Efficient Adaptive Information Extraction for Knowledge Graphs

Fina Polat[1], Paul Groth[1], and Jan-Christoph Kalo[1]

University of Amsterdam, Netherlands
{f.yilmazpolat, p.t.groth, j.c.kalo}@uva.nl
https://indelab.org/

Abstract. Extraction of knowledge graphs from unstructured text faces a persistent trade-off between rigid, domain-specific pipelines and flexible but resource-intensive Large Language Models. While traditional pipelines suffer from cascading error propagation, frontier LLMs are often impractical due to privacy concerns and high operational costs. In this demo, we present a schema-adaptive framework for both Named Entity Recognition (NER) and Relation Extraction (RE) that is privacy-aware. By optimizing the system to run locally on consumer-level hardware, sensitive data can be processed without the need for external API calls. Our system is based on a fine-tuned 14.7B-parameter LLM, specialized through conversational alignment on the Text2KGBench benchmark using a declarative instruction architecture. This alignment strategy enables zero-shot schema adaptation by allowing the model to conform to dynamic ontologies across diverse domains, ranging from astronomy to music. Deployment instructions of the model can be found at: https://github.com/EnexaProject/phi4-ie-demo. A video showcasing the system in action can be viewed at: https://youtu.be/XGw8txwBfwk

Keywords: Knowledge Graph Construction · Ontology-Based Information Extraction · Large Language Models.

1 Introduction

Knowledge Graph Construction (KGC) remains a cornerstone of the Semantic Web [5], traditionally relying on multi-stage information extraction pipelines. Current research highlights a paradigm shift where Large Language Models (LLMs) replace the traditional fragmented pipeline with cohesive, instruction-driven information extraction processes [6,2]. The inherent natural language processing and instruction-following capabilities of LLMs allow for *zero-shot* or *few-shot* adaptation, where a model can conform to a new schema simply via prompting [7]. Despite this potential, practical adoption in many sensitive sectors (e.g., healthcare, legal, or industrial AI) is hindered by three primary barriers:

(1) Instruction Adherence. Standard open-weight models often struggle with complex, multi-step instructions and strict output formatting. They frequently introduce conversational fillers or do not respect constraints in zero-shot settings.

(2) Privacy and Governance. Proprietary models (e.g., GPT-4, Claude) require data to be transmitted to third-party servers, which can violate data-sovereignty requirements.
(3) Resource Intensity. High-performance open-source models often require data-center-grade hardware, making local, decentralized, or *edge* semantic processing infeasible [9].

In this paper, we demonstrate a ontology-adaptive, and privacy aware solution to these challenges. We introduce a specialized variant of the 14.7B parameter Phi-4 model [1], optimized via conversational alignment on the Text2KGBench [8] benchmark. By utilizing supervised fine-tuning (SFT) and a structured prompt architecture, the model maintains effective information extraction (IE) cap while adhering to schema constraints. Furthermore, our approach facilitates efficient knowledge extraction on consumer-grade hardware. Using 4-bit quantization, we show that prompt-driven IE can run entirely on local CPUs, removing the need for GPU clusters or external API dependencies.

2 Methodology and Implementation

2.1 Role-Conditioned Extraction.

We adopt a role-conditioned extraction framework that enables ontology-adaptive NER and RE using a single language model. Rather than relying on static pipelines, extraction is framed as a schema-conditioned generation task executed within a chat-based inference setup. The model behavior is controlled by a structured prompt composed of (i) a system role definition,(ii) a task specification (NER or RE), (iii) a target schema, (iv) an extraction format, (v) a one-shot demonstration (optional), (vi) the input text.

The system prompt establishes a persistent IE role, while the user prompt instantiates a parameterized template that explicitly defines output format and reinforce ontology compliance. Demonstration examples expose both schema structure and output serialization, acting as behavioral anchors. Each user query is converted into:

$$\text{Prompt} = f\left(\underbrace{\tau}_{\text{task}}, \underbrace{\sigma}_{\text{schema}}, \underbrace{\phi}_{\text{format}}, \underbrace{\mathbf{x}}_{\text{text}} \left[, \underbrace{e}_{\text{example}} \right]\right)$$

where τ denotes the task specification (NER or RE), σ the target ontology schema defining the entity and relation types to extract, ϕ the extraction format constraining output serialization (e.g., JSON or triple notation), $\mathbf{x}$ the input text to be processed, and e an optional one-shot demonstration example. The square brackets around $[, e]$ denote that the demonstration is conditionally included depending on whether a few-shot setting is employed. Together, these five arguments guide the model's extraction behavior at inference time, with the schema σ serving as the primary axis of adaptation across ontologies.

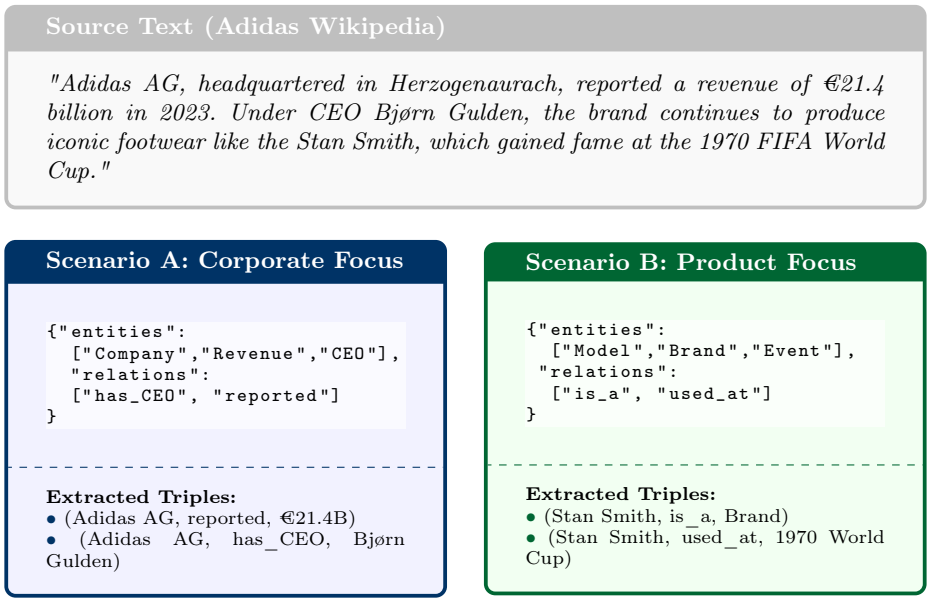

Fig. 1. Target schema acts as an inference-time parameter, directing the extraction model to extract different type of knowledge from identical input text.

This prompt design enables zero-shot schema adaptation without retraining when ontologies change. Figure 1 illustrates this behavior. With a Corporate Focus schema, the model extracts organizational attributes such as revenue and executive roles; when switched to a Product Focus schema, the same text yields product and event-related triples. This confirms that schemas function as an inference-time parameter rather than soft prompts.

2.2 Chat-Based Supervised Fine-Tuning.

For instruction adherence, we fine-tune Phi-4 (14.7B) [1] using the WebNLG [4] corpus which is a curated subset of Text2KGBench [8]. WebNLG is selected for its high-quality, human-authored mappings across 19 ontologies. Training data is converted into chat-formatted supervision consisting of system, user, and assistant turns, with structured extraction outputs as targets. Separate NER and RE prompt instantiations are used while sharing a common template. Supervised fine-tuning is performed using QLoRA [3] for parameter-efficient adaptation in a single epoch. After training, LoRA adapters are merged into the base model, yielding a standalone extraction model compatible with the Hugging Face ecosystem[1].

[1] The model is available at: `https://huggingface.co/FinaPolat/phi4_adaptableIE_v2`

2.3 Local Deployment and Quantized Inference.

To support privacy-preserving and resource-efficient deployment, the merged model weights are converted into the GGUF inference format. The quantized model is deployed locally using Ollama, which provides runtime orchestration, conversational prompt management, and a low-latency command-line interface for interactive extraction workflows. Inference runs entirely on consumer hardware (16 GB RAM). This setup eliminates external API dependencies, helping with data sovereignty.

3 Demonstration Scenario

We demonstrate the model's dynamic schema steering capability through an interactive, *user-in-the-loop* extraction session using command-line interface (CLI). Users specify a target schema and unstructured input text. A single role-conditioned language model is used for both NER and RE. The IE model generates entity–type pairs for NER and subject–predicate–object triples for RE, while guided by the supplied schema. By modifying the schema, the same input text can be reprocessed to extract different ontology-driven information without retraining. All outputs are returned as structured plain-text lists, ready for downstream knowledge graph construction pipelines. During the demo, users will be able to specify their own input text as well as schema.

4 Conclusion

We presented a schema-adaptive, privacy-aware IE system that supports both NER and RE using a single role-conditioned language model. Task behavior is guided explicitly at inference time, enabling flexible reuse of the same model across extraction objectives and domains. By combining supervised fine-tuning with efficient quantization and local CLI-based deployment, the system delivers ontology-driven extraction while helping to perseve data sovereignty. This demo illustrates a practical approach for integrating LLM-based extraction into knowledge graph construction workflows under real-world deployment constraints. Future work includes comprehensive evaluation of the performance of the extraction of this model as well as using this approach with other base LLMs.

5 Acknowledgments

This work is funded by the European Union's Horizon Europe research and innovation programme within the ENEXA project (grant Agreement no. 101070305). We employed Generative AI including Claude and ChatGPT to help check the language of this work.

References

1. Abdin, M., Aneja, J., Behl, H., Bubeck, S., et al.: Phi-4 technical report (2024), https://arxiv.org/abs/2412.08905
2. Bian, H.: Llm-empowered knowledge graph construction: A survey (2025), https://arxiv.org/abs/2510.20345
3. Dettmers, T., Pagnoni, A., Holtzman, A., Zettlemoyer, L.: QLORA: Efficient finetuning of quantized LLMs. In: Proceedings of the 37th International Conference on Neural Information Processing Systems (NeurIPS) (2023), https://dl.acm.org/doi/10.5555/3666122.3666563
4. Gardent, C., Shimorina, A., Narayan, S., Perez-Beltrachini, L.: Creating training corpora for NLG micro-planners. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 179–188 (2017), https://aclanthology.org/P17-1017/
5. Hogan, A., Blomqvist, E., Cochez, M., D'amato, C., et al.: Knowledge graphs. ACM Computing Surveys **54**(4), 1–37 (2021), https://doi.org/10.1145/3447772
6. Li, X.: From fine-tuning to prompting: A paradigm shift in knowledge graph construction (2026), https://dare.uva.nl/search?identifier=d6bbcff4-d7be-4bbc-9fdd-ff038259e397
7. Lu, Y., Liu, Q., Dai, D., Xiao, X., et al.: Unified structure generation for universal information extraction. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 5755–5772 (2022), https://aclanthology.org/2022.acl-long.395/
8. Mihindukulasooriya, N., Tiwari, S., Enguix, C.F., Lata, K.: Text2KGBench: A benchmark for ontology-driven knowledge graph generation from text. In: The Semantic Web – ISWC 2023. pp. 247–265. Springer Nature Switzerland, Cham (2023). https://doi.org/10.1007/978-3-031-47243-5_14
9. Morabito, R., Jang, S.: Smaller, smarter, closer: The edge of collaborative generative ai (2025), https://arxiv.org/abs/2505.16499