, Jonas Adler, Mukund Sundararajan, Sharad Vikram, Dmitry Lepikhin, Michela Paganini, Xavier Garcia, Fan Yang, Dasha Valter, Maja Trebacz, Kiran Vodrahalli, Chulayuth Asawaroengchai, Roman Ring, Norbert Kalb, Livio Baldini Soares, Siddhartha Brahma, David Steiner, Tianhe Yu, Fabian Mentzer, Antoine He, Lucas Gonzalez, Bibo Xu, Raphael Lopez Kaufman, Laurent El Shafey, Junhyuk Oh, Tom Hennigan, George van den Driessche, Seth Odoom, Mario Lucic, Becca Roelofs, Sid Lall, Amit Marathe, Betty Chan, Santiago Ontanon, Luheng He, Denis Teplyashin, Jonathan Lai, Phil Crone, Bogdan Damoc, Lewis Ho, Sebastian Riedel, Karel Lenc, Chih-Kuan Yeh, Aakanksha Chowdhery, Yang Xu, Mehran Kazemi, Ehsan Amid, Anastasia Petrushkina, Kevin Swersky, Ali Khodaei, Gowoon Chen, Chris Larkin, Mario Pinto, Geng Yan, Adria Puigdomenech Badia, Piyush Patil, Steven Hansen, Dave Orr, Sebastien M. R. Arnold, Jordan Grimstad, Andrew Dai, Sholto Douglas, Rishika Sinha, Vikas Yadav, Xi Chen, Elena Gribovskaya, Jacob Austin, Jeffrey Zhao, Kaushal Patel, Paul Komarek, Sophia Austin, Sebastian Borgeaud, Linda Friso, Abhimanyu Goyal, Ben Caine, Kris Cao, Da-Woon Chung, Matthew Lamm, Gabe Barth-Maron, Thais Kagohara, Kate Olszewska, Mia Chen, Kaushik Shivakumar, Rishabh Agarwal, Harshal Godhia, Ravi Rajwar, Javier Snaider, Xerxes Dotiwalla, Yuan Liu, Aditya Barua, Victor Ungureanu, Yuan Zhang, Bat-Orgil Batsaikhan, Mateo Wirth, James Qin, Ivo Danihelka, Tulsee Doshi, Martin Chadwick, Jilin Chen, Sanil Jain, Quoc Le, Arjun Kar, Madhu Gurumurthy, Cheng Li, Ruoxin Sang, Fangyu Liu, Lampros Lamprou, Rich Munoz, Nathan Lintz, Harsh Mehta, Heidi Howard, Malcolm Reynolds, Lora Aroyo, Quan Wang, Lorenzo Blanco, Albin Cassirer, Jordan Griffith, Dipanjan Das, Stephan Lee, Jakub Sygnowski, Zach Fisher, James Besley, Richard Powell, Zafarali Ahmed, Dominik Paulus, David Reitter, Zalan Borsos, Rishabh Joshi, Aedan Pope, Steven Hand, Vittorio Selo, Vihan Jain, Nikhil Sethi, Megha Goel, Takaki Makino, Rhys May, Zhen Yang, Johan Schalkwyk, Christina Butterfield, Anja Hauth, Alex Goldin, Will Hawkins, Evan Senter, Sergey Brin, Oliver Woodman, Marvin Ritter, Eric Noland, Minh Giang, Vijay Bolina, Lisa Lee, Tim Blyth, Ian Mackinnon, Machel Reid, Obaid Sarvana, David Silver, Alexander Chen, Lily Wang, Loren Maggiore, Oscar Chang, Nithya At-

taluri, Gregory Thornton, Chung-Cheng Chiu, Oskar Bunyan, Nir Levine, Timothy Chung, Evgenii Eltyshev, Xiance Si, Timothy Lillicrap, Demetra Brady, Vaibhav Aggarwal, Boxi Wu, Yuanzhong Xu, Ross McIlroy, Kartikeya Badola, Paramjit Sandhu, Erica Moreira, Wojciech Stokowiec, Ross Hemsley, Dong Li, Alex Tudor, Pranav Shyam, Elahe Rahimtoroghi, Salem Haykal, Pablo Sprechmann, Xiang Zhou, Diana Mincu, Yujia Li, Ravi Addanki, Kalpesh Krishna, Xiao Wu, Alexandre Frechette, Matan Eyal, Allan Dafoe, Dave Lacey, Jay Whang, Thi Avrahami, Ye Zhang, Emanuel Taropa, Hanzhao Lin, Daniel Toyama, Eliza Rutherford, Motoki Sano, HyunJeong Choe, Alex Tomala, Chalence Safranek-Shrader, Nora Kassner, Mantas Pajarskas, Matt Harvey, Sean Sechrist, Meire Fortunato, Christina Lyu, Gamaleldin Elsayed, Chenkai Kuang, James Lottes, Eric Chu, Chao Jia, Chih-Wei Chen, Peter Humphreys, Kate Baumli, Connie Tao, Rajkumar Samuel, Cicero Nogueira dos Santos, Anders Andreassen, Nemanja Rakićević, Dominik Grewe, Aviral Kumar, Stephanie Winkler, Jonathan Caton, Andrew Brock, Sid Dalmia, Hannah Sheahan, Iain Barr, Yingjie Miao, Paul Natsev, Jacob Devlin, Feryal Behbahani, Flavien Prost, Yanhua Sun, Artiom Myaskovsky, Thanumalayan Sankaranarayana Pillai, Dan Hurt, Angeliki Lazaridou, Xi Xiong, Ce Zheng, Fabio Pardo, Xiaowei Li, Dan Horgan, Joe Stanton, Moran Ambar, Fei Xia, Alejandro Lince, Mingqiu Wang, Basil Mustafa, Albert Webson, Hyo Lee, Rohan Anil, Martin Wicke, Timothy Dozat, Abhishek Sinha, Enrique Piqueras, Elahe Dabir, Shyam Upadhyay, Anudhyan Boral, Lisa Anne Hendricks, Corey Fry, Josip Djolonga, Yi Su, Jake Walker, Jane Labanowski, Ronny Huang, Vedant Misra, Jeremy Chen, RJ Skerry-Ryan, Avi Singh, Shruti Rijhwani, Dian Yu, Alex Castro-Ros, Beer Changpinyo, Romina Datta, Sumit Bagri, Arnar Mar Hrafnkelsson, Marcello Maggioni, Daniel Zheng, Yury Sulsky, Shaobo Hou, Tom Le Paine, Antoine Yang, Jason Riesa, Dominika Rogozinska, Dror Marcus, Dalia El Badawy, Qiao Zhang, Luyu Wang, Helen Miller, Jeremy Greer, Lars Lowe Sjos, Azade Nova, Heiga Zen, Rahma Chaabouni, Mihaela Rosca, Jiepu Jiang, Charlie Chen, Ruibo Liu, Tara Sainath, Maxim Krikun, Alex Polozov, Jean-Baptiste Lespiau, Josh Newlan, Zeyncep Cankara, Soo Kwak, Yunhan Xu, Phil Chen, Andy Coenen, Clemens Meyer, Katerina Tsihlas, Ada Ma, Juraj Gottweis, Jinwei Xing, Chenjie Gu, Jin Miao, Christian Frank, Zeynep Cankara, Sanjay Ganapathy, Ishita Dasgupta, Steph Hughes-Fitt, Heng Chen, David Reid, Keran Rong, Hongmin Fan, Joost van Amersfoort, Vincent Zhuang, Aaron Cohen, Shixiang Shane Gu, Anhad Mohananey, Anastasija Ilic, Taylor Tobin, John Wieting, Anna Bortsova, Phoebe Thacker, Emma Wang, Emily Caveness, Justin Chiu, Eren Sezener, Alex Kaskasoli, Steven Baker, Katie Millican, Mohamed Elhawaty, Kostas Aisopos, Carl Lebsack, Nathan Byrd, Hanjun Dai, Wenhao Jia, Matthew Wiethoff, Elnaz Davoodi, Albert Weston, Lakshman Yagati, Arun Ahuja, Isabel Gao, Golan Pundak, Susan Zhang, Michael Azzam, Khe Chai Sim, Sergi Caelles, James Keeling, Abhanshu Sharma, Andy Swing, YaGuang Li, Chenxi Liu, Carrie Grimes Bostock, Yamini Bansal, Zachary Nado, Ankesh Anand, Josh Lipschultz, Abhijit Karmarkar, Lev Proleev, Abe Ittycheriah, Soheil Hassas Yeganeh, George Polovets, Aleksandra Faust, Jiao Sun, Alban Rrustemi, Pen Li, Rakesh Shivanna, Jeremiah Liu, Chris Welty, Federico Lebron, Anirudh Baddepudi, Sebastian Krause, Emilio Parisotto, Radu Soricut, Zheng Xu, Dawn Bloxwich, Melvin Johnson, Behnam Neyshabur, Justin Mao-Jones, Renshen Wang, Vinay Ramasesh, Zaheer Abbas, Arthur Guez, Constant Segal, Duc Dung Nguyen, James Svensson, Le Hou, Sarah York, Kieran Milan, Sophie Bridgers, Wiktor Gworek, Marco Tagliasacchi, James Lee-Thorp, Michael Chang, Alexey Guseynov, Ale Jakse Hartman, Michael Kwong, Ruizhe Zhao, Sheleem Kashem, Elizabeth Cole, Antoine Miech, Richard Tanburn, Mary Phuong, Filip Pavetic, Sebastien Cevey, Ramona Comanescu, Richard Ives, Sherry Yang, Cosmo Du, Bo Li, Zizhao Zhang, Mariko Iinuma, Clara Huiyi Hu, Aurko Roy, Shaan Bijwadia, Zhenkai Zhu, Danilo Martins, Rachel Saputro, Anita Gergely, Steven Zheng, Dawei Jia, Ioannis Antonoglou, Adam Sadovsky, Shane Gu, Yingying Bi, Alek Andreev, Sina Samangooei, Mina Khan, Tomas Kocisky, Angelos Filos, Chintu Kumar, Colton Bishop, Adams Yu, Sarah Hodkinson, Sid Mittal, Premal Shah, Alexandre Moufarek, Yong Cheng, Adam Bloniarz, Jaehoon Lee, Pedram Pejman, Paul Michel, Stephen Spencer, Vladimir Feinberg, Xuehan Xiong, Nikolay Savinov, Charlotte Smith, Siamak Shakeri, Dustin Tran, Mary Chesus, Bernd Bohnet, George Tucker, Tamara von Glehn, Carrie Muir, Yiran Mao, Hideto Kazawa, Ambrose Slone, Kedar Soparkar, Disha Shrivastava, James Cobon-Kerr, Michael Sharman, Jay Pavagadhi, Carlos Araya, Karolis Misiunas, Nimesh Ghelani, Michael Laskin, David Barker, Qiujia Li, Anton Briukhov, Neil Houlsby, Mia Glaese, Balaji Lakshminarayanan, Nathan Schucher, Yunhao Tang, Eli Collins, Hyeontaek Lim, Fangxiaoyu Feng, Adria Recasens, Guangda Lai, Alberto Magni, Nicola De Cao, Aditya Siddhant, Zoe Ashwood, Jordi Orbay, Mostafa Dehghani, Jenny Brennan, Yifan He, Kelvin Xu, Yang Gao, Carl Saroufim, James Molloy, Xinyi Wu, Seb Arnold, Solomon Chang, Julian Schrittwieser, Elena Buchatskaya, Soroush Radpour, Martin Polacek, Skye Giordano, Ankur Bapna, Simon Tokumine, Vincent Hellendoorn, Thibault Sottiaux, Sarah Cogan, Aliaksei Severyn, Mohammad Saleh, Shantanu Thakoor, Laurent Shefey, Siyuan Qiao, Meenu Gaba, Shuo yiin Chang, Craig Swanson, Biao Zhang, Benjamin Lee, Paul Kishan Rubenstein, Gan Song, Tom Kwiatkowski, Anna Koop, Ajay Kannan, David Kao, Parker Schuh, Axel Stjerngren, Golnaz Ghiasi, Gena Gibson, Luke Vilnis, Ye Yuan, Felipe Tiengo Ferreira, Aishwarya Kamath, Ted Klimenko, Ken Franko, Kefan Xiao, Indro Bhattacharya, Miteyan Patel, Rui Wang, Alex Morris, Robin Strudel, Vivek Sharma, Peter Choy, Sayed Hadi Hashemi, Jessica Landon, Mara Finkelstein, Priya Jhakra, Justin Frye, Megan Barnes, Matthew Mauger, Dennis Daun, Khuslen Baatarsukh, Matthew Tung, Wael Farhan, Henryk Michalewski, Fabio Viola, Felix de Chaumont Quitry, Charline Le Lan, Tom Hud-

son, Qingze Wang, Felix Fischer, Ivy Zheng, Elspeth White, Anca Dragan, Jean baptiste Alayrac, Eric Ni, Alexander Pritzel, Adam Iwanicki, Michael Isard, Anna Bulanova, Lukas Zilka, Ethan Dyer, Devendra Sachan, Srivatsan Srinivasan, Hannah Muckenhirn, Honglong Cai, Amol Mandhane, Mukarram Tariq, Jack W. Rae, Gary Wang, Kareem Ayoub, Nicholas FitzGerald, Yao Zhao, Woohyun Han, Chris Alberti, Dan Garrette, Kashyap Krishnakumar, Mai Gimenez, Anselm Levskaya, Daniel Sohn, Josip Matak, Inaki Iturrate, Michael B. Chang, Jackie Xiang, Yuan Cao, Nishant Ranka, Geoff Brown, Adrian Hutter, Vahab Mirrokni, Nanxin Chen, Kaisheng Yao, Zoltan Egyed, Francois Galilee, Tyler Liechty, Praveen Kallakuri, Evan Palmer, Sanjay Ghemawat, Jasmine Liu, David Tao, Chloe Thornton, Tim Green, Mimi Jasarevic, Sharon Lin, Victor Cotruta, Yi-Xuan Tan, Noah Fiedel, Hongkun Yu, Ed Chi, Alexander Neitz, Jens Heitkaemper, Anu Sinha, Denny Zhou, Yi Sun, Charbel Kaed, Brice Hulse, Swaroop Mishra, Maria Georgaki, Sneha Kudugunta, Clement Farabet, Izhak Shafran, Daniel Vlasic, Anton Tsitsulin, Rajagopal Ananthanarayanan, Alen Carin, Guolong Su, Pei Sun, Shashank V, Gabriel Carvajal, Josef Broder, Iulia Comsa, Alena Repina, William Wong, Warren Weilun Chen, Peter Hawkins, Egor Filonov, Lucia Loher, Christoph Hirnschall, Weiyi Wang, Jingchen Ye, Andrea Burns, Hardie Cate, Diana Gage Wright, Federico Piccinini, Lei Zhang, Chu-Cheng Lin, Ionel Gog, Yana Kulizhskaya, Ashwin Sreevatsa, Shuang Song, Luis C. Cobo, Anand Iyer, Chetan Tekur, Guillermo Garrido, Zhuyun Xiao, Rupert Kemp, Huaixiu Steven Zheng, Hui Li, Ananth Agarwal, Christel Ngani, Kati Goshvadi, Rebeca Santamaria-Fernandez, Wojciech Fica, Xinyun Chen, Chris Gorgolewski, Sean Sun, Roopal Garg, Xinyu Ye, S. M. Ali Eslami, Nan Hua, Jon Simon, Pratik Joshi, Yelin Kim, Ian Tenney, Sahitya Potluri, Lam Nguyen Thiet, Quan Yuan, Florian Luisier, Alexandra Chronopoulou, Salvatore Scellato, Praveen Srinivasan, Minmin Chen, Vinod Koverkathu, Valentin Dalibard, Yaming Xu, Brennan Saeta, Keith Anderson, Thibault Sellam, Nick Fernando, Fantine Huot, Junehyuk Jung, Mani Varadarajan, Michael Quinn, Amit Raul, Maigo Le, Ruslan Habalov, Jon Clark, Komal Jalan, Kalesha Bullard, Achintya Singhal, Thang Luong, Boyu Wang, Sujeevan Rajayogam, Julian Eisenschlos, Johnson Jia, Daniel Finchelstein, Alex Yakubovich, Daniel Balle, Michael Fink, Sameer Agarwal, Jing Li, Dj Dvijotham, Shalini Pal, Kai Kang, Jaclyn Konzelmann, Jennifer Beattie, Olivier Dousse, Diane Wu, Remi Crocker, Chen Elkind, Siddhartha Reddy Jonnalagadda, Jong Lee, Dan Holtmann-Rice, Krystal Kallarackal, Rosanne Liu, Denis Vnukov, Neera Vats, Luca Invernizzi, Mohsen Jafari, Huanjie Zhou, Lilly Taylor, Jennifer Prendki, Marcus Wu, Tom Eccles, Tianqi Liu, Kavya Kopparapu, Francoise Beaufays, Christof Angermueller, Andreea Marzoca, Shourya Sarcar, Hilal Dib, Jeff Stanway, Frank Perbet, Nejc Trdin, Rachel Sterneck, Andrey Khorlin, Dinghua Li, Xihui Wu, Sonam Goenka, David Madras, Sasha Goldshtein, Willi Gierke, Tong Zhou, Yaxin Liu, Yannie Liang, Anais White, Yunjie Li, Shreya Singh, Sanaz Bahargam, Mark Epstein, Sujoy Basu, Li Lao, Adnan Ozturel, Carl Crous, Alex Zhai, Han Lu, Zora Tung, Neeraj Gaur, Alanna Walton, Lucas Dixon, Ming Zhang, Amir Globerson, Grant Uy, Andrew Bolt, Olivia Wiles, Milad Nasr, Ilia Shumailov, Marco Selvi, Francesco Piccinno, Ricardo Aguilar, Sara McCarthy, Misha Khalman, Mrinal Shukla, Vlado Galic, John Carpenter, Kevin Villela, Haibin Zhang, Harry Richardson, James Martens, Matko Bosnjak, Shreyas Rammohan Belle, Jeff Seibert, Mahmoud Alnahlawi, Brian McWilliams, Sankalp Singh, Annie Louis, Wen Ding, Dan Popovici, Lenin Simicich, Laura Knight, Pulkit Mehta, Nishesh Gupta, Chongyang Shi, Saaber Fatehi, Jovana Mitrovic, Alex Grills, Joseph Pagadora, Tsendsuren Munkhdalai, Dessie Petrova, Danielle Eisenbud, Zhishuai Zhang, Damion Yates, Bhavishya Mittal, Nilesh Tripuraneni, Yannis Assael, Thomas Brovelli, Prateek Jain, Mihajlo Velimirovic, Canfer Akbulut, Jiaqi Mu, Wolfgang Macherey, Ravin Kumar, Jun Xu, Haroon Qureshi, Gheorghe Comanici, Jeremy Wiesner, Zhitao Gong, Anton Ruddock, Matthias Bauer, Nick Felt, Anirudh GP, Anurag Arnab, Dustin Zelle, Jonas Rothfuss, Bill Rosgen, Ashish Shenoy, Bryan Seybold, Xinjian Li, Jayaram Mudigonda, Goker Erdogan, Jiawei Xia, Jiri Simsa, Andrea Michi, Yi Yao, Christopher Yew, Steven Kan, Isaac Caswell, Carey Radebaugh, Andre Elisseeff, Pedro Valenzuela, Kay McKinney, Kim Paterson, Albert Cui, Eri Latorre-Chimoto, Solomon Kim, William Zeng, Ken Durden, Priya Ponnapalli, Tiberiu Sosea, Christopher A. Choquette-Choo, James Manyika, Brona Robenek, Harsha Vashisht, Sebastien Pereira, Hoi Lam, Marko Velic, Denese Owusu-Afriyie, Katherine Lee, Tolga Bolukbasi, Alicia Parrish, Shawn Lu, Jane Park, Balaji Venkatraman, Alice Talbert, Lambert Rosique, Yuchung Cheng, Andrei Sozanschi, Adam Paszke, Praveen Kumar, Jessica Austin, Lu Li, Khalid Salama, Bartek Perz, Wooyeol Kim, Nandita Dukkipati, Anthony Baryshnikov, Christos Kaplanis, XiangHai Sheng, Yuri Chervonyi, Caglar Unlu, Diego de Las Casas, Harry Askham, Kathryn Tunyasuvunakool, Felix Gimeno, Siim Poder, Chester Kwak, Matt Miecnikowski, Vahab Mirrokni, Alek Dimitriev, Aaron Parisi, Dangyi Liu, Tomy Tsai, Toby Shevlane, Christina Kouridi, Drew Garmon, Adrian Goedeckemeyer, Adam R. Brown, Anitha Vijayakumar, Ali Elqursh, Sadegh Jazayeri, Jin Huang, Sara Mc Carthy, Jay Hoover, Lucy Kim, Sandeep Kumar, Wei Chen, Courtney Biles, Garrett Bingham, Evan Rosen, Lisa Wang, Qijun Tan, David Engel, Francesco Pongetti, Dario de Cesare, Dongseong Hwang, Lily Yu, Jennifer Pullman, Srini Narayanan, Kyle Levin, Siddharth Gopal, Megan Li, Asaf Aharoni, Trieu Trinh, Jessica Lo, Norman Casagrande, Roopali Vij, Loic Matthey, Bramandia Ramadhana, Austin Matthews, CJ Carey, Matthew Johnson, Kremena Goranova, Rohin Shah, Shereen Ashraf, Kingshuk Dasgupta, Rasmus Larsen, Yicheng Wang, Manish Reddy Vuyyuru, Chong Jiang, Joana Ijazi, Kazuki Osawa, Celine Smith, Ramya Sree Boppana, Taylan Bilal, Yuma Koizumi, Ying Xu, Yasemin Altun, Nir Shabat, Ben Bariach, Alex Korchemniy, Kiam

Choo, Olaf Ronneberger, Chimezie Iwuanyanwu, Shubin Zhao, David Soergel, Cho-Jui Hsieh, Irene Cai, Shariq Iqbal, Martin Sundermeyer, Zhe Chen, Elie Bursztein, Chaitanya Malaviya, Fadi Biadsy, Prakash Shroff, Inderjit Dhillon, Tejasi Latkar, Chris Dyer, Hannah Forbes, Massimo Nicosia, Vitaly Nikolaev, Somer Greene, Marin Georgiev, Pidong Wang, Nina Martin, Hanie Sedghi, John Zhang, Praseem Banzal, Doug Fritz, Vikram Rao, Xuezhi Wang, Jiageng Zhang, Viorica Patraucean, Dayou Du, Igor Mordatch, Ivan Jurin, Lewis Liu, Ayush Dubey, Abhi Mohan, Janek Nowakowski, Vlad-Doru Ion, Nan Wei, Reiko Tojo, Maria Abi Raad, Drew A. Hudson, Vaishakh Keshava, Shubham Agrawal, Kevin Ramirez, Zhichun Wu, Hoang Nguyen, Ji Liu, Madhavi Sewak, Bryce Petrini, DongHyun Choi, Ivan Philips, Ziyue Wang, Ioana Bica, Ankush Garg, Jarek Wilkiewicz, Priyanka Agrawal, Xiaowei Li, Danhao Guo, Emily Xue, Naseer Shaik, Andrew Leach, Sadh MNM Khan, Julia Wiesinger, Sammy Jerome, Abhishek Chakladar, Alek Wenjiao Wang, Tina Ornduff, Folake Abu, Alireza Ghaffarkhah, Marcus Wainwright, Mario Cortes, Frederick Liu, Joshua Maynez, Andreas Terzis, Pouya Samangouei, Riham Mansour, Tomasz Kępa, François-Xavier Aubet, Anton Algymr, Dan Banica, Agoston Weisz, Andras Orban, Alexandre Senges, Ewa Andrejczuk, Mark Geller, Niccolo Dal Santo, Valentin Anklin, Majd Al Merey, Martin Baeuml, Trevor Strohman, Junwen Bai, Slav Petrov, Yonghui Wu, Demis Hassabis, Koray Kavukcuoglu, Jeff Dean, and Oriol Vinyals. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *Preprint*, arXiv:2403.05530.

Qwen Team. 2024a. Qwen2.5: A party of foundation models.

The Mistral AI Team. 2024b. Un ministral, des ministraux. https://mistral.ai/en/news/ministraux?utm_source=tldrai.

Erik F. Tjong Kim Sang. 2002. Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition. In *COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Alexander Tkachenko, Timo Petmanson, and Sven Laur. 2013. Named entity recognition in Estonian. In *Proceedings of the 4th Biennial International Workshop on Balto-Slavic Natural Language Processing*, pages 78–83, Sofia, Bulgaria. Association for Computational Linguistics.

Thinh Hung Truong, Mai Hoang Dao, and Dat Quoc Nguyen. 2021. COVID-19 Named Entity Recognition for Vietnamese. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Gökhan Tür, Dilek Hakkani-Tür, and Kemal Oflazer. 2003. A statistical information extraction system for turkish. *Natural Language Engineering*, 9(2):181–210.

Ahmet Üstün, Viraat Aryabumi, Zheng Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. 2024. Aya model: An instruction finetuned open-access multilingual language model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15894–15939, Bangkok, Thailand. Association for Computational Linguistics.

Rob van der Goot, Ibrahim Sharaf, Aizhan Imankulova, Ahmet Üstün, Marija Stepanović, Alan Ramponi, Siti Oryza Khairunnisa, Mamoru Komachi, and Barbara Plank. 2021. From masked language modeling to translation: Non-English auxiliary tasks improve zero-shot spoken language understanding. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2479–2497, Online. Association for Computational Linguistics.

Xiao Wang, Weikang Zhou, Can Zu, Han Xia, Tianze Chen, Yuansen Zhang, Rui Zheng, Junjie Ye, Qi Zhang, Tao Gui, Jihua Kang, Jingsheng Yang, Siyuan Li, and Chunsai Du. 2023a. Instructuie: Multi-task instruction tuning for unified information extraction. *Preprint*, arXiv:2304.08085.

Xiao Wang, Weikang Zhou, Can Zu, Han Xia, Tianze Chen, Yuansen Zhang, Rui Zheng, Junjie Ye, Qi Zhang, Tao Gui, et al. 2023b. Instructuie: Multi-task instruction tuning for unified information extraction. *arXiv preprint arXiv:2304.08085*.

Xiaozhi Wang, Ziqi Wang, Xu Han, Wangyi Jiang, Rong Han, Zhiyuan Liu, Juanzi Li, Peng Li, Yankai Lin, and Jie Zhou. 2020. MAVEN: A massive general domain event detection dataset. In *Proceedings of EMNLP 2020*.

Shuhei Watanabe. 2023. Tree-structured parzen estimator: Understanding its algorithm components and their roles for better empirical performance. *ArXiv*, abs/2304.11127.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and

Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

Genta Winata, Alham Fikri Aji, Zheng Xin Yong, and Thamar Solorio. 2023. The decades progress on code-switching research in NLP: A systematic survey on trends and challenges. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2936–2978, Toronto, Canada. Association for Computational Linguistics.

Miriam Winkler, Virginija Juozapaityte, Rob van der Goot, and Barbara Plank. 2024a. Slot and intent detection resources for Bavarian and Lithuanian: Assessing translations vs natural queries to digital assistants. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 14898–14915, Torino, Italia. ELRA and ICCL.

Miriam Winkler, Virginija Juozapaityte, Rob van der Goot, and Barbara Plank. 2024b. Slot and intent detection resources for Bavarian and Lithuanian: Assessing translations vs natural queries to digital assistants. In *Proceedings of The 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Xiangyu Xi, Jianwei Lv, Shuaipeng Liu, Wei Ye, Fan Yang, and Guanglu Wan. 2022. MUSIED: A benchmark for event detection from multi-source heterogeneous informal texts. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2947–2964, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jingjing Xu, Ji Wen, Xu Sun, and Qi Su. 2017. A discourse-level named entity recognition and relation extraction dataset for chinese literature text. *ArXiv*, abs/1711.07010.

Liang Xu, Qianqian Dong, Cong Yu, Yin Tian, Weitang Liu, Lu Li, and Xuanwei Zhang. 2020. Cluener2020: Fine-grained name entity recognition for chinese. *arXiv preprint arXiv:2001.04351*.

Soyoung Yang, Minseok Choi, Youngwoo Cho, and Jaegul Choo. 2023. HistRED: A historical document-level relation extraction dataset. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3207–3224, Toronto, Canada. Association for Computational Linguistics.

Feng Yao, Chaojun Xiao, Xiaozhi Wang, Zhiyuan Liu, Lei Hou, Cunchao Tu, Juanzi Li, Yun Liu, Weixing Shen, and Maosong Sun. 2022. LEVEN: A large-scale Chinese legal event detection dataset. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 183–201.

Marat Yavrumyan. 2020. ArmTDP-NER: Named entity corpus of modern eastern armenian. Corpus available from https://github.com/myavrum/ArmTDP-NER.

Rustem Yeshpanov, Yerbolat Khassanov, and Huseyin Atakan Varol. 2022. KazNERD: Kazakh named entity recognition dataset. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 417–426, Marseille, France. European Language Resources Association.

Miaoran Zhang, Vagrant Gautam, Mingyang Wang, Jesujoba Alabi, Xiaoyu Shen, Dietrich Klakow, and Marius Mosbach. 2024. The impact of demonstrations on multilingual in-context learning: A multidimensional analysis. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 7342–7371, Bangkok, Thailand. Association for Computational Linguistics.

Yue Zhang and Jie Yang. 2018. Chinese NER using lattice LSTM. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1554–1564, Melbourne, Australia. Association for Computational Linguistics.

Jiawei Zhao, Zhenyu Zhang, Beidi Chen, Zhangyang Wang, Anima Anandkumar, and Yuandong Tian. 2024a. Galore: Memory-efficient llm training by gradient low-rank projection. *Preprint*, arXiv:2403.03507.

Yuze Zhao, Jintao Huang, Jinghan Hu, Xingjun Wang, Yunlin Mao, Daoze Zhang, Zeyinzi Jiang, Zhikai Wu, Baole Ai, Ang Wang, Wenmeng Zhou, and Yingda Chen. 2024b. Swift:a scalable lightweight infrastructure for fine-tuning. *Preprint*, arXiv:2408.05517.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Mengna Zhu, Zijie Xu, Kaisheng Zeng, Kaiming Xiao, Mao Wang, Wenjun Ke, and Hongbin Huang. 2024. CMNEE:a large-scale document-level event extraction dataset based on open-source Chinese military news. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3367–3379, Torino, Italia. ELRA and ICCL.

Jingming Zhuo, Songyang Zhang, Xinyu Fang, Haodong Duan, Dahua Lin, and Kai Chen. 2024. ProSA: Assessing and understanding the prompt sensitivity of LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1950–1976, Miami, Florida, USA. Association for Computational Linguistics.

Caleb Ziems, William Held, Jingfeng Yang, Jwala Dhamala, Rahul Gupta, and Diyi Yang. 2023. Multi-VALUE: A framework for cross-dialectal English NLP. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 744–768, Toronto, Canada. Association for Computational Linguistics.

Mikel Zubillaga, Oscar Sainz, Ainara Estarrona, Oier Lopez de Lacalle, and Eneko Agirre. 2024. Event extraction in Basque: Typologically motivated cross-lingual transfer-learning analysis. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6607–6621, Torino, Italia. ELRA and ICCL.

Yuxin Zuo, Wenxuan Jiang, Wenxuan Liu, Zixuan Li, Long Bai, Hanbin Wang, Yutao Zeng, Xiaolong Jin, Jiafeng Guo, and Xueqi Cheng. 2024. Alignxie: Improving multilingual information extraction by cross-lingual alignment. *Preprint*, arXiv:2411.04794.

Östling, Robert, Sjons, Johan, and Bjerva, Johannes. 2013. Sic - stockholm internet corpus.

## A Task Description

- **Named Entity Recognition** Identifying and classifying named entities from input text into predefined categories such as person names, organizations, etc.
- **Relation Extraction** Identifying relation triplets that consist of the entity pairs and their relation types among predefined categories from the input text.
- **Slot Filling** Identifying slots that correspond to predefined parameters of a user's query from the input utterance.
- **Event Detection** Identifying and classifying event triggers that most clearly express specific event occurrences from the input text.
- **Event Extraction** Identifying and classifying both event triggers and event arguments that relate to specific event occurrences from the input text.

## B Data Construction - Additional Details

**Data Collection** To search for datasets, we focus on multilingual and monolingual datasets available for each task via querying search engines[18][19], paperswithcode and ACL Anthology. As we prioritize high-quality instances for evaluations, we only selected datasets that were manually annotated or those initially tagged by heuristics but were then re-annotated by humans. In total, we collected 215 datasets each corresponding to one of the five tasks considered. Besides *standard* languages, we also included lesser-known variants such as Bavarian German[20] or Malaysian English[21] in our collections, which serves to facilitate dialectal NLP (Ziems et al., 2023; Le and Luu, 2023). Moreover, apart from single-language instances, we also collected instances with code-switching (Winata et al., 2023) to better cover possible real-world scenarios.

**Data Validation & Conversion** Although we obtained a large number of data sources, these were presented in various inconsistent formats and not ready for downstream evaluation or training. Thus, we first converted each of them to our pre-defined generative formats. Simultaneously, we also applied validation operations to facilitate consistency such as ensuring exact locations of label spans in the input and merging label fields for the same text instance. This was a very time-consuming process as we had to write processing scripts for each dataset individually and did not have a reliable way to automate it. When dealing with ambiguous guidelines or datasets lacking metadata, we further consulted the datasets' authors for clarifications.

**De-duplication** While a number of datasets explicitly provide repeated labels[22], many do not. For unification's sake, we applied set operations to retain only unique labels in each input text. For each task, in order to increase diversity of training texts, we further limit occurrence of each training input text to once. For example, if an input text appears in the training split of more than two datasets, only one version would be retained[23]. Besides, we remove duplicate entries (identical text input and schema labels) in each data split.

**De-contamination** For reliable evaluations, test time leakage should be prevented. We first recorded every unique text that appeared in the test splits of all datasets and obtained a global pool of test inputs. We then iterated through all training and development instances and removed those whose input text appeared in this global pool.

## C In-Context Learning Evaluation - Additional Details

**Language-specific instructions do not help:** Using Copilot[24], we translate manually written English instructions into four languages: German, Thai, Spanish, Japanese; and examine whether using language-

[18] https://www.google.com
[19] https://www.bing.com
[20] https://en.wikipedia.org/wiki/Bavaria
[21] https://en.wikipedia.org/wiki/Malaysian_English
[22] The same schema text appears twice or more in the input text
[23] We picked the first one appearing in the `os.listdir()` command
[24] https://www.bing.com/chat

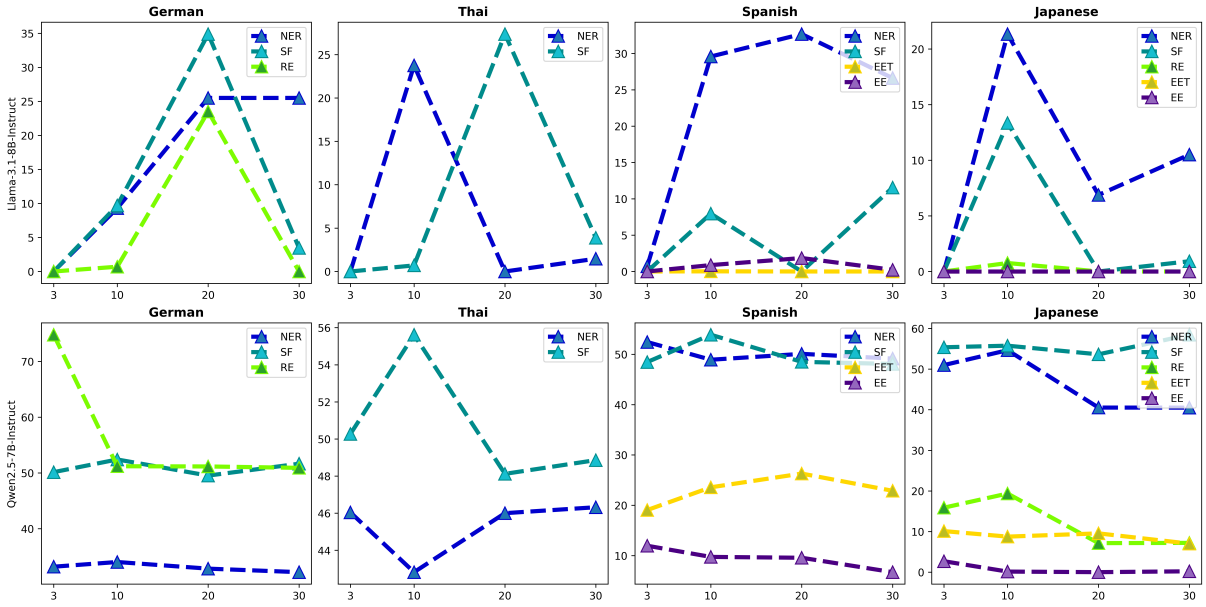


Figure 13: Varying number of language-specific demonstrations (English instructions) with few-shot default prompts (M-LIGHT)

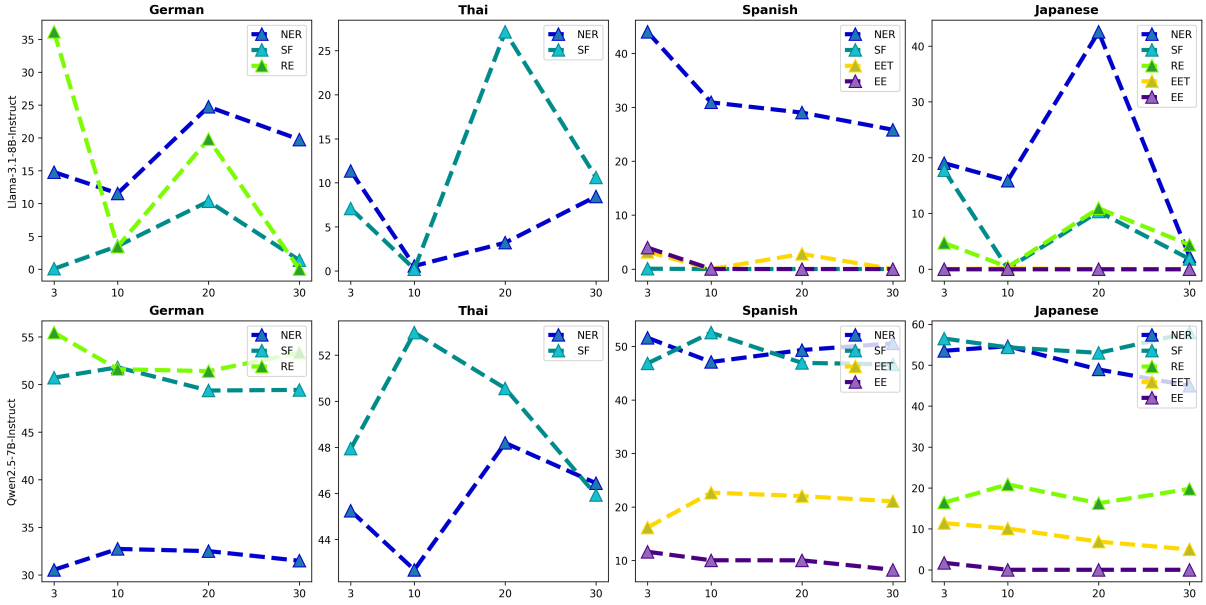


Figure 14: Varying number of language-specific demonstrations (language-specific instructions) with few-shot default prompts (M-LIGHT)

specific instructions would improve models' performance in IE tasks of that same language. Results are shown in Figure 15. In most cases, language-specific instructions do not give better results and even decrease model's performance significantly in some combinations of tasks and languages. For example, models' performance on the Thai language for both NER and SF tasks decrease substantially. For the Spanish language, LLAMA-3.1-8B-INSTRUCT achieves gains on the ED task but loses performance on three other tasks (NER,SF,EE). For the German language, the QWEN-2.5-7B-INSTRUCT model performs worse on all displayed tasks (NER, SF, RE). We hypothesize that this is because existing LLMs are English-dominant, making English instructions more helpful than language-specific alternatives, even when the task at hands involves inputs of that same language. Other reasons might include translation qualities and cross-lingual mismatch in technical terms i.e. equivalence of English jargons might not exist

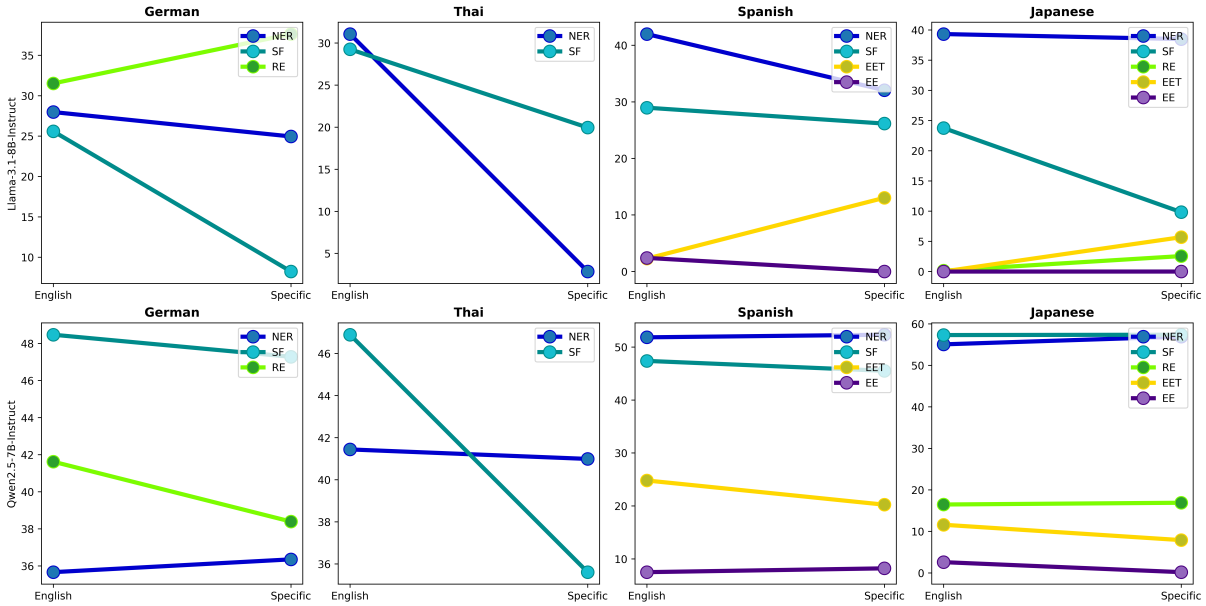


Figure 15: Language-specific instructions with few-shot default prompts (M-LIGHT)

in other languages, making it harder for models to understand the task.

# D Improving reasoning-based IE

## D.1 Settings

In Section 4, we discussed results obtained via incorporating intermediate thoughts into the default prompts. Although there were promising improvements, these improvements vary across tasks and are not consistently better than the default prompts without intermediate thoughts. In Section 5, we further showed that adopting a more optimal instruction and demonstration set can greatly impact prompting performance. In this section, we explore the usage of these more optimal instructions and demonstration sets in improving reasoning-based IE. This is particularly useful as reasoning-based IE possesses better interpretability than black-box IE (i.e. non-reasoning), and the intermediate thoughts can be utilized for further investigations (e.g. debugging or steering).

Concretely, we first repeat the optimization process in Section 5 with the prompt teacher $\Phi$ for few-shot prompting and obtain the more optimal instruction $I_{top}$ and demonstration set $S_{top}$. Then, for each exemplar $(x, y) \in S_{top}$, we inject an intermediate thought $T$ sampled from the thought teacher $\Theta$, creating the new set $S_{top,T} = (x, T, y)$. Since we induced changes in the demonstration set, the instruction $I_{top}$ might have become sub-optimal. Therefore, we re-generate candidate instructions $\{I\}$ and iteratively evaluate[25] $\{I\}$ (with $S_{top,T}$ fixed) on the tuning dataset to find a more optimal instruction $I_{top,T}$. Afterwards, the instruction $I_{top,T}$ and demonstration set $S_{top,T}$ are used at test time with the inference model $\theta$. We hereafter refer to this pipeline as AUTOTHOUGHT.

Next, we describe 3 variants of AUTOTHOUGHT, each differs in the choice of $T$ and $y$.

- AUTOTHOUGHT-1: Sample $T \sim P_\Theta(T|x, y)$, as in Section 4
- AUTOTHOUGHT-2: Sample $T \sim P_\Theta(T|x)$ and keep $y$ as the exemplar's label. In this case there might be mismatches between $T$ and $y$.
- AUTOTHOUGHT-3: Sample $T \sim P_\Theta(T|x)$ and replace label $y$ with augmented label $y' = \Theta(x, T)$, thus enabling consistency between $T$ and $y'$.

Additionally, we experiment with prompt optimization using augmented demonstrations $(x, T, y')$ only, where each intermediate thought is sampled as $T \sim P_\Phi(T|x)$ and label $y$ is replaced with augmented

[25]Note that we also adopt minibatching to approximate evaluations as in Section 5 for efficiency's sakes

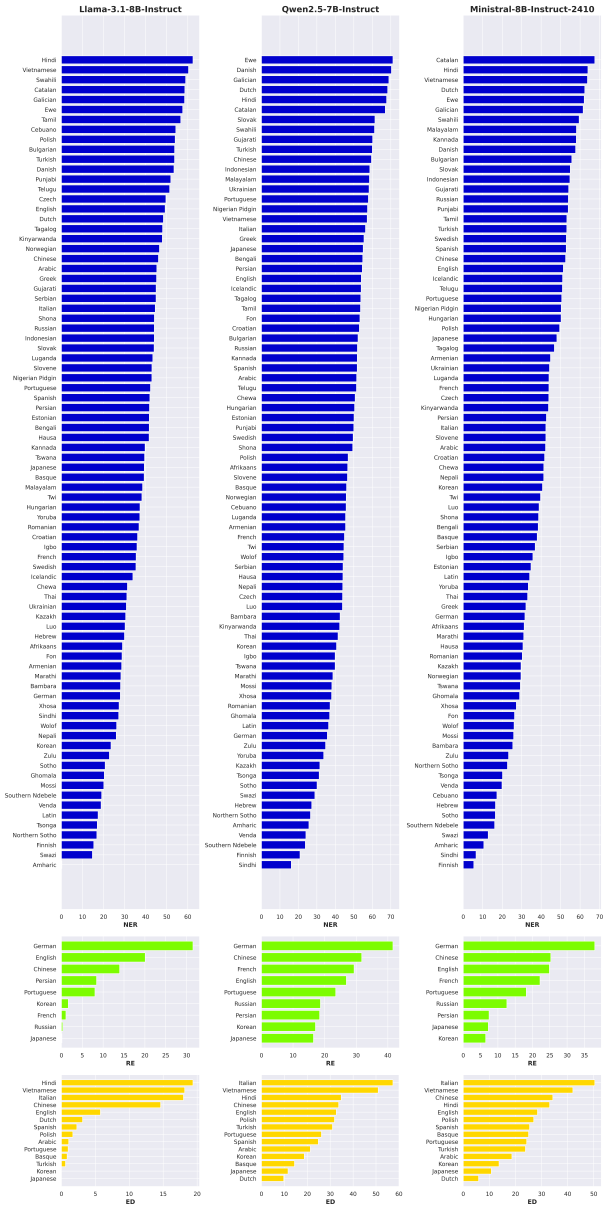


Figure 16: Language-wise results with few-shot default prompts (M-LIGHT)

label $y' = \Phi(x, T)$. Particularly, candidates for demonstration sets $S$ are first constructed by sampling from these augmented demonstrations. Afterwards, we perform instruction generation and subsequent Bayesian optimization based on these candidates (similar to Section 5). We refer to this variant as AUTOTHOUGHT-4.

**Hyperparameters** We maintain the same hyperparameters and tuning datasets as in Section 4 and 5.

**Models** For $\Phi$ (prompt teacher), we use LLAMA-3.1-8B-INSTRUCT and set $\Phi \equiv \theta$ (inference model). For $\Theta$ (thought teacher), we use LLAMA-3.1-70B-INSTRUCT.

## D.2 Results and Analyses

**AutoThought variants consistently improve over default prompts but underperform non-reasoning optimized prompts** Compared to DEFAULT-COT (Sec. 4) which underperforms default prompts on RE and EE tasks, all AUTOTHOUGHT variants perform better than default prompts on $5/5$ tasks (Figure 18), which again highlights the necessity for proper instructions and demonstrations. However, we observe that AUTOTHOUGHT variants usually perform worse than OPTIM (Sec. 5), with one exception in the RE task where AUTOTHOUGHT-4 outperforms OPTIM. This shows that while reasoning-based prompting has the

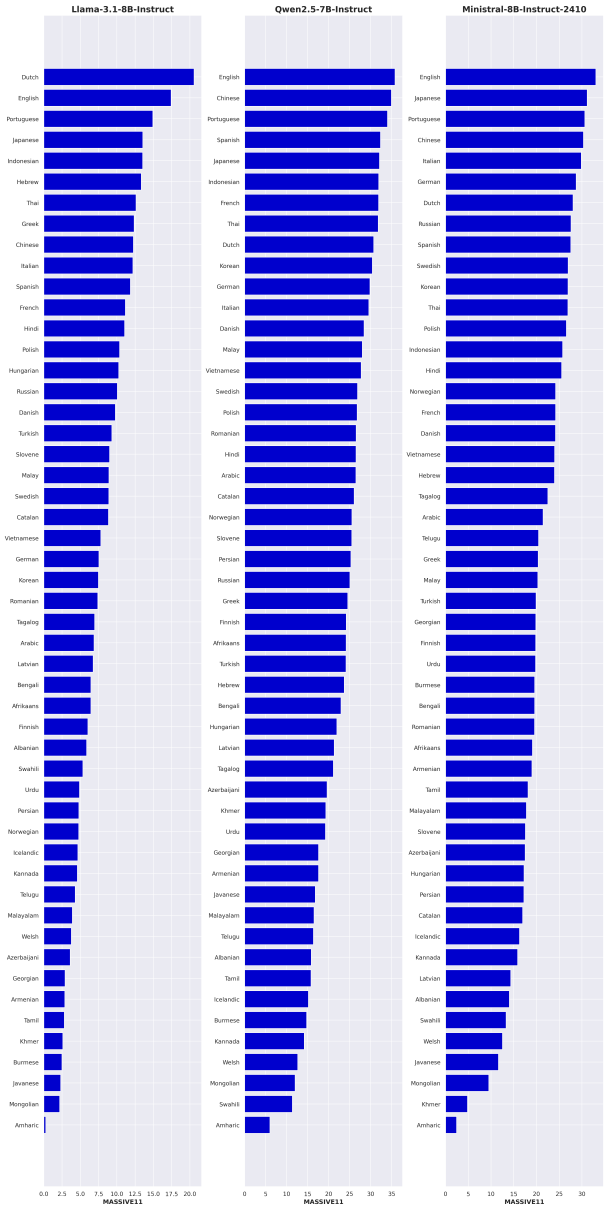


Figure 17: Language-wise results with few-shot default prompts (MASSIVE11 in M-HEAVY)

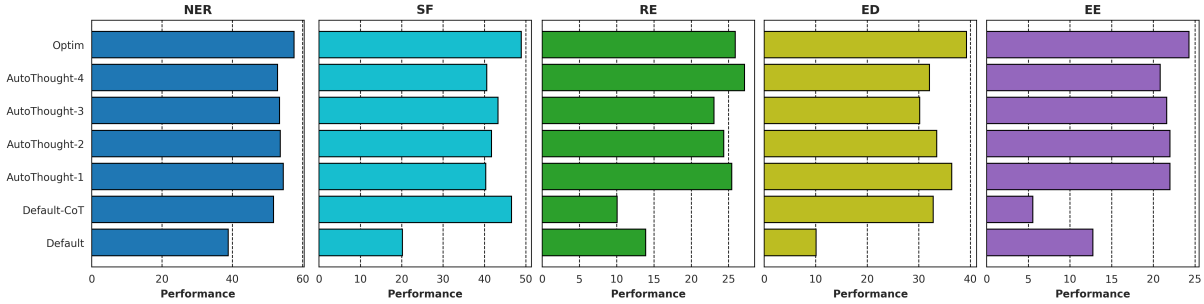


Figure 18: Reasoning-based and non-reasoning few-shot optimized prompts (M-LIGHT)

potential to outperform non-reasoning prompting in IE, in practice non-reasoning optimized prompting often leads to better performance.

**AutoThought-1 performs the best among reasoning-based prompting approaches** We measure the average performance over 5 tasks of each prompting approach in Table 4. Among reasoning-based

| | Avg. (5 tasks) |
|---|---|
| Optim (Sec. 5) | 39.17 |
| AutoThought-4 | 34.68 |
| AutoThought-3 | 34.32 |
| AutoThought-2 | 35.04 |
| **AutoThought-1** | **35.73** |
| Default-CoT | 29.32 |
| Default | 19.15 |

Table 4: Average LF1 scores over 5 tasks. We highlight the best reasoning-based prompting approach.

approaches, we find that AUTOTHOUGHT-1 attains the highest score, suggesting that sampling thought $T \sim P_\Theta(T|x, y)$ might be the better option overall. We note that there still exists a significant gap between AUTOTHOUGHT-1 and OPTIM, necessitating further changes to make reasoning-based optimized prompting as good as non-reasoning optimized prompting for IE.

## E Supervised Fine-tuning Evaluation

### E.1 Settings

In this section, we experiment with supervised fine-tuning LLMs on M-LIGHT. To measure LLMs' zero-shot abilities, we only train on 8 source languages and conduct inference on all 96 languages.

**Training Languages** We choose 8 languages with the highest number of samples in the training split of M-LIGHT: English, Chinese, German, Turkish, Spanish, Thai, Italian, Japanese. Note that development data for checkpoint selection also employ the same set of languages.

**Training Methods** For models whose sizes $\leq 32B$, we use LORA (Hu et al., 2022). For larger models, we use QLORA with 4-bit quantization (Dettmers et al., 2023).

**Hyperparameters** For LORA/QLORA, we use $r = 64$, $\alpha = 128$, and insert adapters on all linear modules. For LORA, we use mixed precision training with BFLOAT16. Each model is trained with an accumulated batch size of 48, with initial learning rate $5e-5$ and a linear warmup schedule with 10000 warmup steps. Checkpoints are evaluated every 3000 steps and the checkpoint with the lowest development loss is chosen for test time inference. For optimizer, we use ADAMW (Loshchilov and Hutter, 2019) with $\beta_1 = 0.9$, $\beta_2 = 0.95$, $\epsilon = 1e-08$. For efficiency's sake, we limit training/development samples' length to a maximum of 800. Note that we do not apply any such limit for test time inference. We set the maximum number of training steps as 1000000 in the scheduler, but we observe in practice that all models converge well under 50000 steps, in which case we stop the training early.

**Models** We choose among instruction-following models from the LLAMA-3.1, QWEN2.5, AYA-EXPANSE, MISTRAL and MINISTRAL series.

### E.2 Results and Analyses

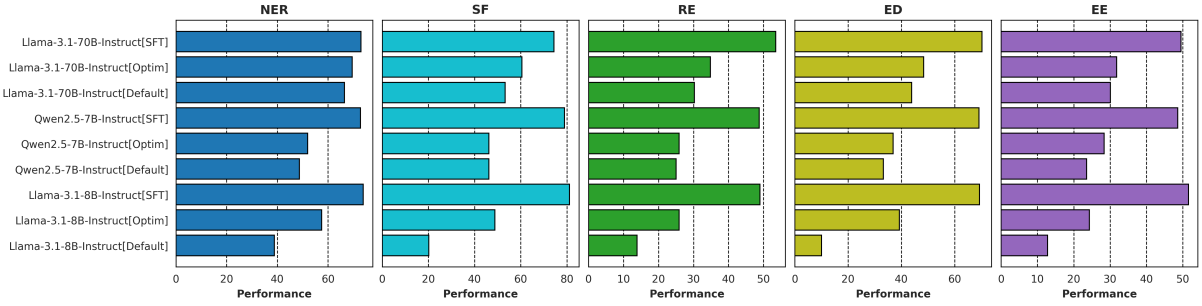


Figure 19: SFT results in the MIXED settings, evaluated on both *seen* and *unseen* languages (M-LIGHT)

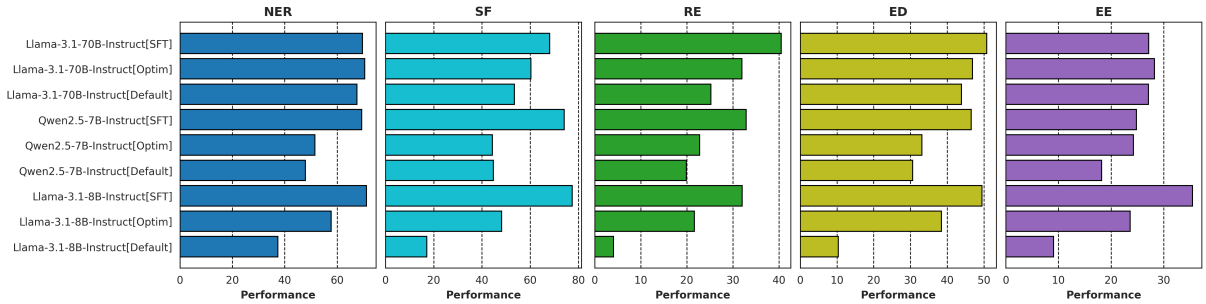


Figure 20: SFT results in the ZERO-SHOT settings, evaluated on *unseen* languages (M-LIGHT)

**SFT largely improves models' improvements in both mixed and zero-shot settings** Here we consider the set of test languages not included in the training data for SFT as *unseen* and showed the results in two settings: *mixed* and *zero-shot* (Figure 19, 20). These settings reflect models' abilities in cross-lingual transfer (Zuo et al., 2024; Pham et al., 2024; Le et al., 2024b; Nguyen et al., 2023a; Asai et al., 2024; Le, 2024), each within a distinct scope. We compare the SFT results with DEFAULT (Sec. 3) and OPTIM (Sec. 5). Overall, we observe substantial improvements of SFT over the two ICL baselines, even when the test set only features languages not included during SFT training.

**But SFT does not always outperform strong ICL prompting** Our training set for SFT consists of $789K$ samples in 5 tasks, with at least $54K$ samples per task. Even after training on such a large number of samples, we notice cases where the SFT model still underperforms a strong ICL baseline, particularly with the LLAMA-3.1-70B-INSTRUCT model evaluated zero-shot on the NER and EE tasks (Figure 20) underperforming the OPTIM baseline. Although this may have been partially caused by QLORA's instability, we find this observation rather surprising as there is no such precedent in previous IE works, which was likely due to previous works not adopting optimized prompting.

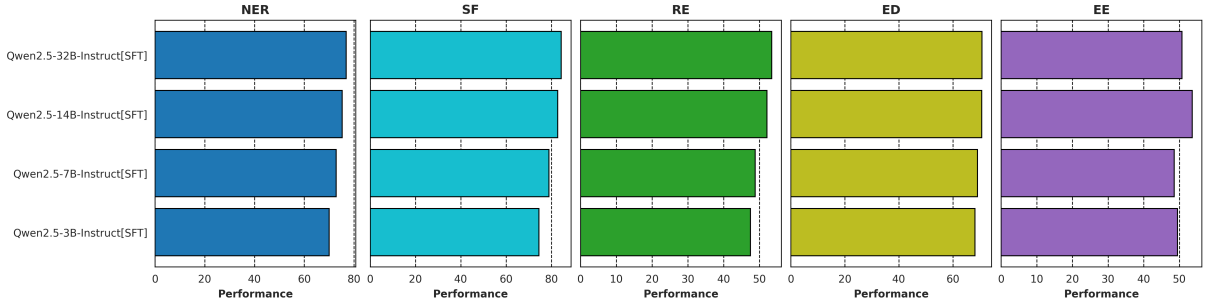


Figure 21: Parameter Scaling with QWEN2.5: SFT results in the MIXED settings, evaluated on both *seen* and *unseen* languages (M-LIGHT)

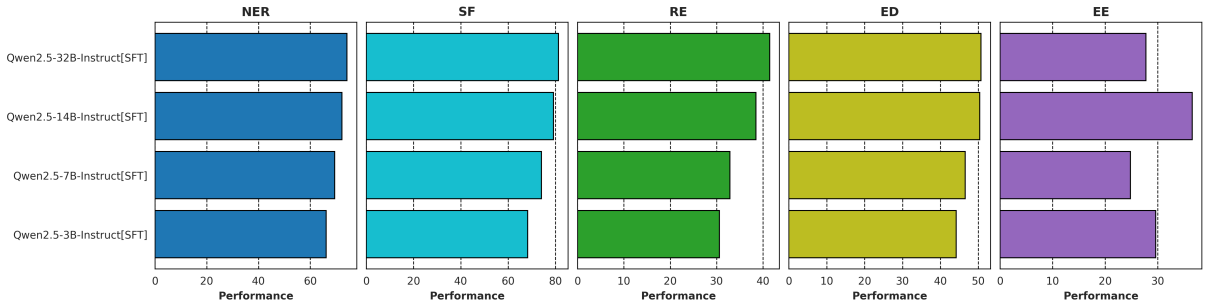


Figure 22: Parameter Scaling with QWEN2.5: SFT results in the ZERO-SHOT settings, evaluated on *unseen* languages (M-LIGHT)

**Scaling laws for SFT generally hold but is less significant than ICL** We observe the influence of

scaling parameters with the QWEN2.5 model series. Although there is an increasing (but non-monotonic) trend, we find the difference less significant than observed in few-shot ICL (Sec. 3).

### E.3 Overall Performance

| | NER | SF | RE | ED | EE | Avg. |
|---|---|---|---|---|---|---|
| Llama-3.1-8B-Instruct | 73.89 | 81.16 | 48.98 | 69.34 | 51.54 | 64.98 |
| Llama-3.1-70B-Instruct | 72.98 | 74.40 | 53.51 | 70.33 | 49.53 | 64.15 |
| Qwen2.5-3B-Instruct | 70.08 | 74.46 | 47.48 | 68.14 | 49.46 | 61.92 |
| Qwen2.5-7B-Instruct | 72.83 | 78.92 | 48.78 | 69.13 | 48.59 | 63.65 |
| Qwen2.5-14B-Instruct | 75.30 | 82.78 | 52.04 | 70.70 | 53.60 | 66.88 |
| Qwen2.5-32B-Instruct | 76.85 | 84.36 | 53.37 | 70.73 | 50.73 | 67.21 |
| aya-expanse-8b | 69.33 | 74.93 | 49.91 | 68.55 | 51.24 | 62.79 |
| Ministral-8B-Instruct-2410 | 63.64 | 67.65 | 42.42 | 64.70 | 44.47 | 56.58 |
| Mistral-7B-Instruct-v0.3 | 64.42 | 69.44 | 36.70 | 61.17 | 36.51 | 53.65 |

Table 5: Results on M-LIGHT (mixed). Models were trained on 8 languages of M-LIGHT.

| | NER | SF | RE | ED | EE | Avg. |
|---|---|---|---|---|---|---|
| Llama-3.1-8B-Instruct | 70.46 | 62.03 | 49.65 | 50.12 | 47.31 | 55.91 |
| Qwen2.5-3B-Instruct | 66.67 | 54.59 | 49.08 | 50.99 | 45.43 | 53.35 |
| Qwen2.5-7B-Instruct | 69.27 | 60.64 | 49.42 | 53.76 | 43.30 | 55.28 |
| Qwen2.5-14B-Instruct | 71.60 | 63.81 | 53.31 | 49.95 | 48.92 | 57.52 |
| Qwen2.5-32B-Instruct | 73.79 | 67.20 | 54.88 | 52.66 | 47.15 | 59.14 |
| aya-expanse-8b | 66.58 | 55.28 | 49.62 | 41.69 | 43.27 | 51.29 |
| Ministral-8B-Instruct-2410 | 59.94 | 50.24 | 45.36 | 42.98 | 37.69 | 47.24 |
| Mistral-7B-Instruct-v0.3 | 60.66 | 48.40 | 40.56 | 43.20 | 33.80 | 45.32 |

Table 6: Results on M-HEAVY (mixed). Models were trained on 8 languages of M-LIGHT.

We report the overall results of each model on M-LIGHT and M-HEAVY in Table 5 and 6.

## F PEFT Ablation

### F.1 Settings

In this section, we move beyond LORA and explore 6 other *parameter-efficient fine-tuning* (PEFT) methods: FOURIERFT (Gao et al., 2024), BOFT (Liu et al., 2024b), GALORE (Zhao et al., 2024a), VERA (Kopiczko et al., 2024), LISA (Pan et al., 2024), DORA (Liu et al., 2024a). Our purpose is to discover if there exists a better PEFT method for multilingual IE than LORA - which remains the standard methods of most works.

**Hyperparameters** Generally training settings remain the same as in Appendix E. However, while LORA models always converge under 40000 steps, we find this to not be the case with several other PEFT methods. For practical comparisons, we only allow each method to train up-to a maximum of 70000 steps and select the best checkpoint obtained until then for test time inference. Regarding method-specific hyperparameters, for FOURIERFT, we set $n_{frequency} = 2000$ and $scaling = 300$. For BOFT, we set $block\_size = 4$ and $dropout = 0.0$. For GALORE, we set $rank = 128$, $update\_projection\_gap = 200$

and $scale = 1.0$. For LISA, we set number of activated layers as 2 and step interval as 20. For VERA, we set $rank = 256$ and $d_{initial} = 0.1$.

### F.2 Results and Analyses

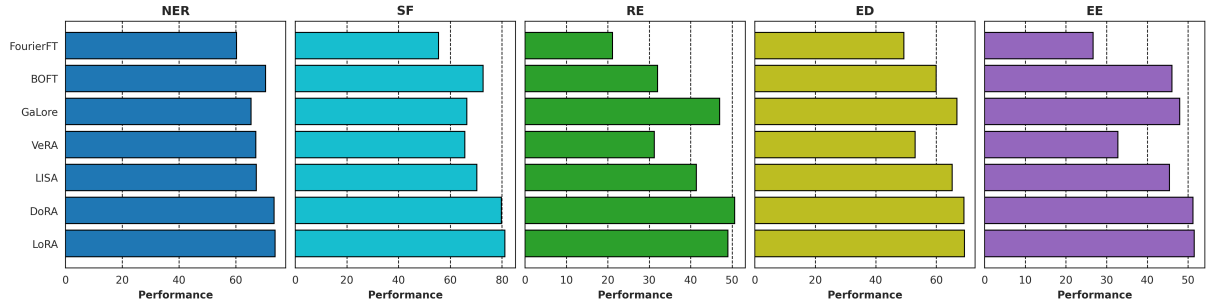


Figure 23: Evaluation of LORA and 6 other PEFT methods on M-LIGHT (*mixed*)

**Other PEFT methods do not outperform LoRA** Results are shown in Figure 23. Surprisingly, no evaluated PEFT method manages to surpass LORA, which suggests that LORA still remains a strong baseline for SFT in multilingual IE.

## G Multi-Task vs Single-Task

### G.1 Settings

In Appendix E, we explored training with multi-task samples. In this section, we examine the performance of training with individual tasks only i.e. one LORA adapter per task.

**Models** We use LLAMA-3.1-8B-INSTRUCT and QWEN2.5-7B-INSTRUCT.

### G.2 Results and Analyses

| | NER | SF | RE | EET | EE | Avg. |
|---|---|---|---|---|---|---|
| LLAMA-3.1-8B-INSTRUCT-SINGLE | 73.29 | 83.68 | 47.81 | 69.12 | 49.34 | 64.65 |
| LLAMA-3.1-8B-INSTRUC-MULTI | 73.89 | 81.16 | 48.98 | 69.34 | 51.54 | **64.98** |
| QWEN2.5-7B-INSTRUCT-SINGLE | 73.43 | 80.40 | 51.00 | 68.69 | 47.23 | **64.15** |
| QWEN2.5-7B-INSTRUCT-MULTI | 72.83 | 78.92 | 48.78 | 69.13 | 48.59 | 63.65 |

Table 7: Multi- and Single-Task SFT results on M-LIGHT (mixed). We highlight the highest average score for each model.

**Multi-task training behaves differently depending on each task and the base model** For LLAMA-3.1-8B-INSTRUCT, we find multi-task training to often help improve results ($4/5$ tasks). For QWEN2.5-7B-INSTRUCT we find single-task training to often achieve better results ($3/5$) tasks. Thus, the effect of multi-task training is highly dependent on the specific combination of task and base model.

## H Simulated Preference Training

### H.1 Settings

In this section, we experiment with preference training for multilingual IE. Since we have ground truth labels in M-LIGHT, we can directly treat these as positive references. For negative references, we bootstrapped predictions from the LLAMA-3.1-8B-INSTRUCT SFT model and filtered out those with perfect LF1 scores, ultimately obtaining $125K$ negative references. We then pair these with the ground truth labels for preference training. We treat this as *simulated preference training* as the concept of *preference* is not clearly defined for IE.

**Training Loss** We explore DPO (Rafailov et al., 2023), CPOSIMPO (Meng et al., 2024) and ORPO (Hong et al., 2024).

**Hyperparameters** We set a learning rate of $5e-7$ and $\beta = 0.01$ for DPO and ORPO. For CPOSIMPO, we use a learning rate of $1e-7$, $\beta = 2.5$, $\alpha = 1$ and $\gamma = 1.375$. Training adopts LORA adapters similar to Appendix E.

**Model** We start training from the LLAMA-3.1-8B-INSTRUCT SFT model.

### H.2 Results and Analyses

| | NER | SF | RE | ED | EE | Avg. |
|---|---|---|---|---|---|---|
| LLAMA-3.1-8B-INSTRUCT-SFT | 73.89 | 81.16 | 48.98 | 69.34 | 51.54 | 64.98 |
| LLAMA-3.1-8B-INSTRUCT-SFT-DPO | 73.45 | **82.01** | 51.74 | **69.94** | 53.38 | 66.10 |
| LLAMA-3.1-8B-INSTRUCT-SFT-ORPO | **74.14** | 81.45 | **52.22** | 69.46 | **53.59** | **66.17** |
| LLAMA-3.1-8B-INSTRUCT-SFT-CPOSIMPO | 71.83 | 80.67 | 51.43 | 68.93 | 51.64 | 64.90 |

Table 8: Preference training results on M-LIGHT (mixed). We highlight the highest score in each column.

**Preference training can help, but it depends on the loss** We show results in Table 8. We find that the bootstrapped preference pairs can be used to further improve performance, but it also depends on the loss e.g. DPO and ORPO show clear improvements, while CPOSIMPO often underperforms the SFT baseline.

## I Implementation - Additional Details

**Score Aggregation** For each data split $D(T, Test, L)$ in the test set (either M-LIGHT or M-HEAVY) of task $T$, where $L$ represents a language in $D$. We denote $NI(D(T, Test, L))$ and $LF1_\theta(D(T, Test, L))$ accordingly as the number of instances in $D(T, Test, L)$ and the $LF1$ scores of model $\theta$ on $D(T, Test, L)$. We then calculate the overall score of each task by adopting a weighted average of $LF1_\theta(D(T, Test, L))$ with $NI(D(T, Test, L))$ as the weight, marginalizing over all possible $D$ and $L$. Note that we rely on the number of instances instead of samples, making the results independent of $split\_num$ value. We choose to adopt weighted average instead of macro average as a number of data splits have very small number of samples, causing high variances and make results less stable for macro averaging.

**Demonstration/Tuning Dataset** List of English demonstration/tuning datasets (Sec. 3, 4, 5): **SF**: XSID (van der Goot et al., 2021; Aepli et al., 2023; Winkler et al., 2024a,b); **NER**: CONLL2003 (Tjong Kim Sang and De Meulder, 2003); **RE**: NEWS_CROSSRE (Bassignana and Plank, 2022); **ED**: ACE2005 (Doddington et al., 2004); **EE**: ACE2005 (Doddington et al., 2004)

List of Japanese demonstration datasets (Sec. 3): **NER**: COARSE_ANYTHINGNER (Luo et al., 2024); **RE**: ROR (Bou et al., 2020); **SF**: MASSIVE11 (FitzGerald et al., 2023); **EE**: MEE (Pouran Ben Veyseh et al., 2022a); **ED**: MEE (Pouran Ben Veyseh et al., 2022a)

List of Thai demonstration datasets (Sec. 3): **NER**: THAINER22 (Mr.Wannaphong); **SF**: MASSIVE11 (FitzGerald et al., 2023)

List of German demonstration datasets (Sec. 3): **NER**: COARSE_ANYTHINGNER (Luo et al., 2024); **RE**: COFUN (Foroutan et al., 2024); **SF**: MASSIVE11 (FitzGerald et al., 2023)

List of Spanish demonstration datasets (Sec. 3): **NER**: COARSE_ANYTHINGNER (Luo et al., 2024); **SF**: MASSIVE11 (FitzGerald et al., 2023); **EE**: MEE (Pouran Ben Veyseh et al., 2022a); **ED**: MEE (Pouran Ben Veyseh et al., 2022a)

**Hyperparameters** We use greedy decoding (temperature 0) for test time inference of all LLMs. For non-reasoning prompts, we set max generation length as 256. For reasoning-based prompts, we set max generation length as 1024. We generally do not set any threshold for inputs' lengths and only truncate them when they either exceed the LLM's context limit or there is not enough KV cache memory (Kwon et al., 2023).

**Frameworks** Experiments relied on VLLM (Kwon et al., 2023), MS-SWIFT (Zhao et al., 2024b),

TRANSFORMERS (Wolf et al., 2020) and PYTORCH (Paszke et al., 2019). To implement prompt optimization, we utilize the DSPY library (Khattab et al., 2024).

**Accelerators** All experiments were conducted with 6 A100 PCIE and 1 H100 HBM3 GPUs.

## J Language Distribution in MASSIE

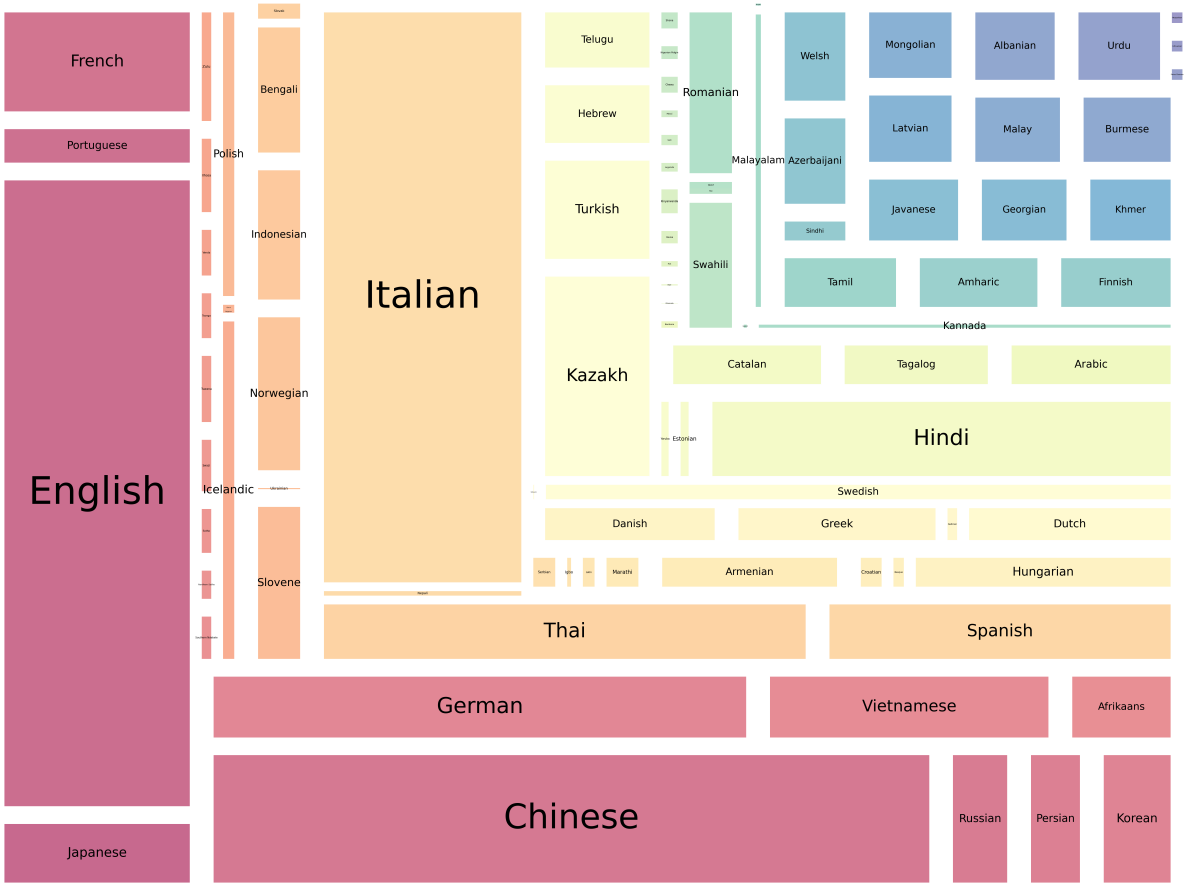


Figure 24: Language distribution in M-HEAVY

## K List of languages

Table 9: List of languages in MASSIE

| Code | Language | Family |
|---|---|---|
| afr | Afrikaans | Indo-European |
| amh | Amharic | Afro-Asiatic |
| ara | Arabic | Afro-Asiatic |
| aze | Azerbaijani | Turkic |
| bam | Bambara | Mande |
| bbj | Ghomala | Atlantic-Congo |
| ben | Bengali | Indo-European |
| bul | Bulgarian | Indo-European |
| cat | Catalan | Indo-European |
| ceb | Cebuano | Austronesian |
| ces | Czech | Indo-European |
| cym | Welsh | Indo-European |

| Code | Language | Family |
|---|---|---|
| dan | Danish | Indo-European |
| deu | German | Indo-European |
| ell | Greek | Indo-European |
| eng | English | Indo-European |
| est | Estonian | Uralic |
| eus | Basque | Language isolate |
| ewe | Ewe | Atlantic-Congo |
| fas | Persian | Indo-European |
| fin | Finnish | Uralic |
| fon | Fon | Atlantic-Congo |
| fra | French | Indo-European |
| glg | Galician | Indo-European |
| gsw | Swiss German | Indo-European |
| guj | Gujarati | Indo-European |
| hau | Hausa | Afro-Asiatic |
| heb | Hebrew | Afro-Asiatic |
| hin | Hindi | Indo-European |
| hrv | Croatian | Indo-European |
| hun | Hungarian | Uralic |
| hye | Armenian | Indo-European |
| ibo | Igbo | Atlantic-Congo |
| ind | Indonesian | Austronesian |
| isl | Icelandic | Indo-European |
| ita | Italian | Indo-European |
| jav | Javanese | Austronesian |
| jpn | Japanese | Japonic |
| kan | Kannada | Dravidian |
| kat | Georgian | Kartvelian |
| kaz | Kazakh | Turkic |
| khm | Khmer | Austroasiatic |
| kin | Kinyarwanda | Atlantic-Congo |
| kor | Korean | Koreanic |
| lat | Latin | Indo-European |
| lav | Latvian | Indo-European |
| lit | Lithuanian | Indo-European |
| lug | Luganda | Atlantic-Congo |
| luo | Luo | Nilotic |
| mal | Malayalam | Dravidian |
| mar | Marathi | Indo-European |
| mon | Mongolian | Mongolic |
| mos | Mossi | Atlantic-Congo |
| msa | Malay | Austronesian |
| mya | Burmese | Sino-Tibetan |
| nap | Neapolitan | Indo-European |
| nbl | Southern Ndebele | Atlantic-Congo |
| nep | Nepali | Indo-European |
| nld | Dutch | Indo-European |
| nor | Norwegian | Indo-European |
| nso | Northern Sotho | Atlantic-Congo |
| nya | Chewa | Atlantic-Congo |
| pan | Punjabi | Indo-European |

| Code | Language | Family |
|---|---|---|
| pcm | Nigerian Pidgin | Indo-European |
| pol | Polish | Indo-European |
| por | Portuguese | Indo-European |
| ron | Romanian | Indo-European |
| rus | Russian | Indo-European |
| slk | Slovak | Indo-European |
| slv | Slovene | Indo-European |
| sna | Shona | Atlantic-Congo |
| snd | Sindhi | Indo-European |
| sot | Sotho | Atlantic-Congo |
| spa | Spanish | Indo-European |
| sqi | Albanian | Indo-European |
| srp | Serbian | Indo-European |
| ssw | Swazi | Atlantic-Congo |
| swa | Swahili | Niger-Congo |
| swe | Swedish | Indo-European |
| tam | Tamil | Dravidian |
| tel | Telugu | Dravidian |
| tgl | Tagalog | Austronesian |
| tha | Thai | Kra-Dai |
| tsn | Tswana | Atlantic-Congo |
| tso | Tsonga | Atlantic-Congo |
| tur | Turkish | Turkic |
| twi | Twi | Atlantic-Congo |
| ukr | Ukrainian | Indo-European |
| urd | Urdu | Indo-European |
| ven | Venda | Atlantic-Congo |
| vie | Vietnamese | Austroasiatic |
| wol | Wolof | Atlantic-Congo |
| xho | Xhosa | Atlantic-Congo |
| yor | Yoruba | Atlantic-Congo |
| zho | Chinese | Sino-Tibetan |
| zul | Zulu | Atlantic-Congo |

## L List of datasets

- **SF**: MASSIVE11 (FitzGerald et al., 2023), MTODS (Schuster et al., 2019), MTOP (Li et al., 2021a), XSID (van der Goot et al., 2021; Aepli et al., 2023; Winkler et al., 2024a,b), BAVARIAN_XSID (van der Goot et al., 2021; Aepli et al., 2023; Winkler et al., 2024a,b)

- **NER**: MPHAYANER (Mbuvha et al., 2023), NERWIKIJAPANESE (Inc., 2021), EVEREST (Niraula and Chapagain, 2022), WEB_NYTK (Simon and Vad'asz, 2021), MAHASOCIALNER (Chaudhari et al., 2024), KLUENER (Park et al., 2021), WEBSITES_MIMGOLD (Ing'olfsd'ottir et al., 2020), CONLL2003 (Tjong Kim Sang and De Meulder, 2003), CANTEMIST (Miranda-Escalada et al., 2020), DDT_UNER (Mayhew et al., 2024), TOPRES19TH_HIPE2022 (Ehrmann et al., 2022), ANATEM (Pyysalo and Ananiadou, 2014), SSJ500K-SYN.UD_SUK (Arhar Holdt et al., 2024), CODEMIXEDTWEET (Singh et al., 2018a), TLUNIFIED (Miranda, 2023), GJA_UNER (Mayhew et al., 2024), LST20 (Boonkwan et al., 2020), FINDVEHICLE (Guan et al., 2024), ELNER4 (Bartziokas et al., 2020), HINERORIGINAL (Murthy et al., 2022), NEWSEYE_HIPE2022 (Ehrmann et al., 2022), NOB_NORNE (Jørgensen et al., 2020), BOOKS_MIMGOLD (Ing'olfsd'ottir et al., 2020), CODEMIXEDMSAEA (Aguilar et al., 2018), CODEMIXEDSPAENG (Aguilar et al.,

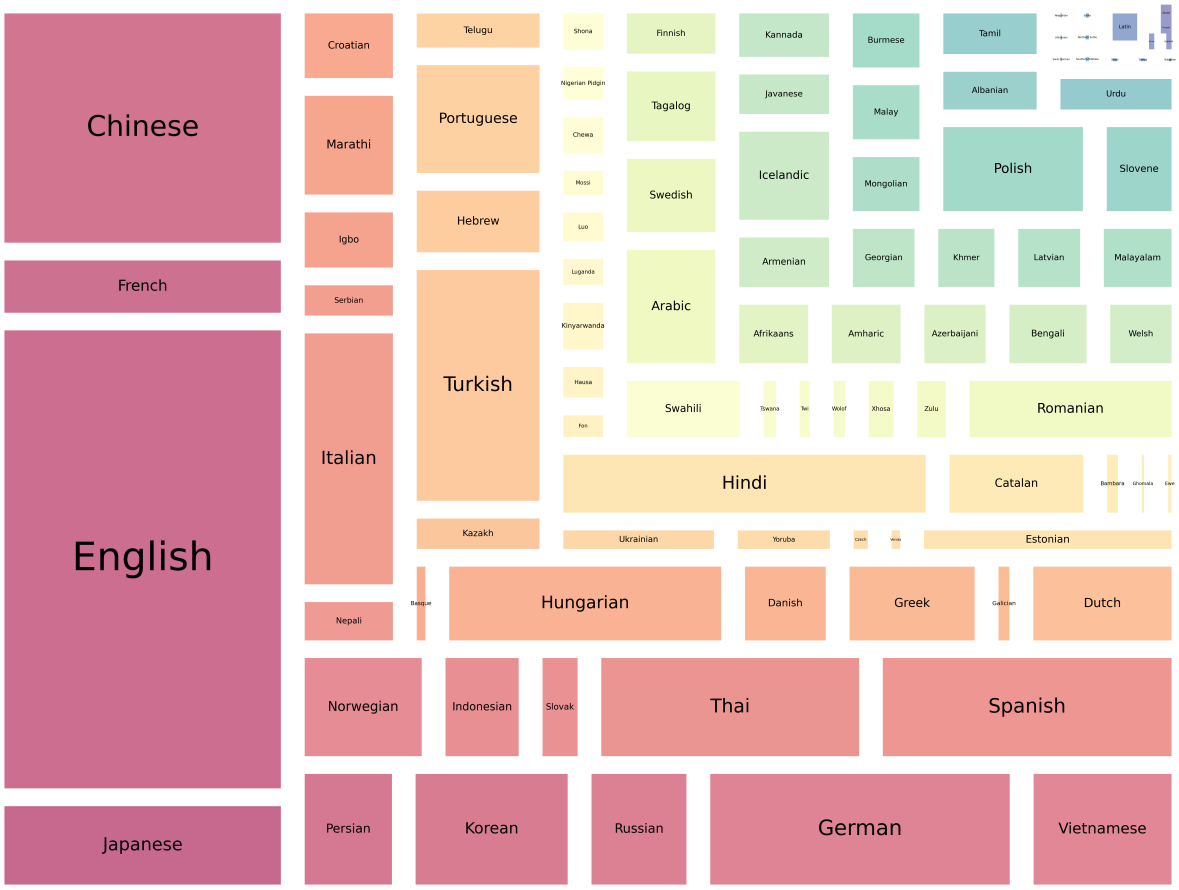


Figure 25: Language distribution in M-LIGHT

2018), BLOG_MIMGOLD (Ing'olfsd'ottir et al., 2020), ONTONOTES5 (Pradhan et al., 2013), NERUK20 (Chaplynskyi and Romanyshyn, 2024), CODEMIXEDSOCIAL (Singh et al., 2018b), ASIABIBI_SLAVICNER (Piskorski et al., 2024), TURKUNER (Luoma et al., 2020), GERMANLERFINE (Leitner et al., 2020), NCBIDISEASE (Doğan et al., 2014), CW_HERODOTOS (Erdmann et al., 2019), SENTENCE_ESTNER (Tkachenko et al., 2013; Sirts, 2023), CZECHHISTORY10 (Hubkov'a et al., 2020), PHONER (Truong et al., 2021), WEIBONER2ND (Peng and Dredze, 2015), FENEC (Millour et al., 2024), SETSR (Batanovi'c et al., 2023), SCHOOLESSAYS_MIMGOLD (Ing'olfsd'ottir et al., 2020), GSDSIMP_UNER (Mayhew et al., 2024), ANCORA (Armengol-Estap'e et al., 2021), DARNERCORP (Moussa and Mourhir, 2023), HIPE2020_HIPE2022 (Ehrmann et al., 2022), WIKIPEDIA_NYTK (Simon and Vad'asz, 2021), NERPNERPROSA (Hoesen and Purwarianti, 2018), PIONER (Ghukasyan et al., 2018), AMHARICNER (Jibril and Tantuğ, 2023), BOSQUE_UNER (Mayhew et al., 2024), PLINYELDER_HERODOTOS (Erdmann et al., 2019), WN_EVALITA2023 (Paccosi and Palmero Aprosio, 2022), NEMOCORPUS (Bareket and Tsarfaty, 2021), RELDINORMTAGNER-HR (Ljube si'c et al., 2023), USELECTION2020_SLAVICNER (Piskorski et al., 2024), ELEXISWSD.UD_SUK (Arhar Holdt et al., 2024), NORDSTREAM_SLAVICNER (Piskorski et al., 2024), NCHLT (Eiselen, 2016), MAHANER (Litake et al., 2022), PPORTALNER (Silva and Moro, 2024), TALBANKEN_UNER (Mayhew et al., 2024), NAAMPADAM (Mhaske et al., 2023), SIC (Östling, Robert et al., 2013), FINED_ANYTHINGNER (Luo et al., 2024), RYANAIR_SLAVICNER (Piskorski et al., 2024), BREXIT_SLAVICNER (Piskorski et al., 2024), TRG_UNER (Mayhew et al., 2024), ADJUDICATIONS_MIMGOLD (Ing'olfsd'ottir et al., 2020), IGBONER (Chukwuneke et al., 2022), NEWS_NYTK (Simon and Vad'asz, 2021), CONLL2002 (Tjong Kim Sang, 2002), AI_CROSSNER (Liu et al., 2020), EIEC (Alegria et al., 2004), HR500K (Ljube si'c and Samard zi'c, 2023), MILLIYET (Tür et al., 2003), SIMONERO (Mitrofan and Mititelu, 2020), EWT_UNER

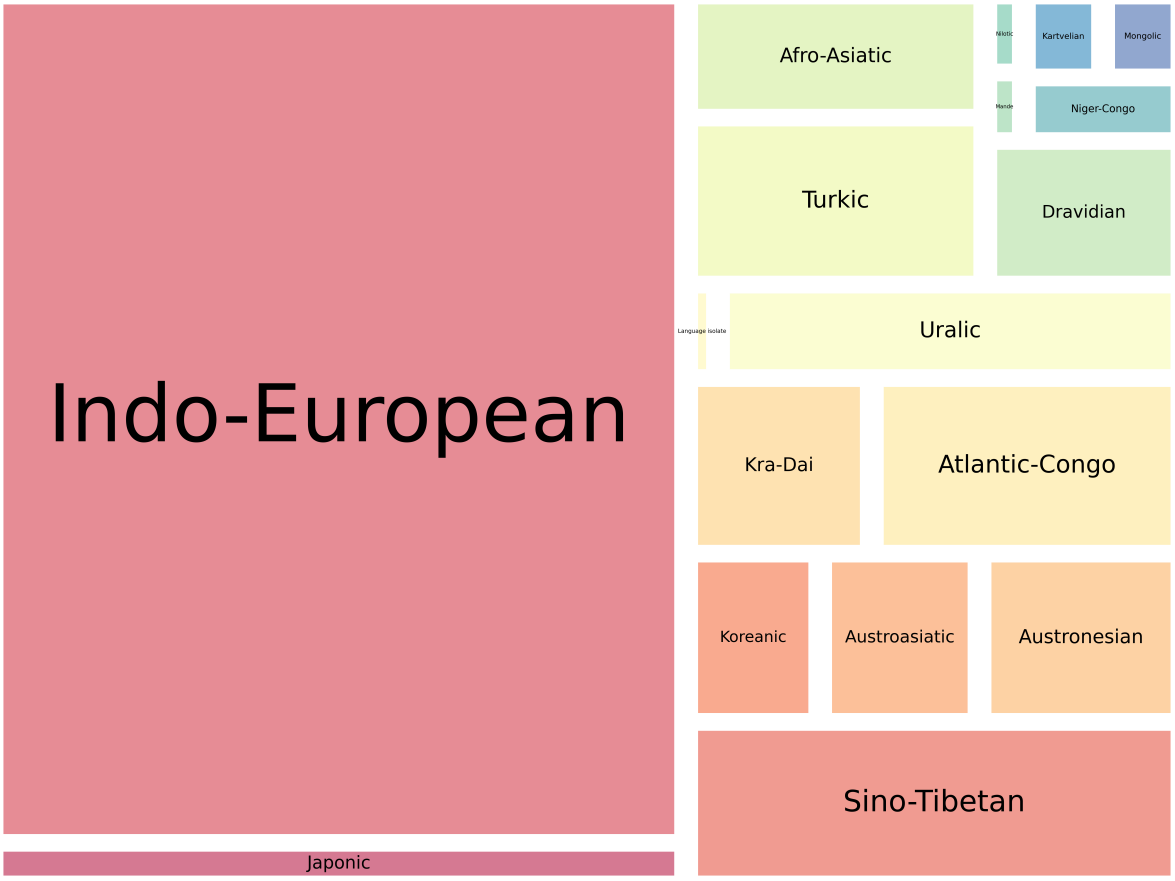


Figure 26: Family distribution in M-LIGHT

(Mayhew et al., 2024), DDT (Hvingelby et al., 2020), RESUMENER (Zhang and Yang, 2018), CLUENER (Xu et al., 2020), FIC_EVALITA2023 (Paccosi and Palmero Aprosio, 2022), RELDI (Ljubesi'c et al., 2023), THAINER22 (Mr.Wannaphong), CARANER (Al-Thubaity et al., 2022), LEGALNERO (Păiș et al., 2022), SCIENCEWEB_MIMGOLD (Ing'olfsd'ottir et al., 2020), SLIGALICIAN (Agerri et al., 2018), ELNER18 (Bartziokas et al., 2020), SENTICOREF.UD_SUK (Arhar Holdt et al., 2024), ENPNER (Blouin et al., 2024), FINERDATA (Ruokolainen et al., 2019), WRITTEN-TO-BE-SPOKEN_MIMGOLD (Ing'olfsd'ottir et al., 2020), LEGAL_NYTK (Simon and Vad'asz, 2021), SENTENCE_ESTNERNEW (Sirts, 2023), RADIOTVNEWS_MIMGOLD (Ing'olfsd'ottir et al., 2020), MUSIC_CROSSNER (Liu et al., 2020), WEBMEDIA_MIMGOLD (Ing'olfsd'ottir et al., 2020), SET_UNER (Mayhew et al., 2024), GERMANLERCOARSE (Leitner et al., 2020), RUSSIA-UKRAINEWAR_SLAVICNER (Piskorski et al., 2024), COVID-19_SLAVICNER (Piskorski et al., 2024), NNO_NORNE (Jørgensen et al., 2020), FBL_MIMGOLD (Ing'olfsd'ottir et al., 2020), PARLIAMENTARY (Jonkers, 2016), LREC16KB (Garcia, 2016), GSD_UNER (Mayhew et al., 2024), BC5CDR (Li et al., 2016), COARSE_ANYTHINGNER (Luo et al., 2024), EMAILS_MIMGOLD (Ing'olfsd'ottir et al., 2020), HARVEYNER (Chen et al., 2022), RONEC (Dumitrescu and Avram, 2020), MBL_MIMGOLD (Ing'olfsd'ottir et al., 2020), PUD_UNER (Mayhew et al., 2024), GERMEVAL2014 (Benikova et al., 2014), LENERBR (Luz de Araujo et al., 2018), WIESP2022 (Grezes et al., 2022), BANGLABIOMED (Sazzed, 2022), AQMAR (Mohit et al., 2012), UGNAYAN_UNER (Mayhew et al., 2024), MUSICNER (Epure and Hennequin, 2023), WIKIGOLDSK (Suba et al., 2023), KAZNERD (Yeshpanov et al., 2022), SCIENCE_CROSSNER (Liu et al., 2020), E3CCORPUS (Magnini et al., 2020), SINER (Ali et al., 2020), GW_HERODOTOS (Erdmann et al., 2019), LAWS_MIMGOLD (Ing'olfsd'ottir et al., 2020), LETEMPS_HIPE2022 (Ehrmann et al., 2022), WIKI_BARNER (Peng et al., 2024), LITERATURE_CROSSNER (Liu et al., 2020), BENGALINER

(Rahman Rifat et al., 2019), AJMC_HIPE2022 (Ehrmann et al., 2022), PERSIANNER (Poostchi et al., 2016, 2018), FICTION_NYTK (Simon and Vad'asz, 2021), THAINNER (Buaphet et al., 2022), PLINYYOUNGER_HERODOTOS (Erdmann et al., 2019), BC2GMCORPUS (Smith et al., 2008), ADG_EVALITA2023 (Paccosi and Palmero Aprosio, 2022), SNK_UNER (Mayhew et al., 2024), HINERCOLLAPSED (Murthy et al., 2022), SONAR_HIPE2022 (Ehrmann et al., 2022), BIOSWEDISH (Simon Almgren, 2016), MASAKHANER20 (Adelani et al., 2022), GVYORUBA (Alabi et al., 2020), UDJAPANESEGSD (Omura and Asahara, 2018; Asahara et al., 2018), VKNER (Orasmaa et al., 2022), POLITICS_CROSSNER (Liu et al., 2020), TWEET_BARNER (Peng et al., 2024), IDNERNEWS2K (Khairunnisa et al., 2020), ARMTDP (Yavrumyan, 2020), NERTELUGU (Reddy et al., 2018), JANES-TAG.UD_JANES (Lenardi c et al., 2022), OVID_HERODOTOS (Erdmann et al., 2019)

- **RE**: LITERATURE_CROSSRE (Bassignana and Plank, 2022), SL1_HISTRED (Yang et al., 2023), RUSERRC (Bruches et al., 2020), DOCUMENT_ICORPUS (Shinohara et al., 2022), COFUN (Foroutan et al., 2024), CODEMIXEDMIXRED (Kong et al., 2024), SL2_HISTRED (Yang et al., 2023), MUSIC_CROSSRE (Bassignana and Plank, 2022), JACRED (Ma et al., 2024), SENTENCE_VLSP2020 (Nguyen et al., 2020; Nguyen and Man, 2020), BIORED (Luo et al., 2022), TFH (Chen et al., 2020), POPCORN (Giordano et al., 2024), MEN (Chanthran et al., 2024), SANWEN (Xu et al., 2017), RUREBUS (Ivanin et al., 2020; Artemova et al., 2020), ADEV2 (Gurulingappa et al., 2012), DOCUMENT_VLSP2020 (Nguyen et al., 2020; Nguyen and Man, 2020), SCIENCE_CROSSRE (Bassignana and Plank, 2022), PERLEX (Asgari-Bidhendi et al., 2020), NEWS_CROSSRE (Bassignana and Plank, 2022), CORPUSER (Collovini et al., 2016), POLITICS_CROSSRE (Bassignana and Plank, 2022), REDOCRED (Tan et al., 2022), ROR (Bou et al., 2020), NEREL (Loukachevitch et al., 2021), AI_CROSSRE (Bassignana and Plank, 2022), SEMEVAL2018 (G'abor et al., 2018), SCIERC (Luan et al., 2018), SEMEVAL2010 (Hendrickx et al., 2010), INSTRUCTIE (Gui et al., 2023), CONLL04 (Roth and Yih, 2002), SENTENCE_ICORPUS (Shinohara et al., 2022), SL0_HISTRED (Yang et al., 2023), RURED (Gordeev et al., 2020)

- **ED**: SENTENCE_DISASTER (Sahoo et al., 2020), LEVEN (Yao et al., 2022), SENTENCE_MUSIED (Xi et al., 2022), DOCUMENT_DEIE (Ren et al., 2024), EUSIE (Zubillaga et al., 2024), EVENTDNA (Colruyt et al., 2022), RAMS (Ebner et al., 2020), WIKIEVENTS (Li et al., 2021b), SENTENCE_DEIE (Ren et al., 2024), EVENTNET-ITA (Rovera, 2024), DOCUMENT_MUSIED (Xi et al., 2022), DOCUMENT_DISASTER (Sahoo et al., 2020), CMNEE (Zhu et al., 2024), LIFEEVENTDIALOG (Chen et al., 2023; Li et al., 2017), ACE2005 (Doddington et al., 2004), MINION (Pouran Ben Veyseh et al., 2022b), CASIE (Satyapanich et al., 2020), BKEE (Nguyen et al., 2024b), PHEE20 (Sun et al., 2024, 2022), MEE (Pouran Ben Veyseh et al., 2022a), CHED (Congcong et al., 2023), MAVEN (Wang et al., 2020)

- **EE**: SENTENCE_DISASTER (Sahoo et al., 2020), DOCUMENT_DEIE (Ren et al., 2024), EUSIE (Zubillaga et al., 2024), EVENTDNA (Colruyt et al., 2022), RAMS (Ebner et al., 2020), WIKIEVENTS (Li et al., 2021b), SENTENCE_DEIE (Ren et al., 2024), EVENTNET-ITA (Rovera, 2024), DOCUMENT_DISASTER (Sahoo et al., 2020), CMNEE (Zhu et al., 2024), LIFEEVENTDIALOG (Chen et al., 2023; Li et al., 2017), ACE2005 (Doddington et al., 2004), CASIE (Satyapanich et al., 2020), BKEE (Nguyen et al., 2024b), PHEE20 (Sun et al., 2024, 2022), MEE (Pouran Ben Veyseh et al., 2022a)

Note that a number of event datasets are used for both ED and EE experiments, thus appearing in the listing of both tasks.