

Edge Knowledge Bases for Sustainable and Explainable AI

Aparna S. Varde¹, Emil Njor¹, Francesco Daghero², Mahyar Turchi Moghaddam²

1. Center for Software Technology, SDU Vejle

2. Dept. of Software Engineering, SDU Odense

MMMI, University of Southern Denmark (SDU), Denmark

ORCID: 0000-0002-3170-2510 (Varde), 0000-0003-0219-1624 (Njor),

0000-0001-9595-7216 (Daghero), 0000-0001-5028-7546 (Moghaddam)

{apva,njor,fdag,mtmo}@mmmi.sdu.dk

Abstract

We propose a vision of **EdgeKBs** (Edge Knowledge Bases) integrating KBs into Edge AI systems. EdgeKBs can infuse commonsense and domain knowledge within Edge devices to guide them while learning from training data. By doing so, we not only bring explainability and performance benefits of KBs to the Edge, but also enable more efficient, sustainable Edge AI systems. We envisage three approaches to develop EdgeKBs: Slicing & Dicing; Tailoring to Fit; and Dynamic EdgeKBs. These EdgeKBs will bring Knowledge-Guided Machine Learning (KGML) to the Edge to support next-generation applications, e.g., Physical AI, cobots, and local natural language interfaces.

1 Introduction

Background: As AI expands beyond centralized cloud environments to embedded & network-edge settings, performing inference near data sources can deliver quality gains: reduced latency, greater bandwidth efficiency, and enhanced privacy (Varghese et al., 2016). Advances in knowledge-intensive natural language processing (NLP) show that explicit external memory can improve factuality & provenance compared to purely parametric models (well-known in **Retrieval Augmented Generation (RAG)**) (Lewis et al., 2020), (Wu et al., 2022). More generally, the paradigm *Knowledge-Guided Machine Learning (KGML)* “aims to use both scientific knowledge and data in ML frameworks to achieve better generalizability, scientific consistency and explainability” (Karpatne et al., 2026). KGML studies (Serra et al., 2025), (Austin-Gabriel et al., 2025) add interpretability and raise accuracy as well as efficiency (enhancing sustainability).

Related Work: In KBs, approaches such as the classical Cyc (Lenat, 1995) and WordNet (Fellbaum, 1998) as well as the recent ConceptNet 5.5 (Speer et al., 2017) for commonsense knowledge (CSK) and CANDLE for cultural CSK (Nguyen

et al., 2023), indicate that structured knowledge can be curated at scale. CSK extraction, compilation and use in machine intelligence can be vital (Razniewski et al., 2021), (Tandon et al., 2018). Yet, KB design is rarely co-designed with Edge deployment constraints (memory, latency, maintenance). Emerging RAG frameworks well-optimize local vector retrieval in text generation, but are bounded by unstructured data representations and passive storage (Kang et al., 2026; Solanki et al., 2026; Qin et al., 2024; Zhu et al., 2024). Recent work (Liu et al., 2022) covers resource constraints in physics-informed neural networks.

Novelty and Contributions: Motivated by the state-of-the-art and excited by the gaps to be filled, this vision paper proposes to build **EdgeKB** (Edge Knowledge Bases) to enable **EdgeKGML** (Knowledge-Guided Machine Learning on Edge devices). Such EdgeKBs must provide a small, focused, practical system for managing knowledge that works well on Edge devices. Building these new EdgeKBs entails novel paradigms of Slice and Dice, Build from Scratch, and Dynamic EdgeKBs (elaborated in the paper). They contribute to sustainable and explainable AI for knowledge extraction, storage, reasoning, and updating for devices with limited resources, e.g. cobots, inbolls, NLP-based virtual assistants, and other next-generation systems. This initiative can enable researchers to address further issues such as:

(1). **Model Compression** : Can large-scale commonsense and domain knowledge be compressed in compact EdgeKBs without sacrificing relational information for reliable local inference?

(2). **EdgeKB Maintenance**: How should EdgeKBs be updated, valued, and orchestrated to enhance latency, privacy, robustness, and explainability under resource constraints?

We elaborate on our proposed approaches and experimental plans in this vision paper.

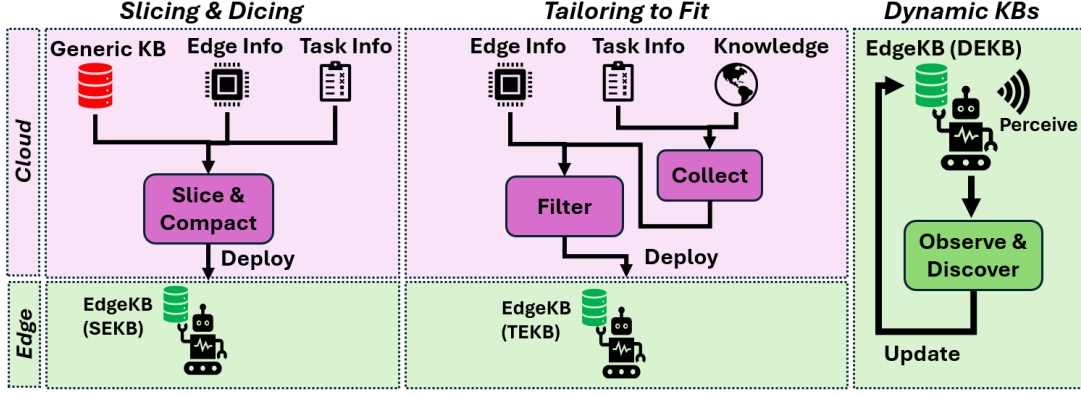


Figure 1: Overview of proposed approaches for EdgeKB development.

2 Proposed EdgeKB Approaches

We propose three archetype approaches to construct EdgeKBs. They are not meant to be mutually exclusive, e.g., some EdgeKBs might be created by their combination. See Figure 1 for a visualization of the approaches.

2.1 Approach 1: Slicing and Dicing

As a first approach, we propose to slice one or more existing KBs to derive an EdgeKB that fits within Edge device constraints and context-dependencies. We call the resulting KB a **Sliced EdgeKB (SEKB)**. It can be built by one of the following methods.

(i) **Deploy a Multifaceted KB:** A multifaceted KB such as Dice (Chalier et al., 2020b,a) can be applied to slice commonsense KBs (ConceptNet (Speer et al., 2017), WebChild (Tandon et al., 2014), Quasimodo (Romero et al., 2019), CANDLE (Nguyen et al., 2023) etc.) or domain KBs (ChEBI (Degtyarenko et al., 2007), GeneOntology (Aleksander et al., 2023) etc.) Dice has clauses (Plausibility, Typicality, Remarkability, Saliency) helpful to curate knowledge from other KBs.

$$\rho(e_1, p) \wedge \lambda(e_1, e_2) \wedge \neg \rho(e_2, p) \implies \mu(e_1, p) \quad (1)$$

$$\tau(e, p) \wedge \mu(e, p) \implies \sigma(e, p) \quad (2)$$

Consider the above equations (1) and (2) from Dice, potentially useful to build the SEKB by slicing. Here, ρ means *plausible*, λ is *related*, μ is *remarkable*, τ is *typical*, σ is *salient*; e, e_1, e_2 are entities, p being properties. These equations can express facts, e.g., in domestic robotics, if it is *plausible* to find an *air-fryer* item as a property of a *kitchen* and *not plausible* to find it in a related entity *laundry room*, an *air-fryer* is *remarkable* for a *kitchen*. By calculating scores for each clause, aggregated scores can be computed to extract info from other

KBs with reference to context, e.g., retain info on "salient" entities only or on all "plausible" ones.

While we exemplify the method with Dice, other multifaceted KBs, e.g., Octopus (Yang et al., 2002) can be used as well.

(ii) **Harness LLMs:** Large language models such as GPT-4o (OpenAI, 2024), and DeepSeek-R1 (Guo et al., 2025) can also be applied for slicing existing Knowledge Bases (KBs). We can conduct LLM slicing using either of the following options.

(a) *Semantic Chunking:* This is a method analogous to semantics-preserving RAG chunks (Puspi-tasari et al., 2026). Here, relevant KBs are passed to a given LLM with instructions to split them based on general concepts, e.g., "everything relevant to domestic robotics". Hence, this method can slice big chunks of knowledge for the SEKB.

(b) *Knowledge Graph Extraction:* This method is similar to Microsoft GraphRAG (Edge et al., 2024). Here, an LLM can be prompted to identify more specific entities and relationships, thus knowledge slices can be mapped into structured SPO (subject-predicate-object) triples. LLM prompts can mention specific entities, e.g., only "kitchen", "bedroom", or "living room", along with specific relationships, e.g., "containment". It can then extract SPO triples such as "refrigerator cools food", i.e., $S=\text{refrigerator}, P=\text{cooling}, O=\text{food}$. The most relevant SPO triples can then be stored in the SEKB.

2.2 Approach 2: Tailoring to Fit

Edge AI applications often work in unique environments which necessitate tailored models & data to achieve peak performance (Njor et al., 2024). Thus, while some EdgeKBs can be extracted from slices of existing traditional-scale KBs, there is often a need for an EdgeKB tailor-made from scratch for the Edge AI to fit unique characteristics of applica-

tions. We call this a **Tailored EdgeKB (TEKB)**.

One primary driver for tailoring a new EdgeKB from scratch will likely be the lack of an appropriate existing KBs to adapt. Perhaps the application is one that only exists for Edge AI, such as an AI system deployed on an IoT sensor, which judges whether captured data is simple enough to be processed locally, or whether it is so complex that it must be sent to a larger machine. Another example can be EdgeKBs for wearable devices in medicine (e.g., intelligent pacemakers) or drones on the Edge with natural language interaction.

In some cases, KBs suitable for the application may exist, but they are constructed or licensed in a way that prevents slicing them to fit Edge AI constraints. In other cases, designing an EdgeKB from scratch will allow for faster development. No matter the reason, tailoring an EdgeKB to fit the needs of the application enables highly domain-specific, specialized data structures and custom semantics to reduce data storage and resource consumption.

We aim to distinguish between two TEKB types:

(i) **Personalized EdgeKB:** This type of TEKB is bespoke to a specific Edge AI system. It tightly fits device constraints, and is designed to leverage any specialized hardware for maximum performance.

(ii) **General EdgeKB:** This TEKB type, on the other hand, is designed for the Edge, but not necessarily a specific Edge AI system. It will be resource-efficient, but will follow ecosystem-wide norms rather than system-specific constraints & hardware.

We propose that tailoring an EdgeKB from scratch (TEKB) entails the following. (1). Define constraints of the Edge AI system (Maximum memory use, energy consumption, latency). (2). Evaluate which data structures and custom semantics can support better resource efficiency for the EdgeKB. (3). Address whether the EdgeKB will be able to query other KBs on a network, or if it must act autonomously. (4). Consider how AI systems and users will query the EdgeKB, e.g., through natural language interfaces. (5). Decide whether the EdgeKB should be personalized or general-purpose.

2.3 Approach 3: Dynamic EdgeKBs

Rather than remaining static artifacts, some EdgeKBs should be able to adapt to environments in which they are deployed. Enabled by increasingly capable Edge AI models, e.g., Small Language Models (SLMs) (Wang et al., 2025) and lightweight foundation models, Edge devices may

evolve by continuously identifying, deriving, organizing and refining knowledge to maintain up-to-date information for their goals. We call the evolved KB a **Dynamic EdgeKB (DEKB)**, and claim that this evolution can occur through two complementary paradigms as explained next.

(i) **KG refinement:** This paradigm of refining a knowledge graph (KG) entails dynamically adapting the respective KBs initially obtained offline, updating outdated information, refining existing relations, and incorporating context-dependent variations observed during operation. It preserves the original KB structure while ensuring that its content gets aligned with the current environment.

(ii) **KG derivation:** In contrast, this other paradigm of deriving a knowledge graph delegates management of the given KB entirely to the Edge AI system. The Edge model autonomously discovers new entities and relations, creates application-specific abstractions, validates knowledge against local observations, and continuously restructures the KB as per evolving requirements.

2.4 Cross-Approach Considerations

Rather than storing all acquired information locally, often unfeasible due to the devices' resource constraints, some Edge AI systems may be connected to larger KBs over a network. Knowledge that is often accessed or essential for real-time decision-making or privacy may be retained locally, while less relevant information may be compressed, summarized, or migrated to nearby fog nodes and cloud repositories.

Conversely, when new tasks emerge, relevant knowledge may be proactively retrieved from remote nodes and materialized on-device before it is required. Such adaptive placement strategies effectively transform knowledge into a managed resource, analogous to memory or computation in traditional edge systems.

3 EdgeKBs for EdgeKGML

By introducing EdgeKBs to Edge AI, we can bring many benefits that other AI paradigms have harvested to the Edge. Future Edge AI systems supported by EdgeKBs will be more trustworthy & explainable, as their decisions will be supported by verified information. We also expect that EdgeKBs will make Edge AI achieve better predictive performance. Uniquely for Edge AI, this increase in predictive performance can be traded for resource

Table 1: Evaluation Metrics for Edge Knowledge-Guided Machine Learning Systems

Category	Metric	Description	Unit
Hardware	Static Memory (Storage)	Disk/Flash space required to store AI model parameters and the KB index.	MB / GB
	Dynamic Memory (RAM)	Peak working memory consumption during concurrent KB retrieval and model inference.	MB / GB
	Energy Consumption	Total energy consumed per query/batch, critical for battery-constrained edge devices.	mJ / J
	Latency	End-to-end processing time from input reception to final output generation.	ms
AI Model	Accuracy	Proportion of correct predictions or classifications out of all evaluated samples.	%
	F1-Score / Precision	Performance metrics for robustness, especially useful for imbalanced edge datasets.	0–1
Unified	Storage Ratio (KB :AI)	Ratio of static memory footprint of the KB versus the AI Model parameters.	Ratio
	Compute Ratio (KB :AI)	Proportion of processing time (or FLOPs) spent on KB retrieval versus AI model inference.	Ratio
	KB Hit/Relevance Rate	Percentage of inferences where the KB successfully provided useful context.	%

efficiency by using smaller AI models (Banbury et al., 2024), allowing EdgeKBs to make Edge AI more sustainable. By pushing structured, adaptable knowledge directly to resource-constrained devices, EdgeKBs unlock a new class of applications that fundamentally improve performance, trustworthiness, and sustainability of local AI deployments. Such applications can potentially entail EdgeKGML, deploying knowledge from EdgeKBs. We synopsise some of these applications here.

Physical AI: While current Edge AI systems can accurately detect & track objects (Yrjänäinen et al., 2020), adequate interaction with the physical world needs a deeper understanding of affordances (Zhang et al., 2016), task constraints, and safety regulations. In domains like manufacturing, logistics, and warehouse automation, EdgeKBs can bridge this gap. Crucially, this enables next-gen Physical AI, e.g., domestic cobots & motion-tracking inbolts (Kolvig-Raun, 2026). Embedding EdgeKBs (commonsense, domain knowledge) in such systems can bring them closer to thresholds of human cognition.

Edge Natural Language Interfaces: Natural language interfaces operating in high-stakes domains, e.g. industrial machinery, cannot rely solely on generic LLMs (Zheng et al., 2025) that may be unavailable at critical moments. EdgeKBs can enable deployment of highly specialized conversational agents, intrinsically grounded in specific maintenance workflows and safety protocols.

Private Smart Assistants: Smartphone-based assistants (e.g., text-based / voice-operated (Sateli et al., 2013), (Kalvakurthi et al., 2023)) can demon-

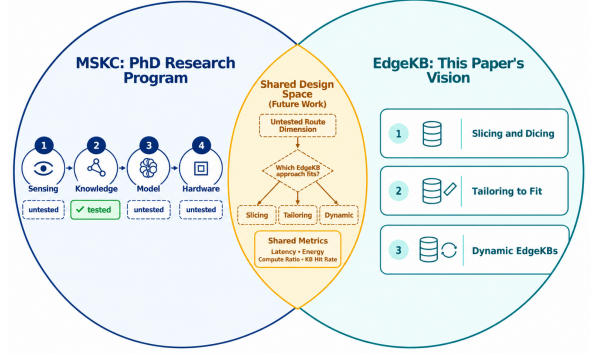


Figure 2: Experimentation Example — MSKC (minimum sufficient knowledge computation) in targeted applications, e.g., decision support of wind turbines for renewable energy, in conjunction with EdgeKB development: Knowledge routing is partly tested, the rest of the parts can be mounted with EdgeKB components.

strate the immense utility that can be offered by dynamic EdgeKBs in the future. Phones can be initialized with a sliced commonsense knowledge base alongside a compact, domain-specific repository. Runtime adaptation can then build a deeply personalized knowledge layer that can capture user routines, preferences, and frequent interactions.

4 Experimentation Plans

The intersection of Edge AI and knowledge bases can offer ideal opportunities to study the benefits they can achieve together. As EdgeKBs mature, we envision promising research trajectories as follows.

- Investigating how EdgeKBs improve Edge AI

system performance by grounding decisions in real information, reducing hallucinations.

- Applying EdgeKBs to improve resource efficiency, e.g., trade-off in performance gains achieved by EdgeKBs vs. efficiency due to less computationally intensive models.
- Investigating how EdgeKBs constructed on the Edge can pave the way for decentralized construction of larger Cloud KBs
- Exploring how large-scale commonsense & domain knowledge can be compressed into compact EdgeKBs without sacrificing relational info needed for reliable local inference.
- Studying how EdgeKBs updated over time offer privacy, robustness, explainability, and low latency, yet respecting resource constraints.

We aim to evaluate solutions across these trajectories, especially with respect to the metrics in Table 1. An example of an outline for our experimentation appears in Figure 2, indicating how ongoing work on the research sub-problem of a PhD student (Sengupta et al., 2026a) furthering from prior work (Sengupta et al., 2026b) can blend with the EdgeKB vision in this paper.

5 Conclusions and Roadmap

Summary: In this paper, have presented our proposed vision of EdgeKBs being compact, curated knowledge bases co-designed with the resource constraints of Edge devices. We have outlined three complementary paths to building them: (i) slicing existing large-scale KBs, (ii) tailoring new ones from scratch, and (iii) allowing deployed KBs to evolve dynamically. We have further illustrated how EdgeKBs can pave the way to potentially help in bringing Knowledge-Guided Machine Learning to the Edge (i.e., EdgeKGML) from Physical AI to local language interfaces.

The benefits of grounding local inference in explicit, verifiable knowledge are that Edge AI becomes more explainable and trustworthy, while the resulting accuracy gains can be traded for smaller models, thus lowering energy and memory demands; thereby making Edge AI more sustainable. We see offline EdgeKB construction as a stepping stone towards inference-time KBs being assembled on the fly, as time progresses.

Usefulness: It is well-known that data is ever-growing, e.g., a Quettabyte (QB), i.e., a decillion bytes (10^{30} bytes) has already been officially de-

clared through SI Units in 2022 to future-proof global data measurement. Much of the data is multimodal with natural language text, images, video etc. While the big data community aims to build accurate, efficient methods to analyze enormous volumes of data, there are parallel efforts in sustainable AI to achieve the best results with minimal resources, e.g. operating on resource-constrained devices. Likewise, there are trends in adding explainability in learning for twofold contributions: better interpretability and resource-saving. The latter occurs by infusing explanations in learning, derived from KBs (one-time process vs. recurrent learning from huge data). Hence, with the proposed approaches in this paper, we aim to promote KGML along with Edge AI with EdgeKBs.

Our vision paper can be useful in terms of introducing the EdgeKB problem setting, proposing development entailing knowledge base slicing, Edge-native KB construction, and adaptive maintenance. It can envision managing knowledge as a system resource together with computation and memory in next-generation Edge AI, keeping in mind the constantly increasing data and knowledge as we head towards next-generation systems. Our work identifies high-impact application domains in cyber-physical and conversational contexts so as to outline a roadmap from conceptualization to practical feasibility.

We invite the NLP, KB and Edge Computing communities to pursue our research agenda. We anticipate that the approaches, experiments and metrics proposed in this vision paper will be exciting to these research communities.

6 Acknowledgements

Aparna Varde acknowledges a research grant NF25OC011003 from NNF, Denmark. Francesco Daghero and Mahyar Turchi Moghaddam acknowledge the REINS project, supported by Next-Generation AI (Forskningreserven) and Digital Research Centre Denmark (DIREC). We thank Purbak Sengupta, a PhD candidate at CST, MMMI, SDU Vejle, for his inputs related to an illustration in this paper. All the authors are at the Maersk McKinney Moller Institute (MMMI) in the Faculty of Engineering (TEK) at the University of Southern Denmark (SDU).

References

- Suzi A Aleksander, James Balhoff, Seth Carbon, J Michael Cherry, Harold J Drabkin, Dustin Ebert, Marc Feuermann, Pascale Gaudet, Nomi L Harris, and 1 others. 2023. The gene ontology knowledge-base in 2023. *Genetics*, 224(1):iyad031.
- Blessing Austin-Gabriel, Aparna S Varde, and Hao Liu. 2025. Geoinformatics-guided machine learning for power plant classification. In *AAAI Conference on Artificial Intelligence (Bridge Program)*.
- Colby Banbury, Emil Njor, Andrea Mattia Garavagno, Mark Mazumder, Matthew Stewart, Pete Warden, Manjunath Kudlur, Nat Jeffries, Xenofon Fafoutis, and Vijay Janapa Reddi. 2024. Wake vision: A tailored dataset and benchmark suite for tinyml computer vision applications. *arXiv preprint arXiv:2405.00892*.
- Yohan Chalier, Simon Razniewski, and Gerhard Weikum. 2020a. *Dice: A joint reasoning framework for multi-faceted commonsense knowledge*. In *International Semantic Web Conference (ISWC Demo)*, volume 2721, pages 16–20.
- Yohan Chalier, Simon Razniewski, and Gerhard Weikum. 2020b. *Joint reasoning for multi-faceted commonsense knowledge*. In *Conference on Automated Knowledge Base Construction, AKBC 2020, Virtual, June 22-24, 2020*.
- Kirill Degtyarenko, Paula De Matos, Marcus Ennis, Janna Hastings, Martin Zbinden, Alan McNaught, Rafael Alcántara, Michael Darsow, Mickaël Guedj, and Michael Ashburner. 2007. Chebi: a database and ontology for chemical entities of biological interest. *Nucleic acids research*, 36(suppl_1):D344–D350.
- Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. 2024. From local to global: A graph rag approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*.
- Christiane Fellbaum. 1998. Wordnet. *WordNet An Electronic Lexical Database*, page 69.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.
- Vishesh Kalvakurthi, Aparna S Varde, and John Jenq. 2023. Hey dona! can you help me with student course registration? In *AAAI Conference on Artificial Intelligence (Workshop on AI for Education)*.
- Dong Ho Kang, Hyunjoon Lee, Hyeonjeong Cha, Minkyu Choi, and Sungsoo Lim. 2026. Pocket rag: On-device rag for first aid guidance in offline mobile environment. *arXiv preprint arXiv:2602.13229*.
- Anuj Karpatne, Xiaowei Jia, and Vipin Kumar. 2026. Knowledge-guided machine learning: A paradigm shift in ai for science. In *AAAI Conference on Artificial Intelligence*, page 40(46):39715–39717.
- Emil Stubbe Kolvig-Raun. 2026. *Motion guidance tracking inbolt*.
- Douglas B Lenat. 1995. Cyc: A large-scale investment in knowledge infrastructure. *Communications of the ACM*, 38(11):33–38.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.
- Ziyue Liu, Xinling Yu, and Zheng Zhang. 2022. *Tt-pinn: A tensor-compressed neural pde solver for edge computing*. *Preprint*, arXiv:2207.01751.
- Tuan-Phong Nguyen, Simon Razniewski, Aparna S. Varde, and Gerhard Weikum. 2023. *Extracting cultural commonsense knowledge at scale*. In *ACM Web Conference 2023, WWW*, pages 1907–1917. ACM.
- Emil Njor, Mohammad Amin Hasanpour, Jan Madsen, and Xenofon Fafoutis. 2024. A holistic review of the tinyml stack for predictive maintenance. *IEEE Access*, 12:184861–184882.
- OpenAI. 2024. *Gpt-4o system card*. *Preprint*, arXiv:2410.21276.
- Fachrina Dewi Puspitasari, Chaoning Zhang, Jiaquan Zhang, Zhicheng Wang, Hafiz Shakeel Ahmad Awan, Rizwan Qureshi, Jewon Lee, Tae-Ho Kim, and Yang Yang. 2026. Efficient rag with intent-aware retrieval and semantics-preserving chunking. *arXiv preprint arXiv:2606.01240*.
- Ruiyang Qin, Zheyu Yan, Dewen Zeng, Zheng Jia, Dancheng Liu, Jianbo Liu, Ahmed Abbasi, Zhi Zheng, Ningyuan Cao, Kai Ni, and 1 others. 2024. Robust implementation of retrieval-augmented generation on edge-based computing-in-memory architectures. In *Proceedings of the 43rd IEEE/ACM International Conference on Computer-Aided Design*, pages 1–9.
- Simon Razniewski, Niket Tandon, and Aparna S Varde. 2021. Information to wisdom: Commonsense knowledge extraction and compilation. In *Proceedings of the 14th ACM international conference on web search and data mining*, pages 1143–1146.
- Julien Romero, Simon Razniewski, Koninika Pal, Jeff Z. Pan, Archit Sakhadeo, and Gerhard Weikum. 2019. Commonsense properties from query logs and question answering forums. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 1411–1420.

- Bahar Sateli, Gina Cook, and René Witte. 2013. Smarter mobile apps through integrated natural language processing services. In *International Conference on Mobile Web and Information Systems*, pages 187–202. Springer.
- Purbak Sengupta, Aparna Varde, and Emil Njor. 2026a. Sustainable event-triggered knowledge-guided intelligence for decision support. Technical report, University of Southern Denmark.
- Purbak Sengupta, Aparna S. Varde, and Emil Njor. 2026b. Carbon-aware adaptive knowledge injection for sustainable, trustworthy ai in resource-constrained predictive modelling. In *Nordic AI Meet*.
- Angela Serra, Michele Fratello, Antonio Federico, and Dario Greco. 2025. An update on knowledge graphs and their current and potential applications in drug discovery. *Expert opinion on drug discovery*, 20(5):599–619.
- Om Solanki, Lopamudra Praharaj, Deepti Gupta, and Maanak Gupta. 2026. Policy-aware edge llm-rag framework for internet of battlefield things mission orchestration. *arXiv preprint arXiv:2604.09493*.
- Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31, No. 1.
- Niket Tandon, Gerard De Melo, Fabian Suchanek, and Gerhard Weikum. 2014. Webchild: Harvesting and organizing commonsense knowledge from the web. In *ACM WSDM international conference on Web search and data mining*, pages 523–532.
- Niket Tandon, Aparna S Varde, and Gerard de Melo. 2018. Commonsense knowledge in machine intelligence. *ACM SIGMOD Record*, 46(4):49–52.
- Blesson Varghese, Nan Wang, Sakil Barbhuiya, Peter Kilpatrick, and Dimitrios S Nikolopoulos. 2016. Challenges and opportunities in edge computing. In *2016 IEEE international conference on smart cloud (SmartCloud)*, pages 20–26. IEEE.
- Fali Wang, Zhiwei Zhang, Xianren Zhang, Zongyu Wu, Tzuhao Mo, Qiuhaio Lu, Wanjing Wang, Rui Li, Junjie Xu, Xianfeng Tang, and 1 others. 2025. A comprehensive survey of small language models in the era of large language models: Techniques, enhancements, applications, collaboration with llms, and trustworthiness. *ACM Transactions on Intelligent Systems and Technology*, 16(6):1–87.
- Yuxiang Wu, Yu Zhao, Baotian Hu, Pasquale Minervini, Pontus Stenetorp, and Sebastian Riedel. 2022. An efficient memory-augmented transformer for knowledge-intensive NLP tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5184–5196, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Jun Yang, Qing Li, and Yueting Zhuang. 2002. Octopus: aggressive search of multi-modality data using multifaceted knowledge base. In *Proceedings of the 11th international conference on World Wide Web*, pages 54–64.
- Jukka Yrjänäinen, Xingyang Ni, Bishwo Adhikari, and Heikki Huttunen. 2020. Privacy-aware edge computing system for people tracking. In *2020 IEEE International Conference on Image Processing (ICIP)*, pages 2096–2100. IEEE.
- Zuo Zhang, Yaoru Sun, and Glyn W Humphreys. 2016. Perceiving object affordances through visual and linguistic pathways: A comparative study. *Scientific reports*, 6(1):26806.
- Yue Zheng, Yuhao Chen, Bin Qian, Xiufang Shi, Yuanchao Shu, and Jiming Chen. 2025. A review on edge large language models: Design, execution, and applications. *ACM Computing Surveys*, 57(8):1–35.
- Xunyu Zhu, Jian Li, Yong Liu, Can Ma, and Weiping Wang. 2024. A survey on model compression for large language models. *Transactions of the Association for Computational Linguistics*, 12:1556–1577.