First Workshop on Small and Efficient Large Language Models for Knowledge Extraction (SmaLLEXT @ CIKM2025)

Felice Antonio Merra[1], Kristian Skračić[1], Daniele Malitesta[2], Jacek Golebiowski[3] and Pasquale Minervini[4]

[1]*Cognism, London, United Kingdom*
[2]*Université Paris-Saclay, Gif-sur-Yvette, France*
[3]*distil labs, Berlin, Germany*
[4]*University of Edinburgh and Miniml.AI, Edinburgh, United Kingdom*

Abstract

This preface introduces the Proceedings of the "First Workshop on **Sma**ll and Efficient **LL**Ms for Knowledge **Ext**raction)" (**SmaLLEXT**), that was co-located with the 34th ACM International Conference on Information and Knowledge Management (CIKM 2025) and held in Seoul (Korea) on November 14, 2025. SmaLLEXT offered a dedicated satellite event at CIKM 2025 for researchers and practitioners interested in developing small and efficient language models for knowledge extraction, operating with limited memory and low latency while still achieving high accuracy. The workshop featured a keynote speech, an industrial talk, and two paper sessions (with five presented papers). The website is accessible at: https://sites.google.com/view/smallext.

Keywords

Knowledge Extraction, Large Language models (LLMs), Efficient Machine Learning

1. Introduction

The extensive adoption of large language models (LLMs) has driven notable progress in knowledge extraction (KE), as well as in understanding and reasoning over large-scale unstructured data [1]. Nevertheless, KE has become a critical bottleneck for the successful deployment of LLMs in real-world applications. The high computational demands of state-of-the-art LLMs limit their scalability and hinder their practical use, especially in resource-constrained settings such as enterprise systems with strict latency requirements. Domains including finance, healthcare, legal technology, and web-scale analytics require solutions capable of extracting structured information from noisy and heterogeneous data while meeting tight constraints on latency and memory. In such scenarios, even models with around 10 billion parameters often fall short when high-quality outputs are required.

The First Workshop on **Sma**ll and Efficient **LL**Ms for Knowledge **Ext**raction (**SmaLLEXT**), co-located with the 34th ACM International Conference on Information and Knowledge Management [2] (CIKM 2025[1]), addressed these challenges by focusing on the development and application of compact and efficient LLMs for knowledge and information extraction [3, 4]. Recent advances in model compression [5], quantization [6], pruning [7], retrieval-augmented generation (RAG) [8], optimized prompt selection [9], and efficient fine-tuning [10] have demonstrated that smaller LLMs can achieve competitive performance across a range of downstream tasks [11]. However, research on scalable KE remains fragmented across subfields such as natural language processing [12], information retrieval [13, 14], and knowledge representation [15].

Proceedings of the First Workshop on Small and Efficient Large Language Models for Knowledge Extraction (SmaLLEXT) @ CIKM 2025

felice.merra@cognism.com (F. A. Merra); kristian.skracic@cognism.com (K. Skračić); daniele.malitesta@centralesupelec.fr (D. Malitesta); jacek@distillabs.ai (J. Golebiowski); pminervini@ed.ac.uk (P. Minervini)
0009-0003-8429-3487 (F. A. Merra); 0000-0001-9221-8661 (K. Skračić); 0000-0003-2228-0333 (D. Malitesta); 0000-0001-8053-8318 (J. Golebiowski); 0000-0002-8442-602X (P. Minervini)

[1]https://cikm2025.org.

This workshop aimed to bridge these gaps by bringing together the efficiency and extraction research communities within the CIKM venue, with a focus on data-centric and application-driven methodologies. It emphasized real-world industrial and business applications, where large-scale data is abundant but often noisy, dynamic, and challenging to process reliably. The workshop was designed to convene researchers and practitioners from both academia and industry to explore strategies for compressing, distilling, specializing, and accelerating LLMs, while maintaining high extraction accuracy and robustness against hallucinations. It also examined how these models can support structured data extraction from diverse sources, including web content, enterprise data repositories, and multimodal documents.

2. Objectives

The SmaLLEXT workshop promoted both research and practical developments in compact, resource-efficient language models tailored for knowledge extraction (KE). Although LLMs have significantly advanced capabilities in language understanding and generation, their large size, high computational cost, and latency make them impractical for many real-world uses, especially in scenarios where efficiency, domain adaptation, or deployment on edge devices is essential.

The workshop tailored KE from diverse and complex data sources, including unstructured text, multimodal inputs (such as images, PDFs, and tables), and semi-structured documents. It intended to showcase state-of-the-art contributions in areas like model compression, parameter-efficient fine-tuning, retrieval-augmented generation, and hybrid approaches that combine symbolic and neural reasoning. The overarching goal was to enhance the capability, robustness, and deployability of smaller models for KE tasks. Particular attention was given to methods that ensure factual reliability, enable low-latency inference, and adapt effectively to specialized or resource-limited settings.

A key objective of the workshop was to encourage collaboration between academia and industry by exploring emerging applications across sectors such as e-commerce, healthcare, finance, law, education, and scientific research, where reliable and interpretable KE techniques are essential. Through keynotes, invited talks, and paper sessions, the workshop served as a venue for interdisciplinary exchange among communities in natural language processing, machine learning, knowledge representation, and information retrieval.

3. Topics of Interest

As stated in its call for papers, SmaLLEXT focused on works addressing the topics listed below, as well as other related ones:

- **Efficient Architectures for KE:** Design of compact transformer variants, sparse and modular network, lightweight language models.
- **Model Compression and Acceleration for KE:** Quantization, pruning, distillation, and low-rank adaptation methods, hardware-aware model optimization, real-time acceleration techniques.
- **Prompt Engineering for Extraction:** Zero-shot vs. few-shot templates, chain-of-thought prompt design, iterative prompt refinement, instruction phrasing best practices.
- **Retrieval-Augmented and Hybrid Approaches for KE:** Retrieval-augmented generation (RAG), symbolic reasoning to support distilled LLMs, lightweight memory-augmented models.
- **Specialization and Fine-Tuning for KE:** Domain adaptation and continual fine-tuning, parameter-efficient tuning strategies for adaptation, transfer learning approaches for specialized domains.
- **Depth vs. Breadth in Small-Model for KE:** Trade-offs in model capacity for multi-schema coverage, domain-aware adaptive sparsity patterns, hybrid breadth-depth pipelines (specialist modules and generalist core), evaluation protocols for depth vs. breadth.

- **KE Applications on:** Unstructured text (e.g., named entity recognition, relation extraction, entity linking), semi-structured data (e.g., tables, forms, web pages), multimodal data (e.g., images, PDFs, charts, and scansion).
- **Cross-lingual & Low-Resource Scenarios:** Multilingual transfer and cross-lingual prompting, data augmentation for low-resource languages.
- **Evaluation, Robustness, and Interpretability for KE:** Benchmarks and datasets for evaluating the models, interpretability and explainability, challenges of measuring faithfulness and detecting hallucinations.
- **Annotation Strategies and Data Curation:** Crowd-sourced annotation with guidelines, LLM-assisted active learning loops, weak supervision via heuristic labeling, annotation quality control metrics.
- **Systems and Deployment Considerations:** Industry experience in building KE pipelines, real-world deployments, energy-efficient training and deployment, drift detection in extracted knowledge over time, canary deployments and A/B testing.
- **Adversarial Robustness & Security:** Defense against prompt injections, data-poisoning mitigation, robustness to private data memorization (or extraction).
- **Ethical and Societal Implications:** Fairness and bias in specialized small models, privacy-preserving characteristics of small LLMs in KE, memorization of public datasets/information.

4. Program

SmaLLEXT was held in Seoul (Korea) on November 14, 2025. The workshop featured the following presentations/keynotes:

- **Keynote:** *Automating Neural Network Design: Neural Architecture Search for Small and Efficient Models* [16], by Aaron Klein.
- **Paper session #1:**
  - *Exploiting LLMs to Improve Recommendation Diversity* [17], by Basar Yilmaz (Middle East Technical University), Ali Eren Çankaya (Middle East Technical University), Ismail Sengor Altingovde (Middle East Technical University), Pinar Karagoz (Middle East Technical University), and Ismail Hakki Toroslu (Middle East Technical University).
  - *Exploring Approaches for Detecting Memorization of Recommender System Data in Large Language Models* [18], by Antonio Colacicco (Politecnico di Bari), Vito Guida (Politecnico di Bari), Dario Di Palma (Politecnico di Bari), Fedelucio Narducci (Politecnico di Bari), and Tommaso Di Noia (Politecnico di Bari).
  - *Identifying Business Trends from Stock Market Data utilizing Domain-Specialized Sentence-RoBERTa and BERTopic Technique* [19], by Ye Lim Jung (Korea Institute of Science and Technology Information), and Hyoung Sun Yoo (Korea Institute of Science and Technology Information).
- **Paper session #2:**
  - *CliqueParcel: An Approach For Batching LLM Prompts That Jointly Optimizes Efficiency And Faithfulness* [20], by Jiayi Liu (Purdue University), Tinghan Yang (Purdue University), and Jennifer Neville (Purdue University/Microsoft Research).

  - *Self-Adapted Entity-Centric Data Augmentation for Discontinuous Named Entity Recognition* [21], by Wen-Fang Su (Galaxy Software Services), Hsiao-Wei Chou (National Taiwan University of Science and Technology), and Wen-Yang Lin (National University of Kaohsiung).

- **Industrial Talk:** *Machine Learning Without Data: Training Small Expert Agents With Minimal Examples* [22], by Jacek Golebiowski.

5. Website & Proceedings

All workshop material including schedule and news will be found on the 2025 workshop website at https://sites.google.com/view/smallext/home.

6. Organizers and Program Committee

SmaLLEXT 2025 was organized by:

- **Felice Antonio Merra** (Cognism)
- **Kristian Skračić** (Cognism)
- **Daniele Malitesta** (Université Paris-Saclay, CentraleSupélec, Inria)
- **Jacek Golebiowski** (distil labs)
- **Pasquale Minervini** (University of Edinburgh and Miniml.AI)

The program committee of the workshop was composed by: Matteo Antonio Senese (Amazon), Ingo Gühring (Amazon Web Services), Dario Di Palma (Politecnico di Bari), Lennart Schneider (Amazon Web Services), Angelo Salatino (Knowledge Media Institute), Jiahuan Pei (Vrije Universiteit Amsterdam), Alessandro De Bellis (Politecnico di Bari), Claudio Pomo (Politecnico di Bari).

All submitted papers were peer-reviewed by an international Program Committee composed of experts in the relevant research areas. Each submission was evaluated based on its scientific quality, originality, and relevance to the workshop topics by at least 2 reviewers. The acceptance of papers was based solely on their quality and relevance to the workshop, and not on factors such as the reputation or prior publication record of the authors. Note that the workshop also include one keynote and one invited talk, which were not subject to the same peer-review process. All the the papers in the proceedings were, however, peer-reviewed submissions.

References

[1] D. Xu, W. Chen, W. Peng, C. Zhang, T. Xu, X. Zhao, X. Wu, Y. Zheng, Y. Wang, E. Chen, Large language models for generative information extraction: a survey, Frontiers Comput. Sci. (2024).

[2] M. Cha, C. Park, N. Park, C. Yang, S. B. Roy, J. Li, J. Kamps, K. Shin, B. Hooi, L. He (Eds.), Proceedings of the 34th ACM International Conference on Information and Knowledge Management, CIKM 2025, Seoul, Republic of Korea, November 10-14, 2025, ACM, 2025.

[3] R. Zmigrod, P. Shetty, M. Sibue, Z. Ma, A. Nourbakhsh, X. Liu, M. Veloso, "what is the value of templates?" rethinking document information extraction datasets for llms, in: EMNLP (Findings), ACL, 2024.

[4] X. Huang, M. Surve, Y. Liu, T. Luo, O. Wiest, X. Zhang, N. V. Chawla, Application of large language models in chemistry reaction data extraction and cleaning, in: CIKM, ACM, 2024, pp. 3797–3801.

[5] J. Lin, J. Tang, H. Tang, S. Yang, W. Chen, W. Wang, G. Xiao, X. Dang, C. Gan, S. Han, AWQ: activation-aware weight quantization for on-device LLM compression and acceleration, in: MLSys, mlsys.org, 2024.

[6] G. Xiao, J. Lin, M. Seznec, H. Wu, J. Demouth, S. Han, Smoothquant: Accurate and efficient post-training quantization for large language models, in: ICML, volume 202 of *Proceedings of Machine Learning Research*, PMLR, 2023, pp. 38087–38099.

[7] M. Sun, Z. Liu, A. Bair, J. Z. Kolter, A simple and effective pruning approach for large language models, in: ICLR, OpenReview.net, 2024.

[8] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, Q. Guo, M. Wang, H. Wang, Retrieval-augmented generation for large language models: A survey, 2024. arXiv:2312.10997.

[9] L. Schneider, M. Wistuba, A. Klein, J. Golebiowski, G. Zappella, F. A. Merra, Hyperband-based bayesian optimization for black-box prompt selection, in: ICML, volume 267 of *Proceedings of Machine Learning Research*, PMLR / OpenReview.net, 2025.

[10] Z. Han, C. Gao, J. Liu, J. Zhang, S. Q. Zhang, Parameter-efficient fine-tuning for large models: A comprehensive survey, Trans. Mach. Learn. Res. 2024 (2024).

[11] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, G. Lample, Llama: Open and efficient foundation language models, CoRR abs/2302.13971 (2023).

[12] H. Liu, Q. Yin, Z. Wang, C. Zhang, H. Jiang, Y. Gao, Z. Li, X. Li, C. Zhang, B. Yin, W. Wang, X. Zhu, Knowledge-selective pretraining for attribute value extraction, in: EMNLP (Findings), Association for Computational Linguistics, 2023, pp. 8062–8074.

[13] C. Fang, X. Li, Z. Fan, J. Xu, K. Nag, E. Körpeoglu, S. Kumar, K. Achan, Llm-ensemble: Optimal large language model ensemble method for e-commerce product attribute value extraction, in: SIGIR, ACM, 2024, pp. 2910–2914.

[14] J. Gong, W. Chen, H. Eldardiry, Knowledge-enhanced multi-label few-shot product attribute-value extraction, in: CIKM, ACM, 2023.

[15] G. R. Ghosal, T. Hashimoto, A. Raghunathan, Understanding finetuning for factual knowledge extraction, in: ICML, OpenReview.net, 2024.

[16] A. Klein, Automating neural network design: Neural architecture search for small and efficient models, Keynote at First Workshop on Small and Efficient Large Language Models for Knowledge Extraction (SmaLLEXT @ CIKM2025), 2025.

[17] B. Yilmaz, A. E. Cankaya, I. S. Altingovde, P. Karagoz, I. H. Toroslu, Exploiting llms to improve recommendation diversity, in: Proceedings of the First Workshop on Small and Efficient Large Language Models for Knowledge Extraction (SmaLLEXT @ CIKM2025), 2025.

[18] A. Colacicco, V. Guida, D. Di Palma, F. Narducci, T. Di Noia, Exploring approaches for detecting memorization of recommender system data in large language models, in: Proceedings of the First Workshop on Small and Efficient Large Language Models for Knowledge Extraction (SmaLLEXT @ CIKM2025), 2025.

[19] Y. L. Jung, H. S. Yoo, Identifying business trends from stock market data utilizing domain-specialized sentence-roberta and bertopic technique, in: Proceedings of the First Workshop on Small and Efficient Large Language Models for Knowledge Extraction (SmaLLEXT @ CIKM2025), 2025.

[20] J. Liu, T. Yang, J. Neville, Cliqueparcel: An approach for batching llm prompts that jointly optimizes efficiency and faithfulness, in: Proceedings of the First Workshop on Small and Efficient Large Language Models for Knowledge Extraction (SmaLLEXT @ CIKM2025), 2025.

[21] W.-F. Su, H.-W. Chou, W.-Y. Lin, Self-adapted entity-centric data augmentation for discontinuous named entity recognition, in: Proceedings of the First Workshop on Small and Efficient Large Language Models for Knowledge Extraction (SmaLLEXT @ CIKM2025), 2025.

[22] M. Gryka, V. R. Guda, G. Kadlecová, B. Kruszczyński, S. Nowicki, J. Verschueren, U. Zafar, Machine learning without data: Training small expert agents with minimal examples, Industrial Talk at First Workshop on Small and Efficient Large Language Models for Knowledge Extraction (SmaLLEXT @ CIKM2025), 2025.