RoRED: A Romanian Relation Extraction Dataset

George-Andrei Dima *, Ilie Cosmin Biltan, Luciana Morogan *

Posted Date: 11 May 2026

doi: 10.20944/preprints202605.0578.v1

Keywords: relation extraction; Romanian NLP; low-resource languages; dataset construction; distant supervision; machine translation

 doi:10.20944/preprints202605.0578.v1

Article

RoRED: A Romanian Relation Extraction Dataset

George-Andrei Dima [1,2,*], Ilie Cosmin Bilțan [2] and Luciana Morogan [2,*]

[1] Computer Science Department, Faculty of Automatic Control and Computer Science, Politehnica University of Bucharest, 060042 Bucharest, Romania
[2] Computer and Cyber Security Department, Faculty of Informations Systems and Cyber Security, Military Technical Academy 'Ferdinand I', 050141 Bucharest, Romania
* Correspondence: george_andrei.dima@stud.acs.upb.ro

Abstract

Relation extraction is an important task for structuring information from unstructured text. However, Romanian language still lacks dedicated datasets and benchmarks for this task. To address this gap, we introduce RoRED, a Romanian relation extraction dataset built by combining two complementary data construction strategies: translating existing high-quality English resources and applying distant supervision to native Romanian Wikipedia data. We leverage a powerful open-source large language model to automatically translate English examples into Romanian. For the native subset, we align Romanian Wikipedia entities with Wikidata relations to obtain naturally occurring Romanian examples. To better reflect real-world relation extraction scenarios, we also introduce synthetic negative examples generated using existing Romanian named entity recognition models. Finally, we validate the dataset by fine-tuning and evaluating multiple baseline models. Our strongest model, LUKE-RoRED, achieves a macro-F1 score of 0.8744 on the RoRED test set, demonstrating that the dataset can support relation extraction for Romanian. Overall, RoRED provides a strong first native benchmark for Romanian relation extraction.

Keywords: relation extraction; Romanian NLP; low-resource languages; dataset construction; distant supervision; machine translation

1. Introduction

Information extraction is one of the primary goals of natural language processing. Organizations have accumulated large amounts of data in natural language form and seek to process this information for various purposes, such as better understanding its structure, discovering new insights, and reorganizing content to enable faster search and retrieval. A crucial step in this process is identifying relations between entities mentioned in the text.

With the rise of generative large language models, the need for data mining techniques such as retrieval-augmented generation systems has emerged to overcome the limitations of LLM context windows. Relation extraction, together with named entity recognition, forms the foundation of more advanced RAG architectures, such as GraphRAG[1]. Consequently, further research in relation extraction remains important, as the task continues to play a significant role in modern NLP systems.

Relation extraction, as defined in TACRED [2], consists of predicting the relation between two entities in a sentence from a fixed set of predefined relations. There is a significant amount of resources for relation extraction, but they are primarily focused on English, where resources are abundant. Datasets such as NYT[3], TACRED[2], FewRel[4,5] or DocRED[6] power the development of the next generation of relation extraction models in English. At present, there is no dataset designed to support relation extraction research in Romanian.

To bridge this gap, we introduce RoRED, a Romanian relation extraction dataset built from two sources: a machine-translated version of the open-source English dataset FewRel and a native

 doi:10.20944/preprints202605.0578.v1

Romanian dataset constructed from Wikipedia relations following the methodology proposed by FewRel, therefore capturing Romanian cultural details.

We explored multiple model architectures for training a Romanian relation extraction model. We identified LUKE [7,8], which introduces entity-aware embeddings, as a strong baseline. We also explored more recent large language models trained on larger corpora to test the assumption that increased training data improves the ability to understand and perform Romanian language tasks. Finally, we provide two baseline models for Romanian relation extraction, LUKE-RoRED and Qwen-RoRED, and evaluate them through a detailed analysis of their results.

Our main contributions are:

- We introduce the RoRED dataset[1], the first relation extraction corpus for Romanian that includes native Romanian samples.
- We demonstrate the usability of the dataset by providing two baseline models for Romanian relation extraction: LUKE-RoRED [2] and Qwen-RoRED.
- We extensively evaluate the models and show that, with the help of generative AI, existing NLP resources can be extended to low-resource languages.

The rest of the paper is organized as follows. Section 2 reviews related work on relation extraction datasets and models. Section 3 describes the construction of the RoRED dataset, including data sources, processing, and filtering steps. Section 4 presents the baseline models and experimental setup, followed by the results and analysis. Finally, Section 5 concludes the paper and outlines directions for future work.

2. Related Work

Relation Extraction is a continuously evolving task, with datasets differing along several key directions. These include the set of relation labels, ranging from broad categories to highly fine grained relation labels, as well as language coverage, with some datasets focusing primarily on English, others on low resource languages[9], and some designed to be multilingual[10]. Datasets also vary in domain, spanning from open domain settings to highly specialized domains[9,11], and in context size, from single sentences to full documents, as in DocRED[6]. Additionally, datasets differ in their intended use. Some are designed for large scale supervised training, such as TACRED[2], while others target few shot learning scenarios, such as FewRel[4,5]. There are also datasets commonly used for evaluation and comparison purposes, such as RELX[12].

A clear trend in the field is the expansion of relation label inventories, together with a shift toward more granular and schema-driven relations. Early benchmarks such as CoNLL-2004[13] defined only five core relation types: Located-In, Work-For, OrgBased-In, Live-In, and Kill. The dataset consists of sentences extracted from English news articles. The NYT dataset later expanded this inventory substantially through distant supervision, and the commonly used NYT10[3] version has 52 Freebase-derived relations plus an NA label. Example relations from the NYT dataset include nationality, contains, and place_of_birth. TACRED is defining 41 TAC-KBP canonical relation types, together with a no_relation label, including relations such as per:schools_attended and org:members. Finally, FewRel expanded the space to 100 relations, constructed from Wikipedia sentences aligned to Wikidata and then filtered by crowdworkers. Its label set is significantly more fine-grained, including relations such as "member of sports team" or "place served by transport hub".

There have also been significant efforts to expand relation extraction resources beyond English. Many recent datasets adopt a multilingual perspective, while models built on top of them aim to address limited data availability through cross-lingual transfer learning. For instance, ACE 2005[10] includes Chinese and Arabic in addition to English. More recently, MultiTACRED (2023)[14] extends TACRED to 12 languages, including English, Arabic, German, Spanish, French, Finnish, Hindi, Hun-

[1] https://huggingface.co/datasets/andreidima/RoRED-v1
[2] https://huggingface.co/andreidima/mluke-large-relation-extraction-romanian

 doi:10.20944/preprints202605.0578.v1

garian, Japanese, Polish, Russian, Turkish, and Chinese, with approximately 106K translated sentences per language. Similarly, RELX[12], introduced in 2020, provides a cross-lingual relation classification benchmark covering English, French, German, Spanish, and Turkish. RED^{FM}[15] is another multilingual resource for relation extraction. It is built from Wikipedia and Wikidata and introduces two resources: RED^{FM}, a human-revised dataset covering 32 relation types and seven languages, and SRED^{FM}, a larger silver-standard dataset covering 18 languages, 400 relation types, 13 entity types, and more than 40 million triplet instances.

Romanian is represented in MULTI-CROSSRE [16], a multilingual and multi-domain relation extraction dataset obtained by machine-translating the English CrossRE dataset [17] into 26 additional languages. The Romanian portion is therefore a translated subset. In contrast, RoRED is designed specifically for Romanian and adds native Romanian Wikipedia examples obtained through distant supervision.

With respect to algorithms, existing work ranges from cross-lingual approaches, which adapt the input representation while keeping the same English-trained model [18], to multilingual versions of validated English architectures. One such model is mLUKE, the multilingual version of LUKE[7], which uses entity-aware self-attention and is therefore suitable for relation extraction, where the model must reason over a marked pair of entities. Another relevant model is mREBEL[15,19], a multilingual version of REBEL that addresses relation extraction as a sequence-to-sequence task, generating relational triplets directly from the input text.

3. RoRED Dataset

To build a Romanian relation extraction dataset, we selected FewRel[4,5] as the source resource for adaptation into Romanian. Our choice was motivated by its high quality, large number of instances, broad relation coverage, and permissive license, in contrast to datasets such as TACRED[2]. Beyond simple translation, FewRel also provides a methodology for generating new examples. We leveraged the authors' approach to create additional samples from Romanian Wikipedia, thereby grounding the dataset in authentic Romanian text, entities, and facts, and incorporating a stronger cultural component.

RoRED samples follow the same structure as FewRel. FewRel samples contain the following fields: **relation**, **tokens**, **head**, **tail**, and **names**. The fields **relation** and **names** contain information about the relation represented in the dataset sample: the former stores the Wikidata property ID of the relation, while the latter contains two strings, a short description of the relation and a full definition, both retrieved from Wikidata. The text is already tokenized and stored in the **tokens** field, which is later referenced by the relation components. The fields **head** and **tail** describe the subject and object entities of the relation, respectively. Both fields share the same structure, with the following sub-components: **text**, a lowercased substring representing the entity mention; **type**, the corresponding Wikidata entity ID; and **indices**, a list of token span indices mapped to the entity.

In addition, each RoRED example contains a **source** field. This field indicates whether the example comes from the translated FewRel subset, the native Romanian Wikipedia subset, or the synthetically generated negative subset. We keep this information explicitly in the dataset to support source-level analysis, enable controlled experiments on translated versus native data, and facilitate future research that may rely on one source or another.

3.1. Adapting FewRel to Romanian

As discussed in the related work section of this paper, powerful large language models have demonstrated strong capabilities for dataset creation and enhancement, being particularly proficient in translation tasks. Our infrastructure allowed us to use `gpt-oss:120b` [20] to automatically translate the FewRel samples from English into Romanian. The model was hosted on an Ollama[3] server running

[3] https://ollama.com/

 doi:10.20944/preprints202605.0578.v1

on a machine equipped with NVIDIA A100 80GB GPUs. The translation was performed sequentially, and the entire process took approximately 350 hours.

The main challenge in translating a relation extraction dataset is preserving the relation annotations after translation, meaning being able to recompute the new indices of the entities in the translated text. We tested two approaches: marker-based translation and metadata-based translation, as shown in Table 1.

Table 1. Prompt approaches for preserving entity annotations.

Approach	Sample added in prompt
Marker-based	Radiant Baby is a musical about <head>Keith Haring</head>, who was an artist and social activist in <tail>New York City</tail>.
Metadata-based	{"text": "Radiant Baby is a musical about Keith Haring, who was an artist and social activist in New York City.", "head": "Keith Haring", "tail": "New York City"}

In the marker-based translation, we inserted head and tail markers around the corresponding entities in the source text and instructed the model to keep these markers in the translation. In the metadata-based translation, we created a JSON object containing the text and appended auxiliary information about the entities on separate fields next to the sentence to be translated, and requested the model to update those fields with the corresponding entities from the translated text.

We conducted a preliminary evaluation of both prompt formats on 100 examples and found that the marker-based format outperforms the metadata-based one. We observed that the metadata-based approach tends to preserve entities in their original language, even when Romanian translations are available. Additionally, in some cases, the generated entities could not be reliably matched to the text using simple string-based methods. Since the marker-based format produced fewer errors overall, we selected it for translating the dataset.

With regards to prompting techniques, we employed a one-shot approach to increase the likelihood of obtaining correct answers from the LLM. Since GPT-OSS is an instruction-tuned model, we used the chat format. The final prompt consisted of: (1) a system prompt containing Romanian instructions for translating the text and returning the relation annotations; (2) an example user prompt with an English text with head and tail markers sample extracted from the dataset; (3) an example assistant response with the correctly translated text from step 2; and (4) a user prompt with the dataset sample to be processed by the model. Prompts are available in Appendix A.1.

After retrieving the model's answer, we programmatically construct samples that follow the FewRel format. We split the translated text into tokens and then identify and compute the corresponding indices for the two entities. We keep the relation type, relation descriptions, and entity types from the original sample.

3.2. Distant Supervision from Romanian Wikipedia

We constructed the raw Romanian Wikipedia relation extraction dataset by aligning Romanian Wikipedia mentions with Wikidata entities and then projecting Wikidata relations onto co-occurring entity mentions. We used the Romanian Wikipedia articles dump dated January 1, 2026, which was the latest available Romanian Wikipedia articles dump at the time of collection. First, we downloaded the Romanian Wikipedia article dump together with the Romanian Wikipedia page properties dump and the Wikidata truthy relations dump. The truthy dump provides a simplified view of Wikidata that retains the best-ranked non-deprecated statements for each item-property pair, making it suitable for extracting high-confidence entity-relation-entity triples. This allowed us to extract high-confidence (subject, property, object) triples and project them onto co-occurring Wikidata-linked entity mentions in Romanian Wikipedia sentences.

We used the page properties dump to align Romanian Wikipedia pages and titles with Wikidata QIDs. We then parsed the article dump, resolved internal links to Wikidata entities, segmented and tokenized the text with the Romanian Stanza[21] pipeline, and retained sentences containing at least

 doi:10.20944/preprints202605.0578.v1

two resolved entity mentions. For each sentence, we generated all ordered mention pairs and searched the Wikidata relations for matching triples. Each matching property yielded one labeled relation extraction instance, resulting in more than 270,000 automatically labeled examples.

3.3. Post-Processing and Filtering

We applied a supplementary step to check whether the relation could be inferred from the context in the Romanian Wikipedia subset. For each candidate example, we reconstructed the sentence from the token sequence and provided the sentence, head entity, tail entity, and relation name to the self-hosted LLM. The model was prompted to decide whether the relation between the two entities could be inferred from the sentence and to answer strictly with *Yes* or *No*. The prompt used is available in Appendix A.2. We retained only examples for which the relation was marked as inferable, while all others were removed from the final dataset. This filtering step was intended to reduce cases where two entities are related in Wikidata, but the relation is not explicitly expressed in the Romanian sentence.

The raw Romanian Wikipedia dataset required an additional cleaning stage because the automatic extraction process sometimes preserved residual Wikipedia markup. The raw texts were parsed using `mwparserfromhell` [4], and most wiki templates were successfully discarded. We also applied a set of fixed rules to remove leftover wiki artifacts, such as residual image/file syntax, category lines, bullet markers, and empty parentheses. After cleaning, the text was retokenized, and the head and tail entity indices were recomputed by matching the entity text against the cleaned token sequence. Rather than remapping indices after each individual cleaning step, we realigned the entity mentions in the final cleaned sequence. Examples were discarded if they contained non-article or noisy markers, had fewer than 5 or more than 80 tokens, retained leftover markup, or if either the head or tail entity could no longer be found after cleaning.

To prevent highly frequent relations from disproportionately influencing the training process, we applied undersampling after merging the cleaned Romanian Wikipedia examples with the translated FewRel examples. Specifically, we removed relations with fewer than 50 samples and capped each remaining relation at a maximum of 1,400 examples. This threshold was chosen because the translated FewRel subset contains up to 700 examples per relation; therefore, a cap of 1,400 allows us to preserve as many examples as possible while still allowing both sources to contribute equally to the same relation. When a relation appeared in both sources, we split the sample allocation equally between them whenever possible. If one source contained fewer than 700 examples, the remaining allocation was assigned to the other source.

3.4. Negative Samples Generation

To make the dataset more suitable for training a real-world relation extraction model, we synthetically added negative examples labeled as *no_relation*. In practical settings, not every pair of entities appearing in the same sentence expresses one of the target relations. Therefore, the model must learn not only to classify known relation types, but also to recognize cases where no relevant relation is expressed between two entity mentions.

Negative samples were generated automatically from the processed positive dataset. We set the number of generated negative examples equal to the number of positive examples, aiming for a balanced 1:1 distribution between positive and negative instances. For each negative example, a source sentence was selected randomly, with replacement. The sentence was reconstructed from its token sequence and processed with the Romanian spaCy[22] model `ro_core_news_lg`[5] to identify named entity mentions. The detected entities were then matched back to the original tokenized sentence in order to obtain token-level spans compatible with the existing dataset format. For each selected sentence, two distinct and non-overlapping entity mentions were chosen from the detected mentions. Whenever possible, we excluded the original head and tail entities from the candidate pool, reducing

[4] `mwparserfromhell` is a Python parser for MediaWiki markup: https://github.com/earwig/mwparserfromhell.
[5] https://spacy.io/models/ro#ro_core_news_lg

 doi:10.20944/preprints202605.0578.v1

the likelihood of reusing the annotated positive relation pair. The resulting entity pair was stored using the same structure as the positive examples, but with the relation label set to *no_relation*. Because source sentences were sampled with replacement and no explicit deduplication was applied, the same sentence or entity pair may occur in multiple generated negative examples.

3.5. Dataset Statistics

We present the main statistics of the RoRED dataset from three perspectives: split-level composition, source distribution, and relation coverage. This organization facilitates the identification of potential biases in the dataset.

Table 2 presents the overall composition of the dataset across the train and test splits, grouped by source. RoRED contains 146 190 examples in total, split into 116 918 training examples and 29 272 test examples. The dataset includes 101 positive Wikidata relations and one additional `no_relation` label. Overall, it is exactly balanced between 73 095 positive examples and 73 095 synthetic negative examples. In the training split, there are 58 442 positive examples and 58 476 negative examples, while the test split contains 14 653 positive examples and 14 619 negative examples. The mean sentence length is 27.46 tokens in the train split and 27.44 tokens in the test split, suggesting that the random split does not introduce a noticeable distribution shift.

Table 2. Source distribution across train and test splits.

Source	Train		Test		Total	
	Examples	%	Examples	%	Examples	%
`ro_wikipedia`	13 687	11.71	3 431	11.72	17 118	11.71
`translated`	44 755	38.28	11 222	38.34	55 977	38.29
`synthetic_negative`	58 476	50.01	14 619	49.94	73 095	50.00
Total	116 918	100.00	29 272	100.00	146 190	100.00

We further group examples by their source: translated FewRel examples, Romanian Wikipedia examples, and synthetic negative examples. The positive examples come from two sources: 55 977 translated FewRel examples and 17 118 native Romanian Wikipedia examples. The relative contribution of each source is nearly identical in the train and test splits, showing that evaluation is not biased toward a different source composition than the one observed during training.

The positive examples are distributed over 101 relation labels, for a total of 73 095 positive instances. Although all positive relations are represented in both train and test, their frequency is not uniform: the smallest relation contains 51 examples, while the largest reaches the undersampling cap of 1 400 examples. On average, each relation contains 723.71 examples. Table 3 shows the positive relation coverage separately for each source. The translated subset largely preserves the structure of FewRel with 80 relations, where each relation is represented by up to 700 examples. The small deviations from this upper bound are caused by examples removed during translation validation and filtering. This explains the high median support of the translated subset. In contrast, the `ro_wikipedia` subset is collected from naturally occurring Romanian text and is therefore less uniform. It covers more relation types, but the number of examples per relation varies substantially.

Table 3. Relation coverage by source.

Source	Relations	Examples	Max/rel.	Mean/rel.	Median/rel.
`ro_wikipedia`	89	17 118	1 400	192.34	85.00
`translated`	80	55 977	700	699.71	700.00

Table 4 lists the 20 most frequent relations in RoRED, sorted by the number of examples. The most frequent relations are dominated by geographic, administrative, compositional, sequential, type/class, biographical, creative, and organizational relations, which are common in Wikidata-derived data.

 doi:10.20944/preprints202605.0578.v1

Table 4. Top 20 relations by number of examples.

Relation	Examples	Label
P150	1 400	contains administrative territorial entity
P131	1 400	located in the administrative territorial entity
P17	1 400	country
P6885	1 400	historical region
P155	1 400	follows
P156	1 400	followed by
P361	1 400	part of
P527	1 209	has part
P276	1 190	location
P31	1 184	instance of
P706	1 181	located on terrain feature
P206	1 153	located in or next to body of water
P1001	1 035	applies to jurisdiction
P800	961	notable work
P26	931	spouse
P3373	916	sibling
P40	889	child
P27	875	country of citizenship
P159	854	headquarters location
P463	843	member of

The relation overlap between sources is shown in Figure 1. The two positive sources share 68 relations, while each source also contributes relations that are not present in the other: 21 relations appear only in `ro_wikipedia`, and 12 appear only in the translated subset. Romanian Wikipedia adds new relations to the FewRel dataset, bringing the total positive relation coverage to 101. Figure 2 shows the distribution of positive examples across relation labels, with bars stacked by source. The highest-frequency relations reach the undersampling limit of 1 400 examples, and several of them combine examples from both `ro_wikipedia` and the translated subset. A large middle segment is dominated by translated examples, reflecting the balanced structure inherited from FewRel, where relations are represented by up to 700 examples. In contrast, the lower-frequency tail is mostly composed of `ro_wikipedia` relations, showing that the native Romanian source contributes additional relation types but with more variable support. Overall, the figure highlights the complementary role of the two positive sources: translated data provides dense and stable supervision, while Romanian Wikipedia increases relation diversity and adds naturally occurring Romanian contexts.

 doi:10.20944/preprints202605.0578.v1

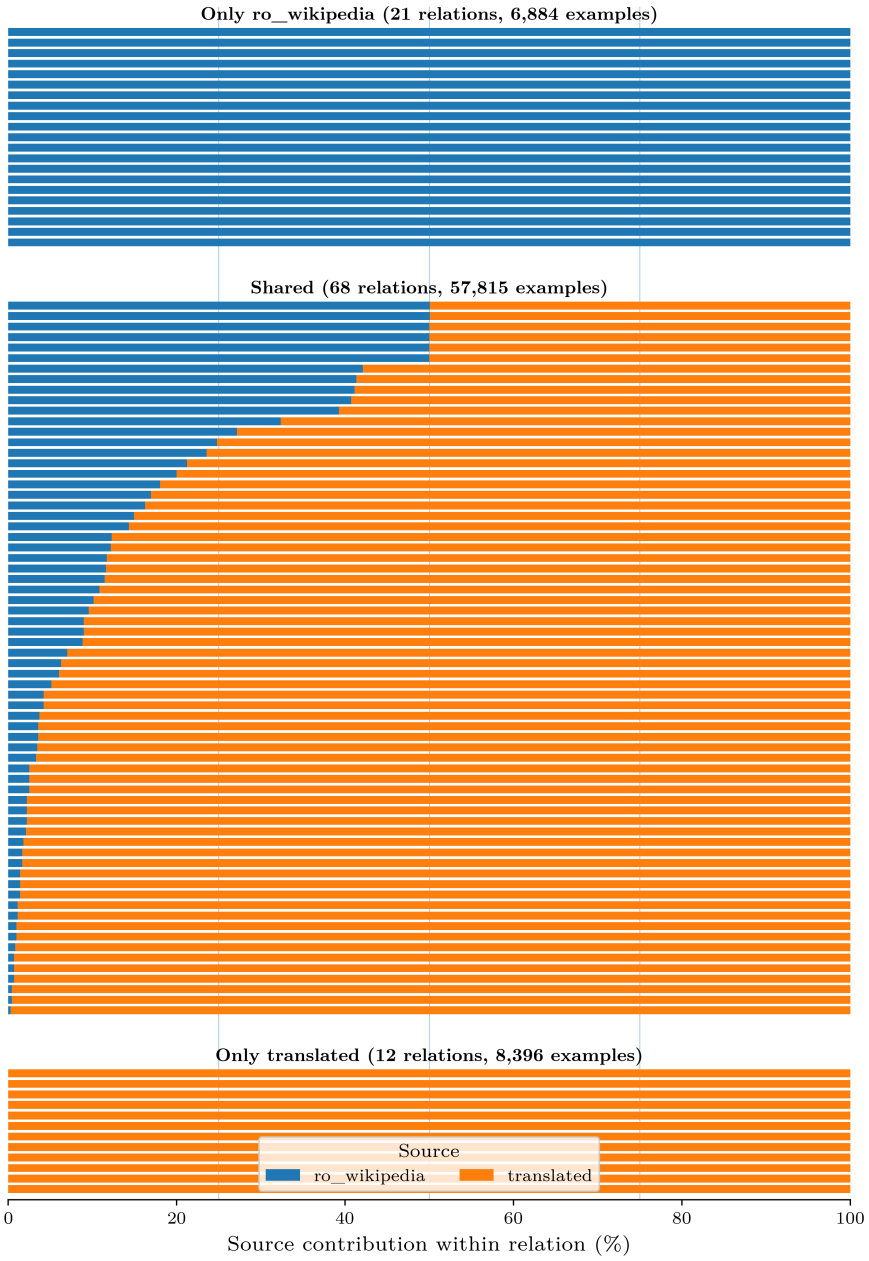

Figure 1. Relation overlap between the data sources.

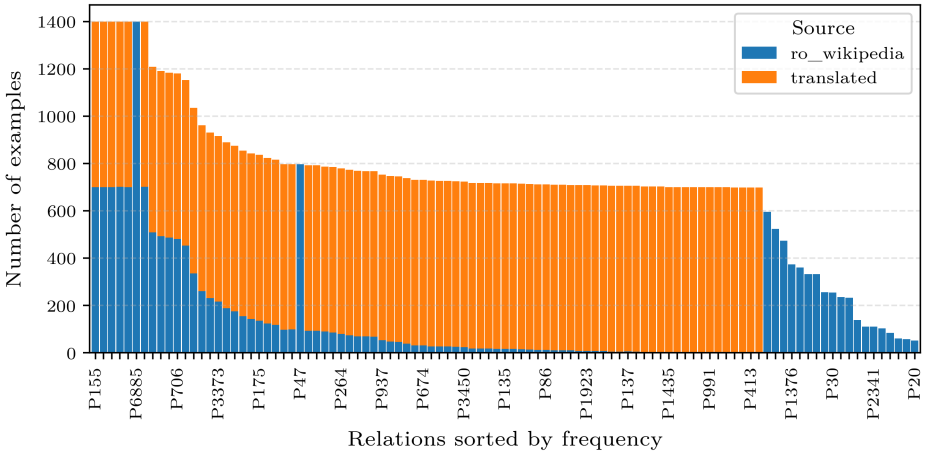

Figure 2. Sample distribution across relations.

 doi:10.20944/preprints202605.0578.v1

We also report token-level statistics in Table 5. The Romanian Wikipedia examples are slightly longer on average than the translated examples, while synthetic negatives remain close to the global average, expected since they were generated from the combined dataset. Entity spans are short, with heads and tails usually consisting of one or two tokens.

Table 5. Sentence and entity-span length statistics.

Source	Examples	Mean tokens	Median tokens
`ro_wikipedia`	17 118	31.48	29.00
`translated`	55 977	25.97	26.00
`synthetic_negative`	73 095	27.66	27.00

Entity span	Mean tokens	Median tokens
Head entity	1.96	2.00
Tail entity	1.82	1.00

3.6. Dataset Limitations

Since we used automatic translation, the dataset is prone to the usual errors that occur in such processes. We carefully analyzed randomly selected samples for translation errors and illustrated representative examples in Table 6. We found that the most common issues are word order errors, as Romanian word order, being a Romance language, can differ significantly from English. Less common errors include grammatical issues in Romanian, such as missing definite articles (e.g., "cosmonaut" → "cosmonautul"), and cases where certain entities were not translated and remained in English, even though Romanian equivalents exist (e.g., "Petrine Sees" → "Scaune Petrine"). None of the errors we encountered significantly impacted the meaning of the sentences, and, at least in the examined cases, the relation between the two entities remained intact and could still be inferred by Romanian native speakers. Therefore, we concluded that the machine translated dataset is adequate for training relation extraction models and thus suitable for the purpose of this paper.

Table 6. Translation errors

Error	**Original translation**	**Correct translation**
Word order	Dušan Trančík (născut pe 26 noiembrie 1946) este un **slovac regizor de film și scenarist**.	Dušan Trančík (născut pe 26 noiembrie 1946) este un **regizor de film și scenarist slovac**.
Missing definite article	Pe 20 august 2012, Malenchenko, împreună cu **cosmonaut** Gennady Padalka, a participat la a cincea sa plimbare spațială din carieră.	Pe 20 august 2012, Malenchenko, împreună cu **cosmonautul** Gennady Padalka, a participat la a cincea sa plimbare spațială din carieră.
Incomplete translation	**Antioch**, Alexandria și Roma au fost pretinse ca având o origine legată de **Peter**, de unde provine termenul „**Petrine Sees**”.	**Antiohia**, Alexandria și Roma au fost pretinse ca având o origine legată de **Petru**, de unde provine termenul „**Scaune Petrine**”.

The Romanian Wikipedia subset can also introduce some limitations. Since it was constructed through distant supervision, its quality depends on the accuracy of entity linking and Wikidata coverage. Errors may arise from ambiguous Wikipedia links, incomplete or outdated Wikidata statements, or sentences that contain several entities with multiple plausible relations. Moreover, the relation distribution shows that Romanian Wikipedia is not uniformly distributed across domains, with geographic, administrative, historical, and biographical information being better represented than other categories. As a result, the native subset improves linguistic authenticity, but it may also introduce biases specific to Romanian Wikipedia.

Negative examples were generated synthetically using named entities detected by the Romanian spaCy model. Consequently, the negative class may be biased toward the entity boundaries and entity

 doi:10.20944/preprints202605.0578.v1

types recognized by this particular NER system. Another limitation is that, since negative examples are automatically created from sentences that already contain positive relation instances, some generated negative pairs may still express valid relations. Such cases would be incorrectly labeled as false negatives.

4. Experiments and Results

4.1. Baseline Model

To provide an initial validation of RoRED and establish strong baselines for Romanian relation extraction, we selected two candidate models for fine-tuning: mLUKE-large[7,8] and Qwen3-0.6B[23]. Although we considered a broader set of architectures, exhaustively evaluating all alternatives was outside the practical constraints of this study. We therefore selected two models that we considered representative, architecturally complementary, and promising for Romanian. We selected mLUKE-large as a non-generative baseline because LUKE-style models use entity-aware self-attention, which is well aligned with relation extraction. In contrast, Qwen3-0.6B was selected as a compact instruction-tuned generative baseline. The Qwen3 family reports strong multilingual and instruction-following capabilities, while the 0.6B variant remains small enough for efficient supervised fine-tuning.

For mLUKE, we used the multilingual checkpoint `studio-ousia/mluke-large`. Although this checkpoint does not explicitly support Romanian, we selected it to assess whether its multilingual pretraining and entity-aware architecture could transfer effectively to Romanian relation extraction. We refer to the resulting model as LUKE-RoRED.

For training LUKE-RoRED, each RoRED example was converted into the input format required by LUKE. The sentence was reconstructed from the token sequence, while the head and tail annotations were mapped to character-level entity spans. These spans were then passed to the LUKE tokenizer, which jointly encodes text tokens and entity tokens. The model was trained to predict one of the 101 positive relation labels plus the additional `no_relation` label. We fine-tuned it for three epochs using a learning rate of 2×10^{-5}, weight decay of 0.01, maximum sequence length of 256 tokens, and batch size of 32. The training process took approximately 27 minutes on an A100 GPU.

We also trained a generative baseline based on Qwen, in order to evaluate whether a compact instruction-tuned language model can learn the RoRED relation extraction task. We used `unsloth/Qwen3-0.6B`, a 600M parameter model, fine-tuned with LoRA for parameter-efficient supervised fine-tuning. In contrast to LUKE, which treats relation extraction as entity-pair classification, Qwen was trained in an instruction-following format. Each example was converted into a chat-style prompt, where the model receives a Romanian sentence with marked head and tail entities and generates the corresponding relation label. This format allows the model to learn the task as a constrained label generation problem. We fine-tuned Qwen for three epochs using a learning rate of 2×10^{-4}, weight decay of 0.01, maximum sequence length of 2048 tokens, and batch size of 64. LoRA adapters were applied to the attention and MLP projection layers, with rank 64, alpha 64, and no dropout.

4.2. Results

We evaluate the proposed baselines on the RoRED test set using accuracy and macro-F1. Table 7 shows that LUKE-RoRED obtains the best performance on both metrics, with a noticeable advantage over the Qwen-based generative baseline. The difference is especially visible in accuracy, suggesting that the entity-aware classification architecture of LUKE-RoRED is better suited to the current RoRED setup and we therefore decided to use LUKE-RoRED for the more detailed source-level and relation-level analysis.

Table 7. Overall performance on the RoRED test set.

Model	Accuracy	Macro-F1
LUKE-RoRED	**0.9187**	**0.8744**
Qwen-RoRED	0.8759	0.8454

 doi:10.20944/preprints202605.0578.v1

Table 8 reports LUKE-RoRED performance separately for the two positive sources, their aggregate, and the negative class. The model obtains similar accuracy on native Romanian Wikipedia examples and translated examples, with a small advantage on the native subset. At the aggregate level, positive-only performance remains strong, reaching 0.8843 accuracy and 0.8842 macro-F1. The negative subset is classified with higher accuracy, showing that LUKE-RoRED can reliably identify many `no_relation` instances.

Table 8. LUKE-RoRED performance on positive sources and negative examples in the RoRED test set.

Subset	Examples	Accuracy	Macro-F1
Positive – `ro_wikipedia`	3 431	0.8913	0.8102
Positive – `translated`	11 222	0.8821	0.7875
Positive – total	14 653	0.8843	0.8842
Negative – `no_relation`	14 619	0.9533	–

We further analyze the relation-level performance of LUKE-RoRED. Table 9 reports both the highest and lowest scoring relations according to F1. The best-performing relations include several labels with only moderate support, showing that some relation types can be learned reliably even with fewer samples. In contrast, the weakest relations include semantically broad or potentially overlapping labels, such as administrative, geographic, ownership, and creative-role relations. This indicates that the remaining errors are not caused only by limited support, but also by relation ambiguity and overlap between related Wikidata properties.

Table 9. Highest and lowest LUKE-RoRED relation-level results on the RoRED test set, ranked by F1.

Relation	Name	Train	Test	Precision	Recall	F1
Highest F1 relations						
P194	legislative body	88	23	1.0000	1.0000	1.0000
P105	taxon rank	596	149	0.9933	0.9933	0.9933
P413	position played on team / speciality	559	140	0.9929	0.9929	0.9929
P2670	has parts of the class	476	119	1.0000	0.9832	0.9915
P59	constellation	581	146	0.9931	0.9863	0.9897
P102	member of political party	580	146	0.9931	0.9795	0.9862
P2094	competition class	560	140	0.9928	0.9786	0.9856
P1435	heritage designation	560	140	0.9720	0.9929	0.9823
P410	military rank	560	140	0.9720	0.9929	0.9823
P412	voice type	560	140	0.9854	0.9643	0.9747
Lowest F1 relations						
P138	named after	82	21	0.7000	0.3333	0.4516
P1366	replaced by	46	12	0.5385	0.5833	0.5600
P58	screenwriter	564	141	0.7258	0.6383	0.6792
P127	owned by	580	145	0.7313	0.6759	0.7025
P36	capital	288	72	0.8333	0.6250	0.7143
P706	located on terrain feature	944	237	0.7004	0.7300	0.7149
P551	residence	565	142	0.7183	0.7183	0.7183
P17	country	1120	280	0.7246	0.7143	0.7194
P131	located in the administrative territorial entity	1120	280	0.7143	0.7321	0.7231
P20	place of death	40	11	0.5789	1.0000	0.7333

Table 10 summarizes the dominant confusion pattern for each of the weakest LUKE-RoRED relations. For each gold relation, the table reports its test support, the overall error rate, and the most frequent incorrect prediction together with its share among that relation's errors. The errors are not uniformly distributed: in several cases, a single predicted relation accounts for more than half of all errors. For instance, P58 (*screenwriter*) is most often predicted as P1877 (*after a work by*), while P551 (*residence*) is frequently confused with P937 (*work location*). These cases suggest that the model often identifies the correct broad semantic domain, but fails to distinguish between finer-grained Wikidata properties.

 doi:10.20944/preprints202605.0578.v1

Table 10. Main confusion patterns for the lowest-performing LUKE-RoRED relations.

Gold relation	Test	Error rate	Main predicted relation	Share of errors
P138 named after	20	65.00%	P50 author	61.54%
P36 capital	70	38.57%	P1383 contains settlement	70.37%
P1366 replaced by	11	36.36%	P156 followed by	50.00%
P58 screenwriter	134	35.82%	P1877 after a work by	64.58%
P127 owned by	142	33.10%	P137 operator	29.79%
P551 residence	132	28.79%	P937 work location	60.53%
P17 country	270	28.52%	P131 located in admin. entity	32.47%
P131 located in admin. entity	261	26.44%	no_relation	30.43%
P706 located on terrain feature	225	26.67%	no_relation	26.67%

Geographic and administrative relations remain an important source of errors. Relations such as P17 (*country*), P131 (*located in administrative entity*), P36 (*capital*), and P706 (*located on terrain feature*) are mostly confused with other location-related labels or predicted as no_relation. This reflects the high semantic overlap between geographic relations, where the distinction often depends on the specific role of the entity pair rather than on surface lexical cues alone.

The support values also show that some relation-level results should be interpreted with caution. For example, P1366 has only 11 test examples, so a small number of errors has a large effect on its score.

Overall, the results establish LUKE-RoRED as the strongest baseline on RoRED, while the competitive Qwen-RoRED scores show that the dataset can also support instruction-style relation classification. The remaining errors are concentrated in a limited set of relation types, suggesting that future work should focus on disambiguating closely related Wikidata properties and improving performance on low-support or semantically broad relations.

5. Conclusions

In this paper, we introduced RoRED, a Romanian relation extraction dataset designed to support supervised relation classification in a low-resource language setting. The dataset combines automatically translated FewRel examples with native Romanian examples extracted from Romanian Wikipedia through distant supervision. This construction strategy allows RoRED to benefit from the consistent number of samples provided by FewRel, while also adding naturally occurring Romanian contexts and entities.

We provide an initial validation of the dataset by training two baseline models: LUKE-RoRED, based on the multilingual LUKE architecture, and Qwen-RoRED, based on an instruction-tuned generative model. The experimental results show that both models can learn the proposed task, with LUKE-RoRED achieving the strongest performance and reaching a macro-F1 score of 0.8744 on the RoRED test set.

Further source-level analysis shows that LUKE-RoRED performs consistently on both translated and native Romanian examples, suggesting no major source bias. At the same time, the relation-level analysis highlights that the remaining errors are concentrated around semantically close Wikidata properties and low-support relations. This indicates that errors are caused both by data scarcity and by the overlapping nature of the relation set.

 doi:10.20944/preprints202605.0578.v1

RoRED still has several limitations due to its construction, such as occasional translation artifacts and noise introduced by distant supervision. Nevertheless, our analysis suggests that these limitations do not prevent the dataset from being useful for training and evaluating Romanian relation extraction models.

Overall, RoRED represents a strong first native resource for Romanian relation extraction and provides a foundation for future work on structured information extraction in Romanian.

In future work, we plan to further improve RoRED by refining the generation of negative samples, evaluating the dataset in scenarios closer to real-world use, and improving the taxonomy of the label set.

Author Contributions: Conceptualization, G.A.D.; methodology, G.A.D.; software, G.A.D. and I.C.B.; validation, G.A.D. and I.C.B.; formal analysis, G.A.D.; investigation, G.A.D.; resources, G.A.D. and L.M.; data curation, G.A.D. and I.C.B.; writing—original draft preparation, G.A.D.; writing—review and editing, I.C.B. and L.M.; visualization, G.A.D.; supervision, L.M.; project administration, L.M.; funding acquisition, L.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Romanian Executive Unit for Financing Higher Education, Research, Development and Innovation (UEFISCDI), grant number PN-IV-P6-6.3-SOL-2024-0090. The article processing charges (APC) were funded by the UEFISCDI grant awarded under the 7Sol(T7)/2024 contract.

Data Availability Statement: The RoRED dataset is publicly available at https://huggingface.co/datasets/andreidima/RoRED-v1.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A. Prompts

Appendix A.1. Prompt Template for FewRel Translation

```
SYSTEM:
You are an agent that translates relation extraction examples into Romanian.
IMPORTANT:
1) Keep the <head></head> and <tail></tail> markers EXACTLY as they are
   (do not translate them, delete them, or move them).
2) Translate all text into Romanian, but keep the content between the markers
   as coherent entities.
3) Respond ONLY with a valid JSON object in the format
   {"text": "the translated sentence, which still includes the markers"}

USER:
{"text":"It starts from <head>Taunsa Barrage</head> at <tail>Indus River</tail> ."}

ASSISTANT:
{"text":"Pornește de la <head>Barajul Taunsa</head> pe <tail>râul Indus</tail>."}

USER:
{"text":"<dataset sample to be translated here>"}
```

Appendix A.2. Relation Filtering Prompt

```
Consider the following entities and the sentence in Romanian language.
Tell me whether the relationship between the head and the tail can be inferred
from the sentence. The answer must be STRICTLY Yes or No.

Sentence: {sentence}
Head: {head}
```

 doi:10.20944/preprints202605.0578.v1


```
Tail: {tail}
Relation (name): {relation_name}
```

Appendix A.3. Prompt Template for Qwen Training

```
SYSTEM:
You are an agent that predicts the semantic relation between the <head> entity
and the <tail> entity in the text provided by the user.
Choose exactly one of the following relation options:
no_relation, screenwriter, nominated_for, developer, director, distributor,
sibling, residence, participant_of, manufacturer, father, owned_by,
headquarters_location, occupant, member_of_political_party, occupation,
country_of_citizenship, participant, location, field_of_work, operator,
instance_of, religion, military_branch, ...(full list of labels)

USER:
<sentence with the target entities marked using <head></head> and <tail></tail>>
```

References

1. Edge, D.; Trinh, H.; Cheng, N.; Bradley, J.; Chao, A.; Mody, A.N.; Truitt, S.; Larson, J. From Local to Global: A Graph RAG Approach to Query-Focused Summarization. *ArXiv* **2024**, *abs/2404.16130*.
2. Zhang, Y.; Zhong, V.; Chen, D.; Angeli, G.; Manning, C.D. Position-aware Attention and Supervised Data Improve Slot Filling. In Proceedings of the Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP 2017), 2017, pp. 35–45.
3. Riedel, S.; Yao, L.; McCallum, A. Modeling Relations and Their Mentions without Labeled Text. In Proceedings of the Machine Learning and Knowledge Discovery in Databases; Balcázar, J.L.; Bonchi, F.; Gionis, A.; Sebag, M., Eds., Berlin, Heidelberg, 2010; pp. 148–163.
4. Han, X.; Zhu, H.; Yu, P.; Wang, Z.; Yao, Y.; Liu, Z.; Sun, M. FewRel: A Large-Scale Supervised Few-Shot Relation Classification Dataset with State-of-the-Art Evaluation. In Proceedings of the Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, - 2018; pp. 4803–4809. https://doi.org/10.18653/v1/D18-1514.
5. Gao, T.; Han, X.; Zhu, H.; Liu, Z.; Li, P.; Sun, M.; Zhou, J. FewRel 2.0: Towards More Challenging Few-Shot Relation Classification. In Proceedings of the Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, 2019; pp. 6251–6256. https://doi.org/10.18653/v1/D19-1649.
6. Yao, Y.; Ye, D.; Li, P.; Han, X.; Lin, Y.; Liu, Z.; Liu, Z.; Huang, L.; Zhou, J.; Sun, M. DocRED: A Large-Scale Document-Level Relation Extraction Dataset. In Proceedings of the Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 2019; pp. 764–777. https://doi.org/10.18653/v1/P19-1074.
7. Yamada, I.; Asai, A.; Shindo, H.; Takeda, H.; Matsumoto, Y. LUKE: Deep Contextualized Entity Representations with Entity-aware Self-attention. In Proceedings of the Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP); Webber, B.; Cohn, T.; He, Y.; Liu, Y., Eds., Online, 2020; pp. 6442–6454. https://doi.org/10.18653/v1/2020.emnlp-main.523.
8. Ri, R.; Yamada, I.; Tsuruoka, Y. mLUKE: The Power of Entity Representations in Multilingual Pretrained Language Models. In Proceedings of the Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); Muresan, S.; Nakov, P.; Villavicencio, A., Eds., Dublin, Ireland, 2022; pp. 7316–7330. https://doi.org/10.18653/v1/2022.acl-long.505.
9. Guan, T.; Zan, H.; Zhou, X.; Xu, H.; Zhang, K. CMeIE: Construction and Evaluation of Chinese Medical Information Extraction Dataset. In Proceedings of the Natural Language Processing and Chinese Computing; Zhu, X.; Zhang, M.; Hong, Y.; He, R., Eds., Cham, 2020; pp. 270–282.
10. Walker, C.; Strassel, S.; Medero, J.; Maeda, K. ACE 2005 Multilingual Training Corpus. LDC Catalog No. LDC2006T06, 2006.
11. Krallinger, M.; Rabal, O.; Lourenço, A. Overview of the BioCreative VI Chemical-Protein Interaction Track. In Proceedings of the Proceedings of the BioCreative VI Workshop, 2017, pp. 141–146.

 doi:10.20944/preprints202605.0578.v1

12. Köksal, A.; Özgür, A. The RELX Dataset and Matching the Multilingual Blanks for Cross-Lingual Relation Classification. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2020, Online, 2020; pp. 340–350. https://doi.org/10.18653/v1/2020.findings-emnlp.32.
13. Roth, D.; Yih, W.t. A Linear Programming Formulation for Global Inference in Natural Language Tasks. In Proceedings of the Proceedings of the Eighth Conference on Computational Natural Language Learning (CoNLL-2004) at HLT-NAACL 2004, Boston, Massachusetts, USA, 6 - 7 2004; pp. 1–8.
14. Hennig, L.; Thomas, P.; Möller, S. MultiTACRED: A Multilingual Version of the TAC Relation Extraction Dataset. In Proceedings of the Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); Rogers, A.; Boyd-Graber, J.; Okazaki, N., Eds., Toronto, Canada, 2023; pp. 3785–3801. https://doi.org/10.18653/v1/2023.acl-long.210.
15. Huguet Cabot, P.L.; Tedeschi, S.; Ngonga Ngomo, A.C.; Navigli, R. RED^{FM}: A Filtered and Multilingual Relation Extraction Dataset. In Proceedings of the Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); Rogers, A.; Boyd-Graber, J.; Okazaki, N., Eds., Toronto, Canada, 2023; pp. 4326–4343. https://doi.org/10.18653/v1/2023.acl-long.237.
16. Bassignana, E.; Ginter, F.; Pyysalo, S.; van der Goot, R.; Plank, B. Multi-CrossRE A Multi-Lingual Multi-Domain Dataset for Relation Extraction. In Proceedings of the Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa); Alumäe, T.; Fishel, M., Eds., Tórshavn, Faroe Islands, 2023; pp. 80–85.
17. Bassignana, E.; Plank, B. CrossRE: A Cross-Domain Dataset for Relation Extraction. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2022; Goldberg, Y.; Kozareva, Z.; Zhang, Y., Eds., Abu Dhabi, United Arab Emirates, 2022; pp. 3592–3604. https://doi.org/10.18653/v1/2022.findings-emnlp.263.
18. Ni, J.; Florian, R. Neural Cross-Lingual Relation Extraction Based on Bilingual Word Embedding Mapping. In Proceedings of the Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP); Inui, K.; Jiang, J.; Ng, V.; Wan, X., Eds., Hong Kong, China, 2019; pp. 399–409. https://doi.org/10.18653/v1/D19-1038.
19. Huguet Cabot, P.L.; Navigli, R. REBEL: Relation Extraction By End-to-end Language generation. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2021; Moens, M.F.; Huang, X.; Specia, L.; Yih, S.W.t., Eds., Punta Cana, Dominican Republic, 2021; pp. 2370–2381. https://doi.org/10.18653/v1/2021.findings-emnlp.204.
20. OpenAI. gpt-oss-120b & gpt-oss-20b Model Card, 2025, [arXiv:cs.CL/2508.10925].
21. Qi, P.; Zhang, Y.; Zhang, Y.; Bolton, J.; Manning, C.D. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In Proceedings of the Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations; Celikyilmaz, A.; Wen, T.H., Eds., Online, 2020; pp. 101–108. https://doi.org/10.18653/v1/2020.acl-demos.14.
22. Honnibal, M.; Montani, I.; Van Landeghem, S.; Boyd, A. spaCy: Industrial-strength Natural Language Processing in Python **2020**. https://doi.org/10.5281/zenodo.1212303.
23. Yang, A.; Li, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Gao, C.; Huang, C.; Lv, C.; et al. Qwen3 Technical Report, 2025, [arXiv:cs.CL/2505.09388].