Beyond Raw Text: Knowledge-Augmented Italian Relation Extraction with Large Language Models

Gianmaria Balducci[1,2,*,†], Elisabetta Fersini[1,*,†] and Enza Messina[1,*,†]

[1]Università Degli studi di Milano-Bicocca), Viale Sarca 336, Milano, 20125, Italia
[2]P.M.I. Reboot S.r.l., Viale Lunigiana 40, Milano, 20125, Italia

Abstract

Relation extraction (RE) is a fundamental NLP task that identifies semantic relationships between entities in text, serving as the foundation for applications such as knowledge graph completion and question answering. In real-world deployments, organizations frequently encounter low-resource scenarios where labeled training data is scarce, making effective RE particularly challenging. Existing approaches often rely on external knowledge sources to augment training data, but such resources can be noisy, incomplete, or misleading for model learning. To address this limitation, we propose an approach that leverages the reasoning capabilities of Large Language Models (LLMs) to generate reliable background knowledge for RE tasks on Italian texts.

Keywords

Relation Extraction, LLMs, Reasoning, Low resources, Italian

1. Introduction

Relation extraction (RE) is a fundamental task in natural language processing that aims to identify and classify relationships between subject and object entities mentioned in text [1]. Formally, given an input sentence $X_i = \{x_1, x_2, \ldots, s, \ldots, o, \ldots, x_n\}$ containing n tokens, where s and o represent head and tail entities respectively, RE systems predict a relation label $Y_i \in \mathcal{Y}$ from a predefined set of relationships (e.g., `founded_by`, `born_in`, and `Work_For`). This capability underlies many critical NLP applications, including knowledge graph completion and question answering systems [2]. Most past approaches focus on adapting standard-scale language models (SLMs) such as BERT[3] to downstream RE tasks [4]. Recent advances in RE have been driven by deep neural networks, with large pre-trained language models achieving state-of-the-art performance. However, despite these advances, several fundamental challenges persist in real-world deployment scenarios. The primary limitation stems from the long-tail distribution of relations in natural datasets. While frequent relations benefit from abundant training examples, the majority of relations suffer from severe data scarcity. This creates a significant bottleneck since deep learning approaches require substantial labeled corpora resources that are often unavailable in low-resource settings [5]. Moreover, while prompt-tuned SLMs and instruction-tuned LLMs have shown remarkable success across various NLP tasks, they exhibit a tendency to memorize rather than truly understand training data [6]. This limitation becomes particularly problematic for semantically complex tasks like RE, which require deep domain-specific knowledge and robust generalization capabilities. To address these limitations and further enhance the effectiveness of RE models, we propose a pipeline based on exploiting the LLMs' reasoning capabilities. The hypothesis is that extending each sample of a given dataset using the knowledge extracted by querying the LLM with specific clarification prompts helps the models trained on these samples, along with clarifications, to understand the task better. We train several models on an Italian dataset, CoNLL04 Italian, translated from the CoNLL04 dataset [7]. Experimental results demonstrate that incorporating LLM-generated background knowledge significantly improves RE performance, particularly in low-resource settings. Subsequently, we conduct an analysis on the contribution that different outlooks that compose the knowledge give to the model's prediction capabilities.

2. Related Work

In the current landscape dominated by Large Language Models (LLMs), Relation Extraction (RE) continues to play a pivotal role. Despite the impressive capabilities of LLMs, they often struggle to fully preserve and accurately interpret implicit relational knowledge—particularly in long-tail scenarios, where entity relations may be subtle or infrequent. These limitations highlight the contin-

CLiC-it 2025: Eleventh Italian Conference on Computational Linguistics, September 24 — 26, 2025, Cagliari, Italy
*Corresponding author.
†These authors contributed equally.
g.balducci1@campus.unimib.it (G. Balducci); elisabetta.fersini@unimib.it (E. Fersini); enza.messina@unimib.it (E. Messina)
https://github.com/jimmypuntoexe/ (G. Balducci)
0009-0009-8752-7502 (G. Balducci); 0000-0002-8987-100X (E. Fersini); 0000-0002-4062-0824 (E. Messina)

ued relevance of RE methods, which explicitly model relationships between entities and thereby enhance LLM performance. Moreover, RE techniques are especially valuable in dynamic domains characterized by the constant emergence of new entities and relation types. Their adaptability makes them well-suited for scalable knowledge extraction from unstructured textual data, fueling ongoing research and development in this area. Recent advances in deep neural networks (DNNs) and pretrained language models (PLMs) have substantially boosted RE performance. Several studies [8, 9] approach RE as a pipeline process: first identifying entities within text, then determining the relationships between identified entity pairs. Earlier RE systems [10, 11] typically relied on external Named Entity Recognition (NER) tools for entity detection, followed by the use of supervised classifiers with hand-engineered features to predict relations. In contrast, more recent approaches assume that entity mentions are pre-identified, focusing solely on relation classification [12, 13]. However, pipeline architectures are prone to error propagation—errors in entity recognition can adversely affect the accuracy of relation classification. Relation Extraction and Classification can be tackled as a generation task: REBEL [14] uses an autoregressive model that outputs each triplet present in the input text. To this end, it employs BART-large [15] as the base model for the seq2seq approach. The Italian LLM ecosystem has recently seen notable expansion, with several new models released or announced that are specifically tailored for the Italian language. Among these is **LLaMAntino-3-ANITA** [16], a fine-tuned version of Meta's LLaMA-3 (8B) [17], adapted through Supervised Fine-Tuning (SFT) and Direct Preference Optimization (DPO) to align with user preferences and reduce biases. Another significant contribution is **Fauno** [18], developed by Sapienza University as the first open-source Italian conversational LLM (7B, with a 13B version forthcoming), trained on a blend of synthetic and technical corpora. **Minerva 7B** [19], created by Sapienza NLP in collaboration with FAIR, CINECA, and Italy's National Recovery and Resilience Plan (PNRR), is trained from scratch on 2.5 trillion tokens (50% Italian), and further enhanced through instruction tuning and safety layers. **Velvet** [20], developed by Almawave, is a family of multilingual LLMs that includes Italian and is built on a proprietary architecture. This wave of Italian LLMs—from academic research efforts to industry-grade solutions—reflects a growing commitment to developing robust, safe, and effective native Italian models. These advances also contribute to improvements in downstream tasks, including RE. For instance, [21] propose an Italian Open Information Extraction framework that leverages LLMs for Open Named Entity Recognition, Open Relation Extraction, and joint tasks via prompt-based instructions. Similarly, [22] combine LLMs with fine-tuned models to extract relations from Italian literary texts. Their approach involves using an LLM to preprocess the text into natural language triples, thereby simplifying the RE task for the fine-tuned model. Existing RE methods also tend to exploit additional knowledge to assist model reasoning. For example, [23] proposes a knowledge-attention encoder that incorporates prior knowledge from external lexical resources like FrameNet and Thesaurus.com into deep neural networks for the relation extraction task. [24] uses enriched sentence-level representations by introducing both structured knowledge from external knowledge graphs and semantic knowledge from the corpus. However, external knowledge can be misleading and vague; external resources don't consider the context and the domain of entities and relations, leading models to misinterpret the meaning of the sentence.

Despite these advances, the potential of Italian LLMs to support and improve downstream RE remains largely underexplored. Given their demonstrated utility, further investigation into their integration with RE workflows is both timely and necessary.

3. Dataset

In this research the proposed approach is evaluated on an Italian translated version of CoNLL04 [7]. The CoNLL04 is a benchmark dataset used for relation extraction tasks. It contains 1,441 sentences, each of which has at least one relation. The sentences are annotated with information about entities and their corresponding relation types [25]. It comprises news articles from The Wall Street Journal and the Associated Press. It encompasses annotations for both entity and relation types, making it versatile for various NLP tasks. The dataset includes relations among entities like people, organizations, locations, and other miscellaneous entities. Relation types are five: *Live_In, Located_In, OrgBased_In, Kill, Work_for.* Relations included: Person-Location, Organization-Person, Person-Person, etc.

Table 1
CoNLL04 benchmark statistics. Every sample is a sentence.

	sentences	entities	relations
train	922	3377	1283
validation	231	893	343
test	288	1079	422
total	1441	5349	2048

This work employ a sophisticated hybrid approach for translating the ConLL04 English relation extraction dataset to Italian while preserving the crucial token-level annotations required for named entity recognition and relation extraction tasks. The translation process operates in three main phases: first, the com-

Table 2
CoNLL04 becnhamrk relation types statistics

relation type	train	validation	test
Live_In	330	91	100
Located_In	247	65	94
OrgBased_In	271	76	105
Kill	179	42	47
Work_for	256	65	76

plete English sentence is translated to Italian using X-ALMA [26], built upon ALMA-R by expanding support from 6 to 50 languages. It utilizes a plug-and-play architecture with language-specific modules, complemented by a carefully designed training recipe. In particular, a 8-bit quantized version due to resource limit constraints is used from the offical repository on Huggingface at *https://huggingface.co/mradermacher/X-ALMA-13B-Group2-GGUF*. The translator model generates fluent Italian text but disrupts the original token alignments. Second, to address the critical challenge of maintaining entity boundaries and types across languages—where direct token-to-token mapping fails due to morphological differences, word order changes, and varying translation lengths, the system employs OpenAI's GPT-4o-mini model [27] to perform intelligent entity alignment by analyzing both the original English tokens and their Italian counterparts, then identifying which specific Italian tokens correspond to each English entity based on semantic understanding rather than positional heuristics. Finally, the system reconstructs the annotated dataset by mapping the spans of the identified Italian entity back to token indices. This step has the main goal to preserve entity types and relation labels while handling edge cases through fallback mechanisms that include proportional mapping and fuzzy string matching when exact alignment fails. This ensures that the resulting Italian dataset maintains the structural integrity necessary for training and evaluating relation extraction models. The comprehensive error handling and multi-stage validation process addresses the inherent complexities of cross-lingual annotation transfer in structured NLP datasets. In each split of the dataset, some translated sentences are removed due to the impossibility of maintaining relation labels. This case is represented by a few sentences that are not well translated, in which one or more entities that were in the relationship label are missing.

Table 1 and Table 3, show the small reduction of sentences (from 1441 to 1407) and consequently of the number of relations and entities. However, in the translation process, entity types and relation types distribution are maintained 2, 4.

Table 3
CoNLL04 Ita version splits statistics

	samples	entities	relations
train	902	3284	1253
validation	224	848	325
test	281	1048	413
total	1407	5180	1991

Table 4
Relation types distribution across the dataset's split

relation type	train	validation	test
Vive_A	322	88	95
Situato_In	243	64	94
OrgLocata_In	256	64	103
Ha_ucciso	178	40	46
Lavora_per	254	69	75

4. Method

4.1. Background

This work considers an LLM as a reliable Knowledge Base (KB). Large Language Models (LLMs) offer significant advantages over external knowledge bases like Wikidata for relation extraction tasks, particularly in their superior ability to interpret sentence semantics and contextual nuances. Unlike Wikidata, which provides static, predefined relations between entities in a structured format, LLMs possess deep contextual understanding that enables them to capture implicit relationships, resolve ambiguities, and interpret complex linguistic phenomena such as metaphors, negations, and conditional statements that traditional knowledge bases cannot handle. LLMs excel at understanding how the same entity pair can express different relations depending on syntactic structure, discourse context, and pragmatic implications—for instance, distinguishing between "CEO of Apple" and "former CEO of Apple" or interpreting temporal and causal relationships that emerge from sentence composition rather than explicit statement. Furthermore, LLMs can handle novel entity combinations and emerging relationships that may not yet exist in manually curated databases, while their training on vast text corpora allows them to recognize subtle linguistic cues and contextual modifiers that determine relation validity and type. This semantic depth proves particularly valuable for relation extraction in domains with complex, evolving terminology or when dealing with informal text where relationships are expressed through natural language patterns rather than formal declarations, making LLMs more robust and adaptable for real-world text analysis scenarios where meaning emerges from the intricate interplay of syntax, semantics, and context. Given a sentence $s = \{w_1, w_2, \ldots, w_n\}$ consisting of n tokens, and

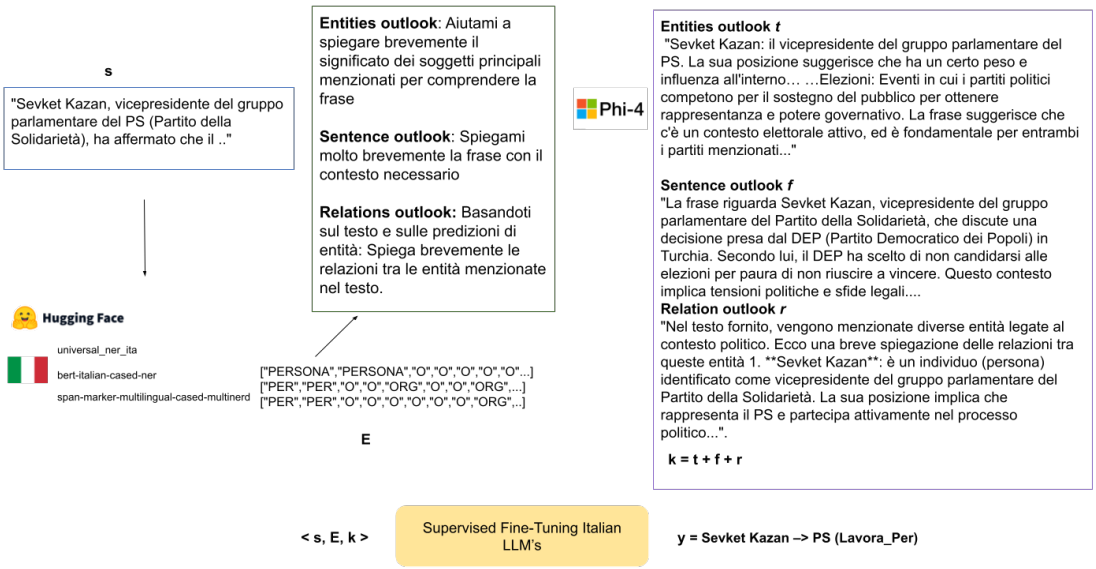

Figure 1: Overview of proposed approach. Starting from the input sentence, the method augment the input with NER predictions and knowledge extracted from Phi4. Subsequently, a supervised fine-tuning with LoRA strategy is performed. LLMs learn to generate the target with a specific notation.

a set of entities $E = \{e_1, e_2, \ldots, e_k\}$ where each entity e_i is defined by its span $(start_i, end_i)$ and type $t_i \in \mathcal{T}$, the relation extraction task aims to identify and classify semantic relationships between entity pairs. Formally, let $\mathcal{R}$ be the set of all possible relation types, including a special *no-relation* type $\emptyset \in \mathcal{R}$. For each ordered pair of entities (e_i, e_j) where $i \neq j$, the relation extraction task seeks to determine the relation type $r_{ij} \in \mathcal{R}$ that holds between e_i (head entity) and e_j (tail entity) within the context of sentence s.

4.2. NER predictions

This step involves in the extension of the input space using state-of-the-art Named Entity Recognition (NER) Italian models. NER is formulated as a sequence labeling task where each token in the input sequence is assigned a label that indicates its role in entity identification and classification. Given an input sentence $s = \{w_1, w_2, \ldots, w_n\}$ consisting of n tokens, the NER task aims to produce a corresponding label sequence $y = \{y_1, y_2, \ldots, y_n\}$ where each label $y_i \in \mathcal{L}$ encodes both the entity type and the token's position within the entity span. In particular, for each of input sentences of the dataset, this work construct a set of NER predictions E comprising annotations from three state-of-the-art multilingual and Italian-specific named entity recognition models. The prediction ensemble includes: (1) `span-marker-multilingual-cased-multinerd`, [28] a SpanMarker model fine-tuned on the MultiNERD. (2) `bert-italian-cased-ner` [29], a cased BERT model specifically trained for Italian NER on the WikiNER Italian dataset plus manually annotated Wikipedia paragraphs, capable of recognizing four entity classes (PER, LOC, ORG, MISC); and (3) `DeepMount00/universal_ner_ita`, an Italian adaptation of GLiNER [30] (Generalist Model for Named Entity Recognition using Bidirectional Transformer) that leverages natural language descriptions to identify arbitrary entity types. Entity types for GLiNER are *"persona", "città", "nazione", "organizzazione", "data", "luogo", "evento", "prodotto"* (*"person", "city", "nation", "organisation", "date", "location", "event", "product"*). Each model processes the tokenized Italian sentences independently, with predictions aligned to the original token boundaries. The resulting prediction set E composed of all the token-level predictions obtained from cited models provides diverse perspectives on entity recognition.

4.3. Knowledge Extraction

Given the extended input (s, E) the aim of this step is to further extend the input, extracting knowledge **k** from LLM. **k** is composed by three different outlooks that are concatenated together to compose the semantic interpre-

tation of a single dataset sample. In particular for a given sentence $s_i \in S$ where S represent the entire corpus of a dataset, $k_i = t_i \oplus f_i \oplus r_i$ where t is the **Entities outlook**, f is the **Sentence outlook** and r is the **Relations outlook**.

- For the Entities outlook we ask to the LLM: "Spiega brevemente il significato dei soggetti principali menzionati per comprendere la frase: {s}" ("Briefly explain the meaning of the main subjects mentioned in order to understand the sentence: {s}").
- Sentence outlook is obtained by asking "Spiegami molto brevemente la frase con il contesto necessario: {s}" ("Explain the sentence to me very briefly, providing the necessary context: {s}").
- Relation outlook is obtained asking "Basandoti sul testo e sulle predizioni di entità: Spiega brevemente le relazioni tra le entità menzionate nel testo. Testo: {s} Predizioni NER {E}" ("Based on the text and entity predictions: Briefly explain the relationships between the entities mentioned in the text. Text: {s} NER predictions {E}")

The model used to extract the Italian knowledge is Phi-4 [31] a 14B parameter state-of-the-art open model, due to the high quality and advanced multilingual reasoning capabilities, even though the small size. In this settings we are able to concatenate the sentence with NER predictions E and knowledge **k** in order to represent the **enriched input** space **<s , E, k >** for a given sentence $s \in S$. Given this input space we employ a parameter-efficient fine-tuning strategy using Low-Rank Adaptation (LoRA) [32] within the PEFT framework [33] for supervised fine-tuning (SFT) of several Italian LLMs.

4.4. Target representation

Relations triplets are composed of a head entity, a tail entity, and a predicate indicating the semantic relationship between a subject entity and the object entity:

"Hideo Kojima ha acquistato una nuova casa a Tokyo." ("Hideo Kojima has purchased a new home in Tokyo.")

The semantic relationship according to CoNLL04 annotation can be (Hideo Kojima, Vive_A, Tokio). Inspired by REBEL triplets linearization [14], we try to minimize the number of tokens in the generation stream in order to decode the output tokens efficiently. A relation triplet is represented by this notation:

$$\textbf{Head Entity -> Tail Entity (Relation type)} \quad (1)$$

Multiple relations are separated by the semicolon character "**;**".
In this work relation extraction is treated as a generation task where the aim is to learn the conditional probability distribution given the input X = <s, E, k> :

$$P(Y|X) = P(y| < s, E, k >) \quad (2)$$

A few Italian LLM's are fine-tuned using LoRA strategy in order to learn to generate the target representation 1. We fine-tune also mREBEL$_{32}$ [34], a multilingual version of REBEL [14]. All models are fine-tuned for 10 epochs. At the end of each epoch, models are evaluated on the validation set, best model on the evaluation set is saved. Translation process, Knowledge extraction step, and training step are executed on the same machine with a NVIDIA GeForce RTX 3090 with 24GB of memory and AMD Ryzen 9 5900X 12-Core Processor.

5. Results

In this section, we present the experimental results of our supervised fine-tuning approach on the Italian ConLL04 dataset. We evaluate multiple Italian large language models under different input configurations to assess the effectiveness of our generative relation extraction framework. We conduct experiments using three configurations:

- **Enriched**: Complete input including sentence, entity predictions, and background knowledge $\langle s, E, k \rangle$
- **Raw**: Input containing only the source sentence $\langle s \rangle$
- **Enriched-Raw**: Model fine-tuned on enriched input but evaluated using only raw sentence input at inference time

The enriched-raw configuration allows us to investigate the implicit knowledge distillation effects, where reasoning capabilities from the enriched training data transfer to simpler inference scenarios.

5.1. Main Results

Table 5 presents the performance comparison across different Italian language models and input configurations. Following standard practice in relation extraction, we report both micro and macro F1 scores, with macro F1 serving as the primary evaluation metric for state-of-the-art comparisons.

5.2. Performance Analysis

LLaMAntino-3 demonstrates superior performance when trained and evaluated on enriched input, achieving 70.6% macro F1 score. This represents a significant improvement over both Minerva-7B (59.6%) and Velvet-14B

Table 5
Performance comparison of supervised fine-tuned Italian LLMs on relation extraction. Input configurations: (enriched) includes entity predictions and background knowledge; (raw) uses only sentence text; (enriched-raw) represents models trained on enriched data but evaluated with raw input only.

Model Configuration	F1 Micro	F1 Macro
mREBEL (enriched)	62.7	63.9
mREBEL (raw)	58.1	59.6
mREBEL (enriched-raw)	49.7	49.12
Minerva-7B (enriched)	57.2	59.6
Minerva-7B (raw)	55.6	57.9
Minerva-7B (enriched-raw)	48.9	51.0
Velvet-14B (enriched)	56.9	60.2
Velvet-14B (raw)	63.0	65.2
Velvet-14B (enriched-raw)	42.6	46.4
LLaMAntino-3 (enriched)	**68.5**	**70.6**
LLaMAntino-3 (raw)	58.4	62.1
LLaMAntino-3 (enriched-raw)	61.1	64.9

(60.2%), despite LLaMAntino-3 being a smaller 8B parameter model. The results indicate that model architecture and training methodology are more critical factors than pure parameter count for this task. The strong performance of mREBEL demonstrates that sequence-to-sequence models, which were previously state-of-the-art for this task, can achieve comparable results to large language models (LLMs). Additionally, mREBEL benefits from enriched input. However, Velvet-14B exhibits the opposite behavior, performing better with raw input (65.2%) than with enriched input (60.2%). This suggests the model may be overfitting to the auxiliary information provided in the enriched input. Comparing LLaMAntino-3 configurations reveals the substantial benefit of enriched input during training. The model trained on enriched data (70.6% macro F1) significantly outperforms the same model trained solely on raw sentences (62.1% macro F1). This demonstrates the value of incorporating entity predictions and background knowledge in the training process. The enriched-raw configuration yields particularly interesting results, achieving 64.9% macro F1 despite using only raw sentence input at inference time. This performance exceeds that of the model trained exclusively on raw input (62.1% macro F1), suggesting an interesting implicit knowledge distillation during training. The model appears to internalize reasoning patterns from the enriched training data, enabling improved performance even when auxiliary information is unavailable at inference time. Table 5.2 shows label-wise performances where the underlying capability of LLaMAantino3-8B to predict well the "Kill" relation, which is the least represented in the training set. These results validate our approach of treating relation extraction as a conditional text generation task and demonstrate the effectiveness of supervised fine-tuning on Italian language models

Table 6
Label wise performances of best model LLaMAantino3-8B (encriched)

relation type	precision	recall	f1
Vive_a	69.9	61.05	65.17
OrgLocata_In	71.95	63.44	67.42
Situato_In	60.63	60.63	60.63
Lavora_Per	62.5	80.0	70.17
Ha_ucciso	86.0	93.4	89.58

for this domain. Error analysis and Ablation study, presented in this section are perfromed on the best model LLaMAntino-3.

5.3. Error Analysis

Error analysis reveals two primary failure modes in the LLaMAntino-3 model's relation extraction performance: **spurious relation generation** (41 instances) and **missed relation detection** (37 instances). The model demonstrates a tendency toward over-generation, particularly struggling with complex sentences containing multiple entities where it produces semantically plausible but factually incorrect relations. Geographic relations (*Situato_In*) show the highest error rates, followed by organizational affiliations (*OrgLocata_In*). Two representative error patterns illustrate these challenges: **Over-generation example**: In the sentence "*Nikita Chruščëv, infuriato, ordinò alle navi dell'Unione Sovietica di ignorare il blocco navale del Presidente Kennedy durante la crisi dei missili cubani*", the model incorrectly generated four identical *Kill* relations between Khrushchev and Kennedy, while missing the correct *Vive_A* relation between Khrushchev and the Soviet Union. This demonstrates the model's tendency to infer dramatic but incorrect relations from contextual conflict scenarios. **Under-detection example**: For the sentence "*MILANO, Italia (AP)*" (Milan, Italy (AP)), the model correctly identified organizational relations for the Associated Press but failed to extract the fundamental *Situato_In* relation between Milan and Italy, suggesting difficulty with implicit geographic knowledge in simple locative constructions. **Out-of-domain hallucination example**: In the sentence "*King venne ucciso il 4 aprile del 1968 a Memphis, nel Tennessee*", the model correctly identified the *Situato_In* relation between Memphis and Tennessee, but additionally generated correct (but counted as wrong) *Evento* relations involving the date "4 aprile del 1968" with Memphis. The *Evento* relation type does not exist in the defined schema, demonstrating the model's tendency to create novel relation categories when encountering temporal-spatial contexts. These patterns indicate that while the generative approach successfully captures complex relational semantics, it requires improved calibration mechanisms,

particularly for handling entity-dense contexts and fundamental geographic relations.

6. Ablation Study

Table 7 presents the performance impact of removing each knowledge component individually. The baseline enriched model achieves 70.6% macro F1, serving as our reference point for measuring component contributions.

Table 7
Ablation study results showing the impact of removing individual knowledge outlook components from the enriched input. Each configuration excludes one specific knowledge type while maintaining entity predictions and the base sentence.

Model Configuration	F1 Micro	F1 Macro
LLaMAntino-3 (enriched)	**68.5**	**70.6**
LLaMAntino-3 without Entity Outlook	60.1	62.6
LLaMAntino-3 without Sentence Outlook	49.0	50.6
LLaMAntino-3 without Relation Outlook	63.1	65.7

The ablation results reveal distinct contribution patterns for each knowledge component: Removing **sentence contextualization** causes the most severe performance degradation, with macro F1 dropping by 20.0 percentage points (70.6% $\rightarrow$ 50.6%). This dramatic decline indicates that contextual sentence understanding is fundamental to relation extraction performance. The sentence outlook provides essential discourse-level information that enables the model to disambiguate entity relationships within specific contextual frameworks. Excluding **entity explanations** results in an 8.0 percentage point decrease (70.6% $\rightarrow$ 62.6%), demonstrating the importance of explicit entity semantics. Entity-focused knowledge helps the model understand the nature and characteristics of mentioned entities, facilitating more accurate relation inference. Removing **relation-specific** explanations leads to a 4.9 percentage point reduction (70.6% $\rightarrow$ 65.7%), showing the smallest but still meaningful impact. While relation outlook provides valuable relational reasoning guidance, the model appears capable of inferring relations from entity and sentence context when this component is absent. The ablation study reveals a clear hierarchy of knowledge component importance: **Sentence Context > Entity Semantics > Relation Guidance**. This hierarchy suggests that: **Contextual understanding** is paramount for relation extraction, as sentences provide the situational framework within which entities interact **Entity semantics** serve as the foundation for identifying potential relation participants and their characteristics **Explicit relational reasoning** provides incremental benefits but is less critical when strong contextual and entity understanding exists. These findings highlight the differential contribution of each component to the overall system performance. The results also suggest potential optimization strategies, where computational resources could be prioritized toward generating high-quality sentence and entity explanations when resource constraints exist.

7. Conclusion

This work presents an effective approach for Italian relation extraction that leverages Large Language Models as reliable knowledge sources to enhance model performance in low-resource scenarios. Our method systematically augments training data with three complementary knowledge components: entity explanations, sentence contextualization, and relation-specific guidance, extracted using Phi-4's reasoning capabilities. The experimental results on the Italian CoNLL04 dataset demonstrate the effectiveness of our approach, with LLaMAntino-3 achieving 70.6% macro F1 when trained on enriched input, representing significant improvements over baseline configurations. The ablation study reveals a clear hierarchy of component importance: sentence context (20.0% performance drop when removed) $>$ entity semantics (8.0% drop) $>$ relation guidance (4.9% drop), highlighting the critical role of contextual understanding in relation extraction. Particularly noteworthy is the implicit knowledge distillation effect observed in the enriched-raw configuration for LLaMAntino-3, trained on enriched data but evaluated with raw input, still outperforms the same model trained exclusively on raw sentences (64.9% vs 62.1% macro F1). This suggests that some reasoning patterns from the enriched training data are internalized by the smallest model.

Limitations: An important limitation is that this approach relies heavily on the choice of LLM from which the knowledge is extracted. It would be interesting to investigate the contribution to the task of several LLM that can be used as knowledge-based. Furthermore the results of SFT depend on the well-formatted prompt used in the training phase. **A promising direction** for future work involves explicit knowledge distillation from enriched input $\langle s, E, k\rangle$ to raw input $\langle s\rangle$. This could be achieved by minimizing the Jensen-Shannon divergence or Kullback-Leibler divergence between the output distributions of models trained on enriched versus raw inputs. Such an approach would enable the deployment of lightweight models that maintain the reasoning capabilities learned from enriched training while operating solely on raw text at inference time, making the system more practical for real-world applications where auxiliary information may not be readily available. The work contributes to the

growing body of research on Italian NLP by providing both a translated benchmark dataset and demonstrating effective strategies for leveraging LLM reasoning in structured prediction tasks. Our findings suggest that carefully designed knowledge augmentation can significantly improve relation extraction performance, particularly in scenarios where training data is limited.

References

[1] X. Zhao, Y. Deng, M. Yang, L. Wang, R. Zhang, H. Cheng, W. Lam, Y. Shen, R. Xu, A comprehensive survey on relation extraction: Recent advances and new frontiers, ACM Comput. Surv. 56 (2024). URL: https://doi.org/10.1145/3674501. doi:10.1145/3674501.

[2] X. Zhao, Y. Deng, M. Yang, L. Wang, R. Zhang, H. Cheng, W. Lam, Y. Shen, R. Xu, A comprehensive survey on relation extraction: Recent advances and new frontiers, 2024. URL: https://arxiv.org/abs/2306.02051. arXiv:2306.02051.

[3] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. URL: https://arxiv.org/abs/1810.04805. arXiv:1810.04805.

[4] I. Yamada, A. Asai, H. Shindo, H. Takeda, Y. Matsumoto, LUKE: Deep contextualized entity representations with entity-aware self-attention, in: B. Webber, T. Cohn, Y. He, Y. Liu (Eds.), Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 6442–6454. URL: https://aclanthology.org/2020.emnlp-main.523/. doi:10.18653/v1/2020.emnlp-main.523.

[5] A. Layegh, A. H. Payberah, A. Soylu, D. Roman, M. Matskin, Wiki-based prompts for enhancing relation extraction using language models, in: Proceedings of the 39th ACM/SIGAPP Symposium on Applied Computing, SAC '24, Association for Computing Machinery, New York, NY, USA, 2024, p. 731–740. URL: https://doi.org/10.1145/3605098.3635949. doi:10.1145/3605098.3635949.

[6] S. Pan, L. Luo, Y. Wang, C. Chen, J. Wang, X. Wu, Unifying large language models and knowledge graphs: A roadmap, IEEE Transactions on Knowledge and Data Engineering 36 (2024) 3580–3599. URL: http://dx.doi.org/10.1109/TKDE.2024.3352100. doi:10.1109/tkde.2024.3352100.

[7] D. Roth, W.-t. Yih, A linear programming formulation for global inference in natural language tasks, in: Proceedings of the Eighth Conference on Computational Natural Language Learning (CoNLL-2004) at HLT-NAACL 2004, Association for Computational Linguistics, Boston, Massachusetts, USA, 2004, pp. 1–8. URL: https://aclanthology.org/W04-2401.

[8] Y. Yuan, X. Zhou, S. Pan, Q. Zhu, Z. Song, L. Guo, A relation-specific attention network for joint entity and relation extraction, in: International joint conference on artificial intelligence, International Joint Conference on Artificial Intelligence, 2021.

[9] T. Zhao, Z. Yan, Y. Cao, Z. Li, Asking effective and diverse questions: A machine reading comprehension based framework for joint entity-relation extraction, in: Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence, 2021, pp. 3948–3954.

[10] S. Pawar, G. K. Palshikar, P. Bhattacharyya, Relation extraction: A survey, arXiv preprint arXiv:1712.05191 (2017).

[11] Z. Huang, W. Xu, K. Yu, Bidirectional lstm-crf models for sequence tagging, arXiv preprint arXiv:1508.01991 (2015).

[12] L. Weber, S anger, m., garda, s. et al.(2021) humboldt@ drugprot: chemical-protein relation extraction with pretrained transformers and entity descriptions, in: Proceedings of the BioCreative VII challenge evaluation workshop, ????, pp. 22–25.

[13] A. Bhartiya, K. Badola, et al., Dis-rex: A multilingual dataset for distantly supervised relation extraction, arXiv preprint arXiv:2104.08655 (2021).

[14] P.-L. Huguet Cabot, R. Navigli, REBEL: Relation extraction by end-to-end language generation, in: M.-F. Moens, X. Huang, L. Specia, S. W.-t. Yih (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2021, Association for Computational Linguistics, Punta Cana, Dominican Republic, 2021, pp. 2370–2381. URL: https://aclanthology.org/2021.findings-emnlp.204/. doi:10.18653/v1/2021.findings-emnlp.204.

[15] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, L. Zettlemoyer, BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, CoRR abs/1910.13461 (2019). URL: http://arxiv.org/abs/1910.13461. arXiv:1910.13461.

[16] M. Polignano, P. Basile, G. Semeraro, Advanced natural-based interaction for the italian language: Llamantino-3-anita, 2024. arXiv:2405.07101.

[17] AI@Meta, Llama 3 model card (2024). URL: https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md.

[18] A. S. F. S. Andrea Bacciu, Giovanni Trappolini, Fauno: The italian large language model that will leave you senza parole!, https://github.com/andreabac3/Fauno-Italian-LLM, 2023.

[19] R. Navigli, S. N. group, Minerva: Italy's first family

of large language models trained on italian texts (2024).
[20] Almawave, Velvet ai: sustainable and high-performance italian multilingual llm, Wikipedia, 2025.
[21] L. Piano, A. Pisu, S. G. Tiddia, S. Carta, A. Giuliani, L. Pompianu, Llimoniie: Large language instructed model for open named italian information extraction (2024).
[22] C. Santini, G. Marozzi, L. Melosi, E. Frontoni, Leveraging large language models to generate a knowledge graph from italian literary texts, in: DH2024 Book of Abstracts, 2024.
[23] P. Li, K. Mao, X. Yang, Q. Li, Improving relation extraction with knowledge-attention, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, 2019, p. 229–239. URL: http://dx.doi.org/10.18653/v1/D19-1022. doi:10.18653/v1/d19-1022.
[24] J. Gao, H. Wan, Y. Lin, Exploiting global context and external knowledge for distantly supervised relation extraction, Knowledge-Based Systems 261 (2023) 110195. URL: https://www.sciencedirect.com/science/article/pii/S0950705122012916. doi:https://doi.org/10.1016/j.knosys.2022.110195.
[25] Y. Tao, Y. Wang, L. Bai, Graphical reasoning: Llm-based semi-open relation extraction, 2024. URL: https://arxiv.org/abs/2405.00216. arXiv:2405.00216.
[26] H. Xu, K. Murray, P. Koehn, H. Hoang, A. Eriguchi, H. Khayrallah, X-alma: Plug play modules and adaptive rejection for quality translation at scale, 2025. URL: https://arxiv.org/abs/2410.03115. arXiv:2410.03115.
[27] OpenAI Team, GPT-4o mini: advancing cost-efficient intelligence, https://openai.com/gpt4o-mini, 2024. Read me. Accessed on 23 Aug. 2024.
[28] lxyuan, span-marker-bert-base-multilingual-cased-multinerd, https://huggingface.co/lxyuan/span-marker-bert-base-multilingual-cased-multinerd, 2023. Fine-tuned SpanMarker model based on bert-base-multilingual-cased for multilingual named entity recognition on MultiNERD dataset.
[29] osiria, bert-italian-cased-ner, https://huggingface.co/osiria/bert-italian-cased-ner, 2023. BERT-based model for Italian Named Entity Recognition, fine-tuned on WikiNER dataset for Person, Location, Organization and Miscellaneous entity classes.
[30] U. Zaratiana, N. Tomeh, P. Holat, T. Charnois, GLiNER: Generalist model for named entity recognition using bidirectional transformer, in: K. Duh, H. Gomez, S. Bethard (Eds.), Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), Association for Computational Linguistics, Mexico City, Mexico, 2024, pp. 5364–5376. URL: https://aclanthology.org/2024.naacl-long.300/. doi:10.18653/v1/2024.naacl-long.300.
[31] M. Abdin, J. Aneja, H. Behl, S. Bubeck, R. Eldan, S. Gunasekar, M. Harrison, R. J. Hewett, M. Javaheripi, P. Kauffmann, J. R. Lee, Y. T. Lee, Y. Li, W. Liu, C. C. T. Mendes, A. Nguyen, E. Price, G. de Rosa, O. Saarikivi, A. Salim, S. Shah, X. Wang, R. Ward, Y. Wu, D. Yu, C. Zhang, Y. Zhang, Phi-4 technical report, 2024. URL: https://arxiv.org/abs/2412.08905. arXiv:2412.08905.
[32] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, Lora: Low-rank adaptation of large language models, arXiv preprint arXiv:2106.09685 (2021).
[33] S. Mangrulkar, S. Gugger, L. Debut, Y. Belkada, S. Paul, Peft: State-of-the-art parameter-efficient fine-tuning methods, https://github.com/huggingface/peft, 2022.
[34] P.-L. Huguet Cabot, S. Tedeschi, A.-C. Ngonga Ngomo, R. Navigli, Red^{fm}: a filtered and multilingual relation extraction dataset, in: Proc. of the 61st Annual Meeting of the Association for Computational Linguistics: ACL 2023, Association for Computational Linguistics, Toronto, Canada, 2023. URL: https://arxiv.org/abs/2306.09802.