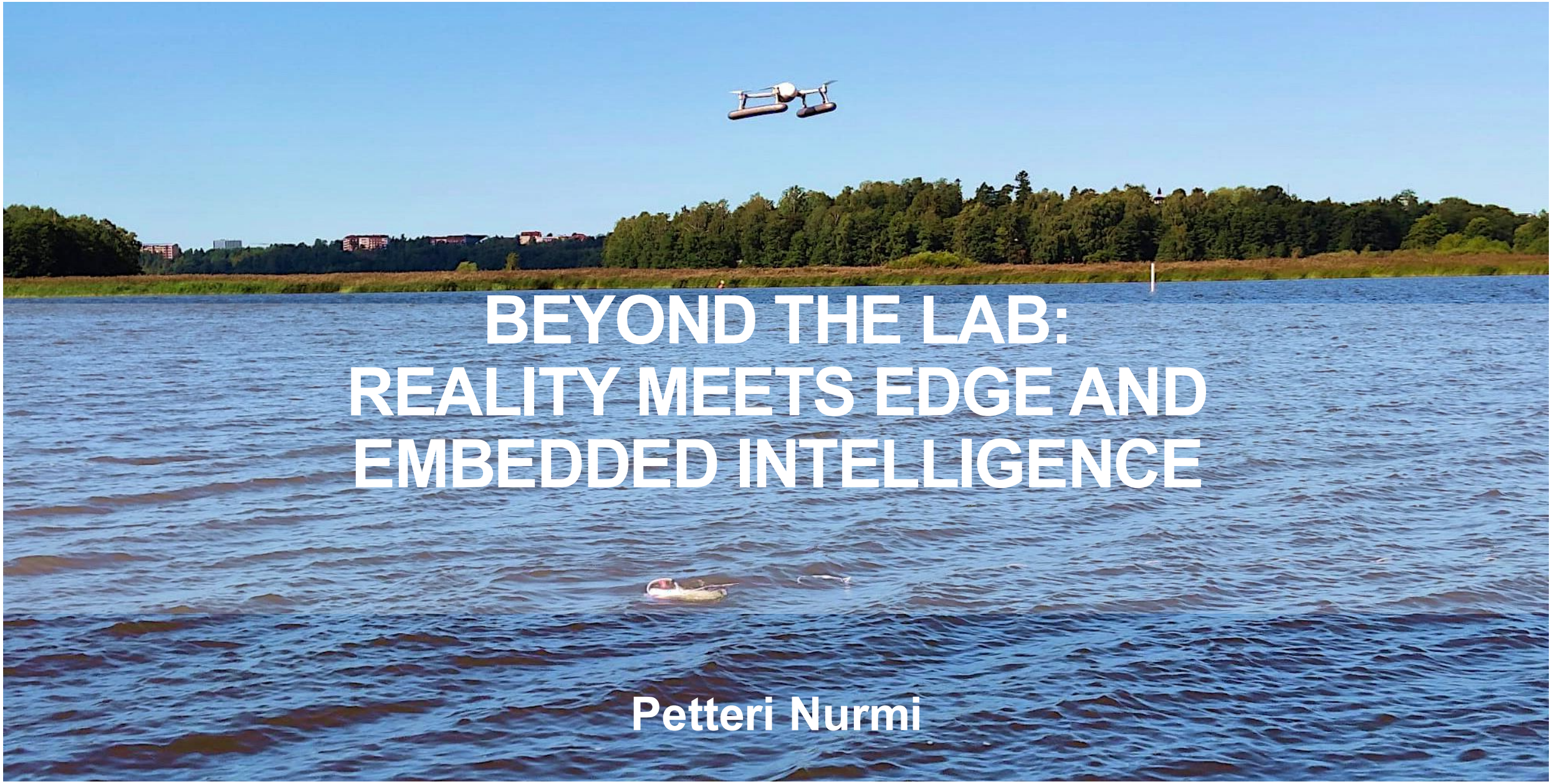


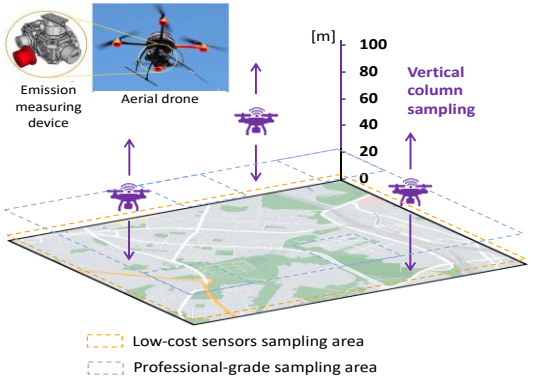
BEYOND THE LAB: REALITY MEETS EDGE AND EMBEDDED INTELLIGENCE

Petteri Nurmi

Internet of Things and Empirics

- Data key to empirical research
- Traditionally based on scientific measurement tools calibrated for measurements
- Modern data increasingly collected using sensor-enabled devices

IoT appliances



Drones (ground, aerial, underwater)

Smartphones and wearables

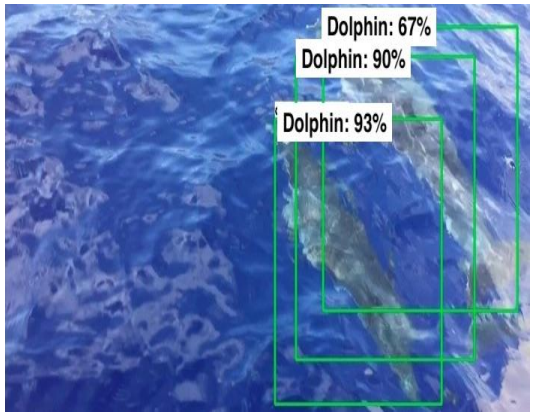
Dedicated Sensor Platforms

Artificial Intelligence of Things (AIoT)

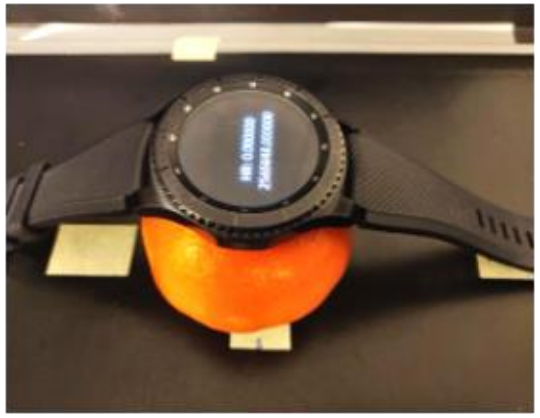
- AI will inevitably start shifting directly to the devices (or the edge)
  1. **Resilience**: Systems can function even without connectivity
  2. **Network Efficiency**: Lower latency, lower bandwidth congestion
  3. **Privacy**: data does not need to leave the devices
  4. **Numbers:** more “edge” devices in the world than human users

- AIoT refers to this integration of AI and Internet of Things

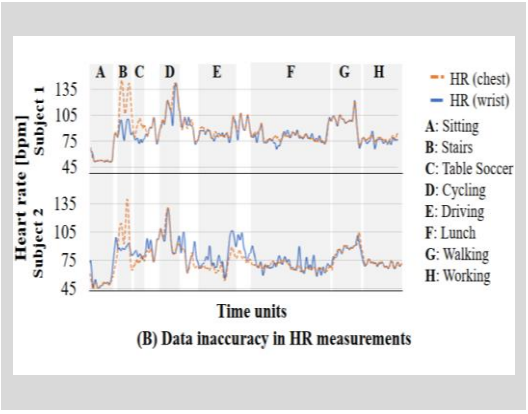
The Why



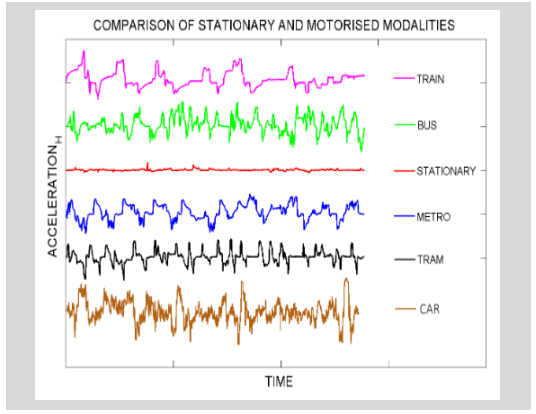
Resilience



Network Efficiency



Privacy



Device Numbers

LLMs to SLMs: From the cloud to the edge

- LLMs among the most complex models and represent the pinnacle of modern AI
- Small Language Models (**SLMs**) bring LLM capacities from the cloud to edge/embedded execution
  - < 10B parameters
  - Often quantized to fit limited memory / storage
  - Can also be compressed variants of larger models
  - Best suited for tasks with a clear structure and formalism

IEEE ComSoc IEEE Open Journal of the Communications Society

Received 13 January 2026; revised 2 March 2026; accepted 23 March 2026. Date of publication 2 April 2026; date of current version 13 April 2026.

Digital Object Identifier 10.1109/OJCOMS.2026.3680064

LLMs and IoT: A Comprehensive Survey on Large Language Models and the Internet of Things

FATEMEH SARHADDI[1,*], NGOC THI NGUYEN[2,*], AGUSTIN ZUNIGA[1,*], MUSFIRA KHAN[1], PAN HUI[1,3] (Fellow, IEEE), SASU TARKOMA[1] (Senior Member, IEEE), HUBER FLORES[4] (Member, IEEE), AND PETTERI NURMI[1]

[1]Department of Computer Science, University of Helsinki, 00014 Helsinki, Finland
[2]National University of Singapore, Singapore
[3]Centre for Metaverse and Computational Creativity, Hong Kong University of Science and Technology (Guangzhou), Guangzhou 511458, China
[4]Institute of Computer Science, University of Tartu, 50409 Tartu, Estonia

CORRESPONDING AUTHOR: F. SARHADDI (fatemeh.sarhaddi@helsinki.fi)

This work was supported by the Academy of Finland under Grant 339614.

*Fatemeh Sarhaddi, Ngoc Thi Nguyen, and Agustin Zuniga contributed equally to this work.

ABSTRACT The Internet of Things (IoT) has emerged as a technological cornerstone, enabling highly interconnected systems that integrate billions of diverse devices. The combined data output of these devices is estimated at hundreds of exabytes per day. The rise of Large Language Models (LLMs) presents new opportunities for managing these devices, processing the data they generate, and addressing the technological and societal challenges posed by IoT ecosystems. However, the integration of LLMs into IoT also introduces significant challenges. This article offers the most comprehensive survey to date on the integration of LLMs into IoT ecosystems. We cover various technical layers, including devices, software engineering, sensing, networking, data processing, privacy and security, and human interaction. We synthesize insights from more than 300 articles, identify open research gaps, and highlight promising future directions, ranging from semantic reasoning and communication to the integration of quantum computing. By addressing the characteristics, challenges, and opportunities associated with LLM integration, this article serves as a foundational resource for researchers and practitioners aiming to enhance IoT functionalities with LLMs while navigating the associated complexities.

INDEX TERMS Large language models (LLMs), Internet of Things (IoT), artificial intelligence, AI, survey.

I. INTRODUCTION

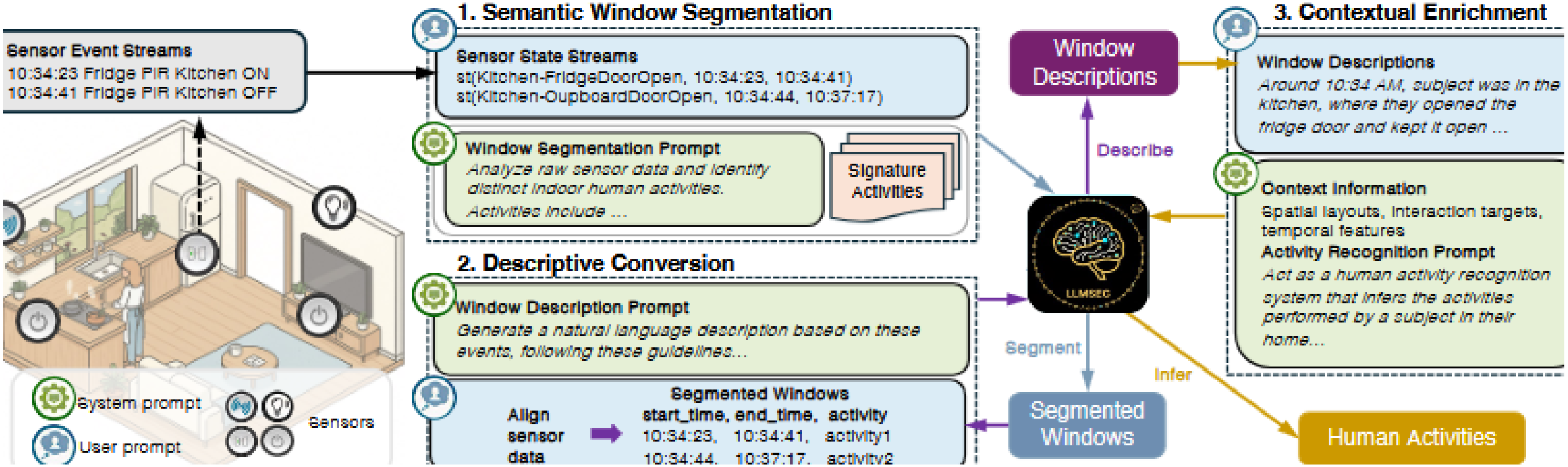
THE rapid proliferation of the Internet of Things (IoT) has led to the emergence of highly interconnected systems that integrate a wide range of devices. These systems provide significant benefits across various sectors, including environmental monitoring, healthcare, agriculture, and smart living environments [1]. Current estimates suggest that the number of connected IoT devices has surpassed 18 billion, with forecasts predicting 37.1 billion devices by 2032 [2]. Managing the vast network of IoT devices and their data is increasingly complex, especially in large-scale ecosystems. Conventional computing methods often struggle to cope with the diversity of IoT devices, the sheer volume of data, and the real-time demands of these environments. Consequently, there is a pressing need for advanced approaches capable of efficiently managing massive volumes of heterogeneous IoT data while facilitating intelligent and context-aware interactions.

Large Language Models (LLMs), recognized for their advanced natural language processing capabilities [3], have emerged as powerful tools for facilitating natural and context-aware interactions. These sophisticated neural networks, pre-trained on extensive datasets, can understand and generate human-like language. Their proficiency in processing large datasets allows tasks such as text summarization, analysis, and generation. These capabilities open up opportunities for meaningful conversations, personalized recommendations, and enhanced decision-making [4].

The integration of LLMs has the potential to enhance IoT ecosystems by managing device diversity, processing generated data, and addressing challenges within IoT environments. By leveraging LLM capabilities, IoT systems can achieve higher levels of intelligence and autonomy, leading to more effective operations. For instance, LLMs facilitate natural human-machine interactions [5], improve data interpretation through contextual understanding [6], [7], and

Sarhaddi, F., Nguyen, N. T., Zuniga, A., Hui, P., Tarkoma, S., Flores, H., & Nurmi, P. (2026). **LLMs and IoT: A Comprehensive Survey on Large Language Models and the Internet of Things.** *IEEE Open Journal of the Communications Society*, *7*, 3585-3616. https://doi.org/10.1109/OJCOMS.2026.3680064

In-home Activity Segmentation and Recognition using LLMs/SLMs



Liu, J., Post, K., Kuchida, R., Zuniga, A., Sarhaddi, F., Flores, H., Nurmi, P., & Nguyen, N. T. (2026). **LLMSEG: Semantic Segmentation and Few-Shot Recognition of Human Activity Recognition Data Using Large Language Models.** In *ACM/IEEE International Conference on Embedded Artificial Intelligence and Sensing Systems* ACM and IEEE. https://doi.org/10.1145/3774906.3802768

In-home Activity Segmentation and Recognition using LLMs/SLMs

- LLMSeg improves **segmentation** performance
  - Outperforms a statistical windowing baseline (88.9% vs. 91.4%)
  - Largest improvements for short and fragmented activities (e.g., snacking 7.5% vs. 28.9%)
- Translates into better **classification** performance
  - Improves on fixed-window (85.8% → 90.7%)
  - Performs similarly to supervised approach (89.9% → 91.4%) but **requires no labels**

Activity	CSDW *Statistical Dynamic*	LLMSEG *Semantic Dynamic*
Leaving home	0.0039	**0.2272**
Personal care	0.7388	**0.8867**
Preparing breakfast	0.5697	**0.7450**
Preparing lunch	0.6858	**0.8025**
Relaxing on couch	0.9731	**0.9750**
Showering	0.9112	**0.9898**
Sleeping	**1.0000**	**1.0000**
Snacking	0.0752	**0.2891**
Weighted F1	0.8890	**0.9140**

Activity	LLM-based			Supervised
	LLMSEG *Dynamic*	Fixed Sliding *(60s/80%)*	Oracle *(Ideal Seg.)*	Bi-LSTM *(60s/80%)*
Leaving home	0.9655	**1.0000**	**1.0000**	**1.0000**
Personal care	0.8870	0.8675	**0.9662**	0.8782
Preparing breakfast	**0.7273**	0.6598	0.6970	0.5871
Preparing lunch	**0.7586**	0.6016	0.6814	0.6861
Relaxing on couch	0.9765	0.9666	**0.9870**	0.9861
Showering	0.8182	0.6239	0.9430	**0.9969**
Sleeping	0.9655	**1.0000**	**1.0000**	0.9916
Snacking	**0.6667**	0.5512	**0.7336**	0.5818
Weighted F1	**0.9068 ± 0.0045**	0.8578 ± 0.0017	**0.9343 ± 0.00997**	0.8997

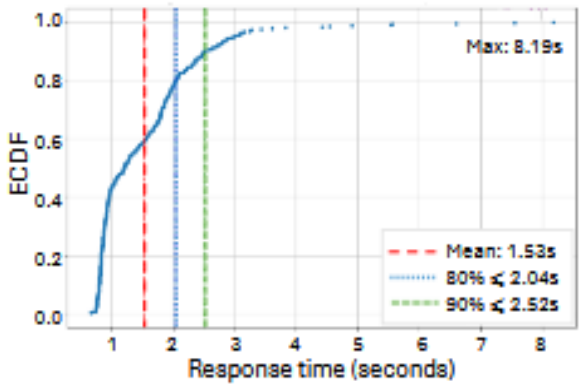
➔ Modern AI techniques can bring clear benefits in real-world tasks

In-home Activity Segmentation and Recognition using LLMs/SLMs

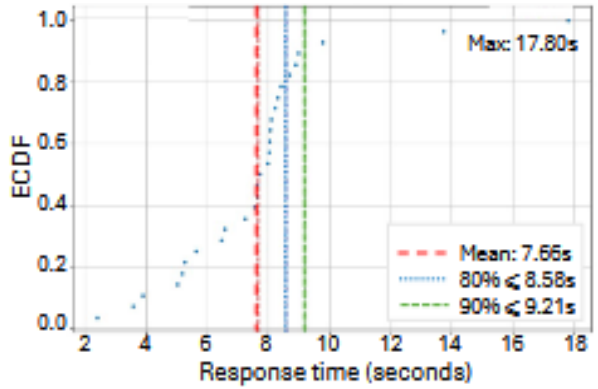
- Deployable on end-user edge/embedded devices
  - Classification ~1.5 seconds on average, 90% of windows done within 2.5 seconds
  - 90% of windows segmented within 10 seconds

Activity	Qwen3 -0.6B	Qwen3 -1.7B	Qwen3 -4B	Llama3.2 -1B	Llama3.2 -3B
Leaving home	0.6000	0.6957	0.7059	0.3284	0.5373
Personal care	0.2564	0.4783	0.4086	0.3171	0.5612
Preparing breakfast	0.4557	0.7907	0.5316	0.4000	0.5739
Preparing lunch	0.5000	0.5882	0.3030	0.4412	0.5517
Relaxing on couch	0.1667	0.7442	0.8060	0.4828	0.8112
Showering	0.2632	0.4138	0.2593	0.3571	0.3294
Sleeping	0.6897	0.6666	0.6486	0.3529	0.8169
Snacking	0.3333	0.0000	0.1739	0.1538	0.0526
Weighted F1	**0.4048**	**0.6036**	**0.5571**	**0.3844**	**0.6201**

- Performance decreases as simpler model used
  - ~ 8 billion parameters sufficient for good performance (see paper for details)
  - ~ 2 billion limit for RPI level devices without optimizations



(a) Activity recognition.

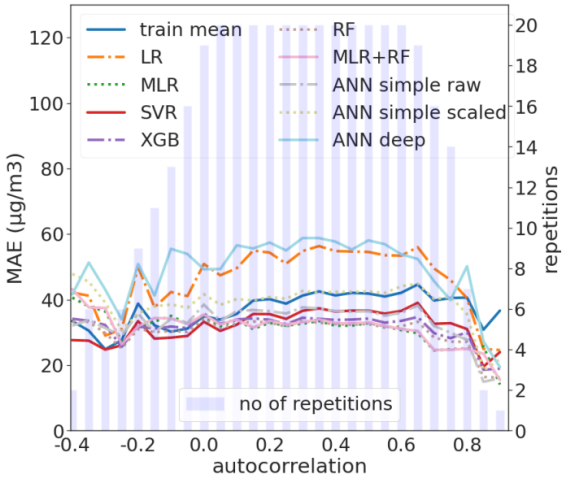


(b) Window segmentation.

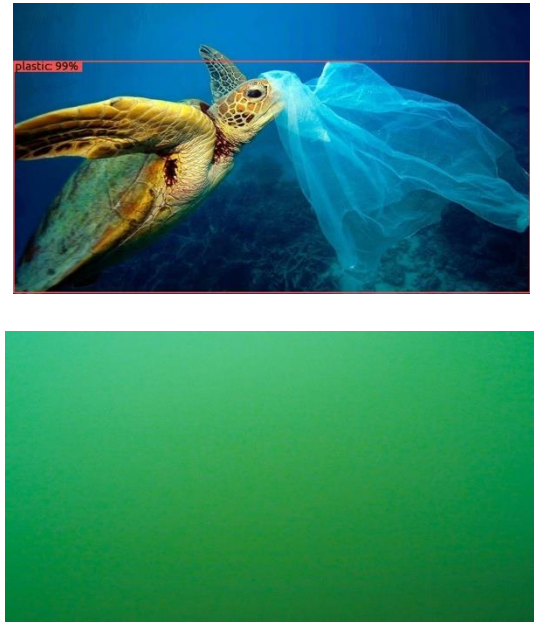
Murphy loves real-world data collection

- We established that AI can increasingly run on embedded devices
- However, real-world is the adversary for these models

Human Bottleneck

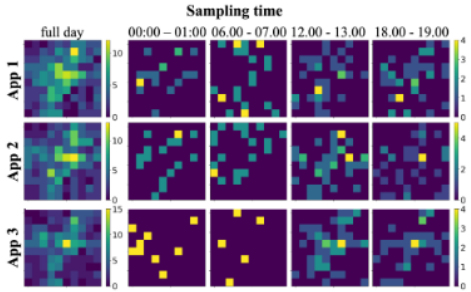


Data diversity



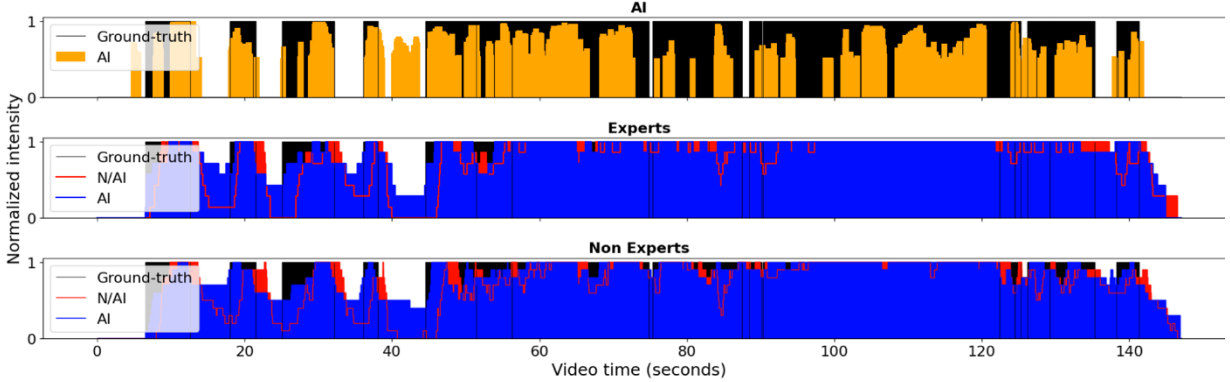
Environment Hostility

Missing data



Hardware heterogeneity

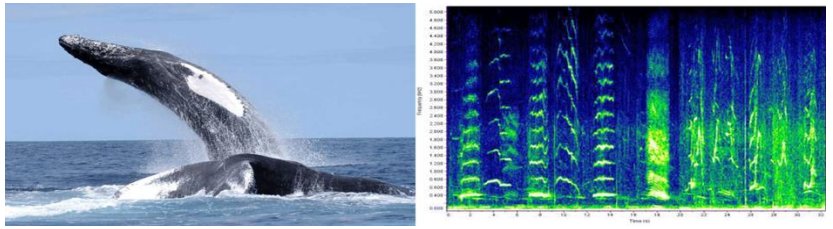
The Human Bottleneck



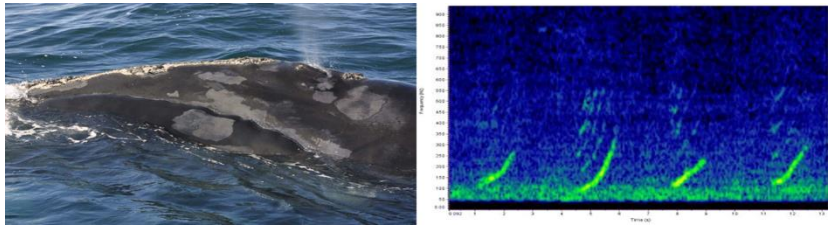
- Tailoring SLMs to function well requires labeled data to fine tune performance
- This data is hardly free to collect
  - Requires manual human effort
  - Can require highly specialized expertise

- Without accurate ground truth data, we cannot have accurate AI models
  - Doesn't matter where they run
  - Garbage-in, Garbage-out

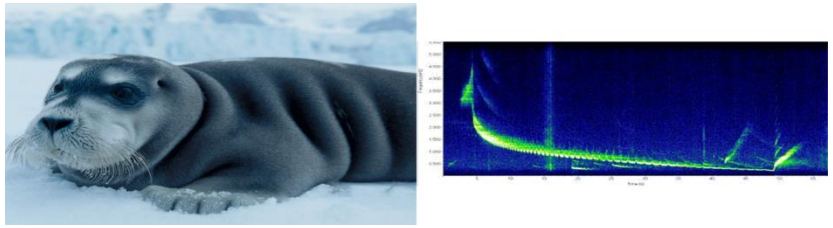
Humpback Whale



North Atlantic Right Whale

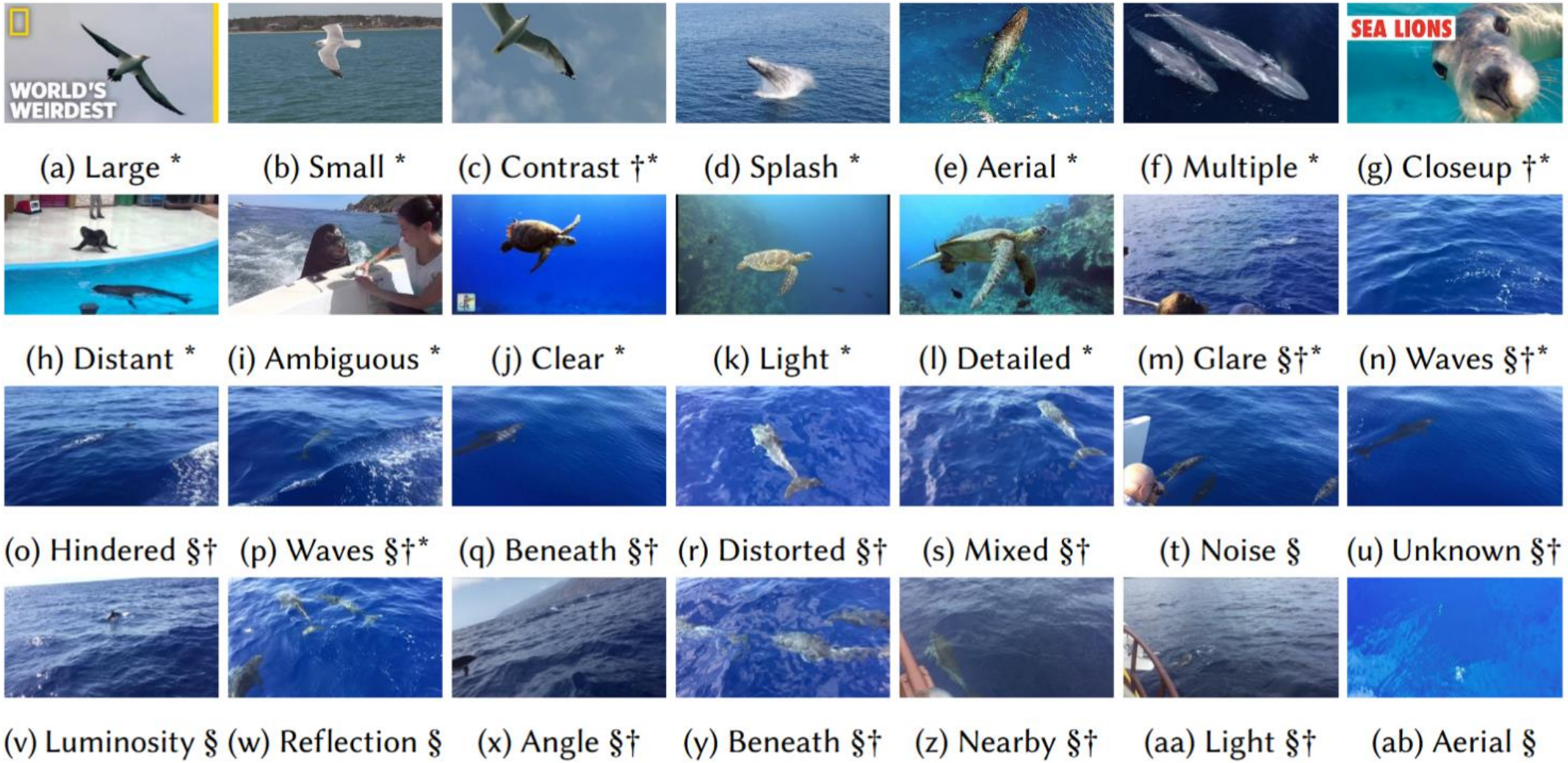


Bearded Seal



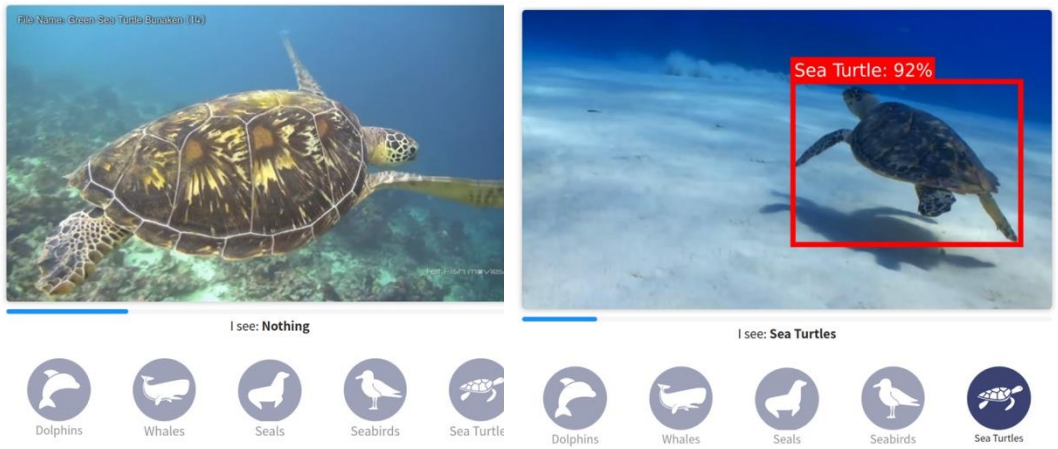
Source: https://www.fisheries.noaa.gov/national/science-data/sounds-ocean

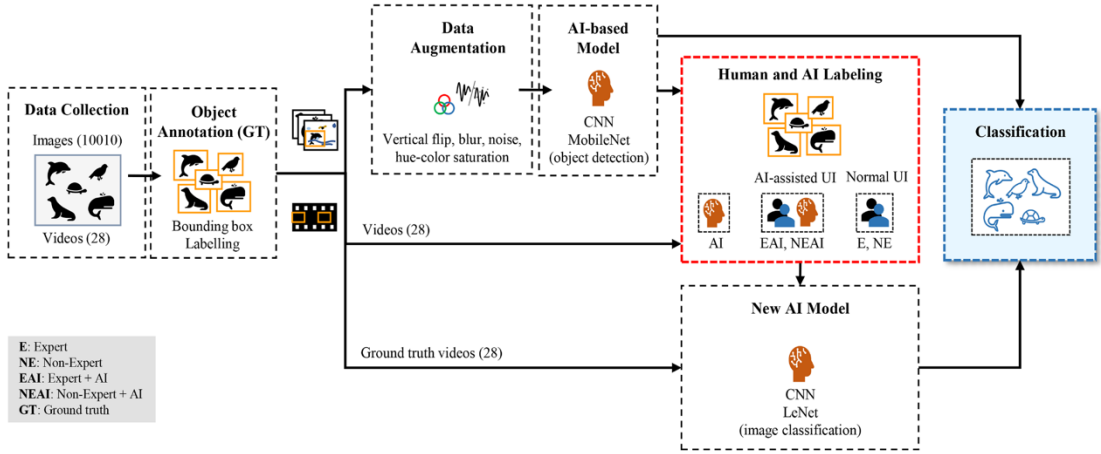
Labels aren’t perfect



Labeling Study

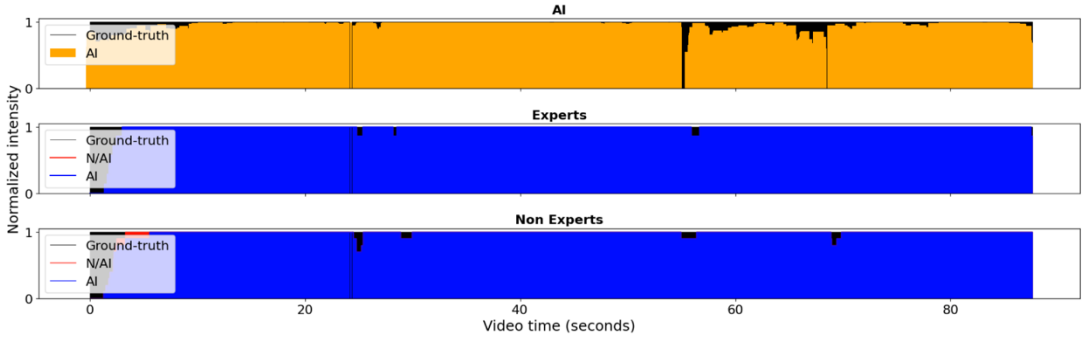
- We conducted a study where:
  1. 4 groups of users annotate video samples (expert vs. regular, AI support vs. w/out)
  2. Accuracy of labels compared and evaluated
  3. AI models are trained on the labeled data
  4. Accuracy of each AI model assessed separately
- Used openly available videos and proprietary video containing marine species observations

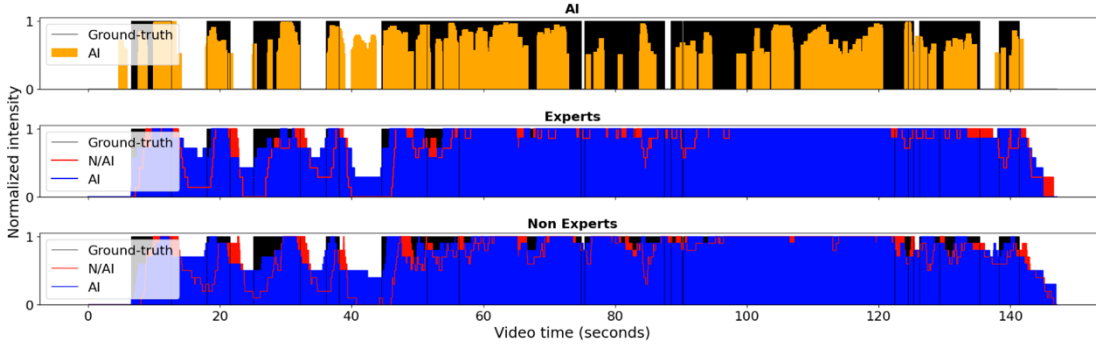




Key findings

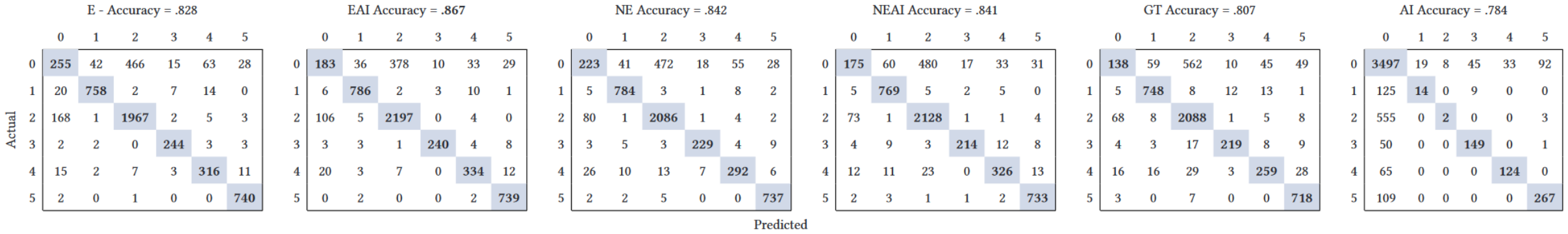
- Main findings for annotation performance:
  1. AI assistance helped **improved** non-experts, making their annotations almost as good as those of expert users
  2. AI assistance had **minimal or slightly negative effect** on experts due to disrupting them from their habitual performance
  3. The pace of video and the difficulty of the observations main factors affecting annotation performance





Key findings

- What about the effect these labelling paradigms have on AI models trained on them?
  1. **Model training with expert + AI** annotations had best overall performance
  2. AI only annotations resulted in worst performance, hence human annotations required
  3. For non-experts, AI improves labelling accuracy but not AI model performance, due to variation in the labels

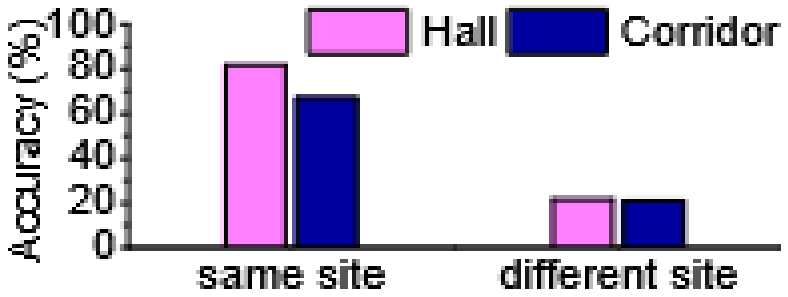


Data Challenge

- We showed that :
  1. Data labels are noisy and the noise affects how the AI models ultimately perform
  2. Real-world introduces noise to labels

- Data variations multiply these effects
  - Spatial variations across locations / environments
  - Temporal dependencies across time

- Scaling gap
  - AI assumes inputs align with training data
  - Variation demands more training data (and labels)



Zhang, J., Tang, Z., Li, M., Fang, D., Nurmi, P. T., & Wang, Z. (2018). CrossSense: Towards Cross-Site and Large-Scale WiFi Sensing. In *Proceedings of the 24th Annual International Conference on Mobile Computing and Networking : MobiCom'18* ACM. https://doi.org/10.1145/3241539.3241570

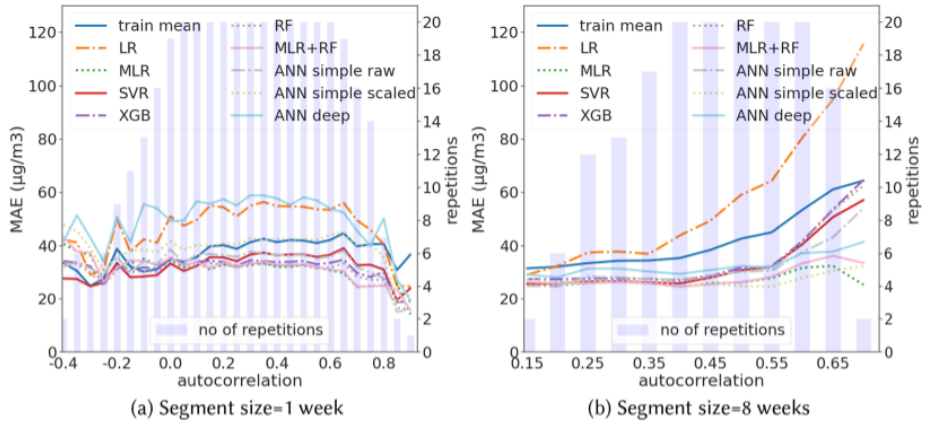
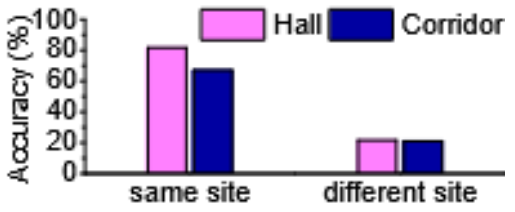
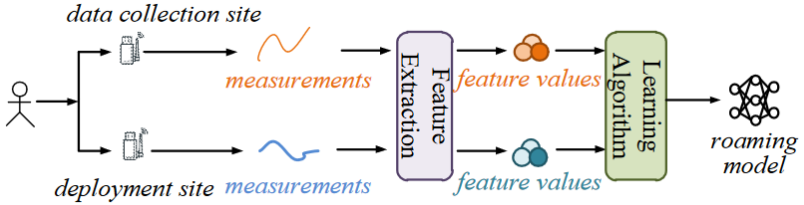


Fig. 2. The accuracy of calibration varies as target variable's (NO_2) autocorrelation between training and testing data changes. When the segment size of training and testing sets increases, i.e., the longer the continuous period of measurements that is considered, model predictions tend to reflect more general pollution levels than the actual required corrections, which results in higher calibration errors. The error in the y-axis is mean absolute error.

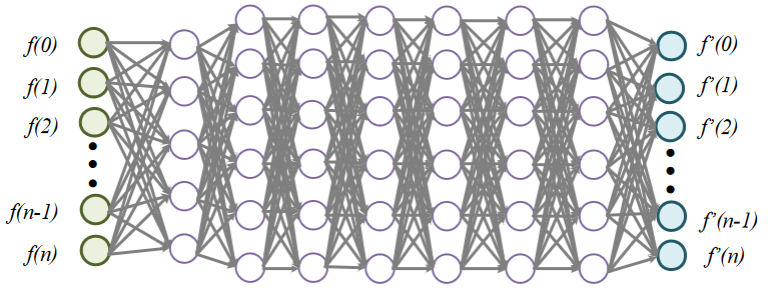
Data Diversity: Spatial Variations



- Wireless sensing example where minor location changes degrade performance significantly (up to ~80% loss)
- Concept drift: need to retrain model when environment changes



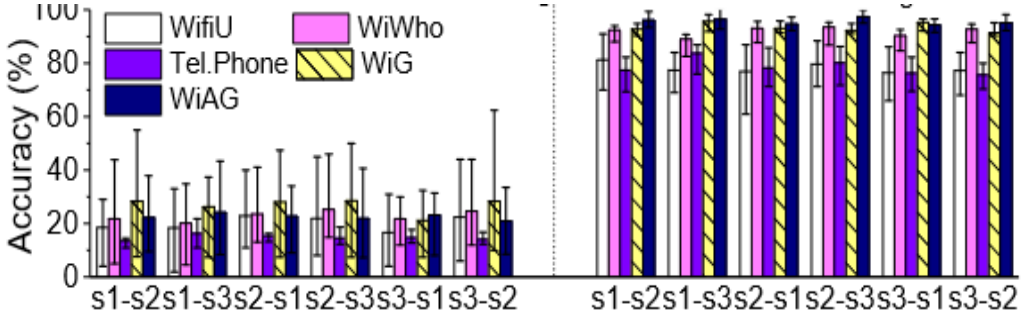
- The CrossSense Solution
  - Roaming model used to bridge measurement locations
  - Mixture-of-Experts approach for improving generality



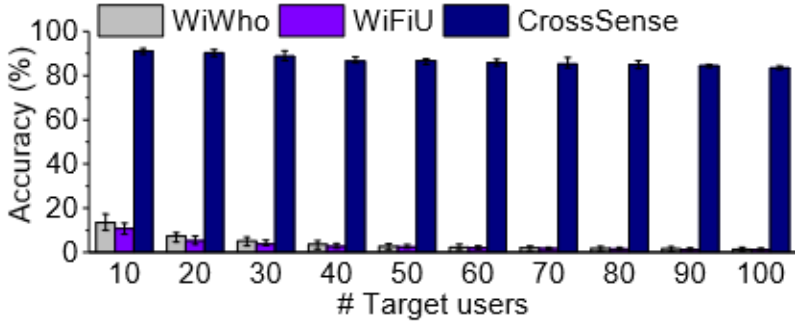
- Modern RAG architectures similarly anchor AI responses
  - MoE offers support for *context dependency*
  - Data not just used for training but as anchor to align outputs

Zhang, J., Tang, Z., Li, M., Fang, D., **Nurmi, P.**, & Wang, Z. (2018). CrossSense: Towards cross-site and large-scale WiFi sensing. In *Proceedings of the 24th Annual International Conference on Mobile Computing and Networking* (pp. 305-320).

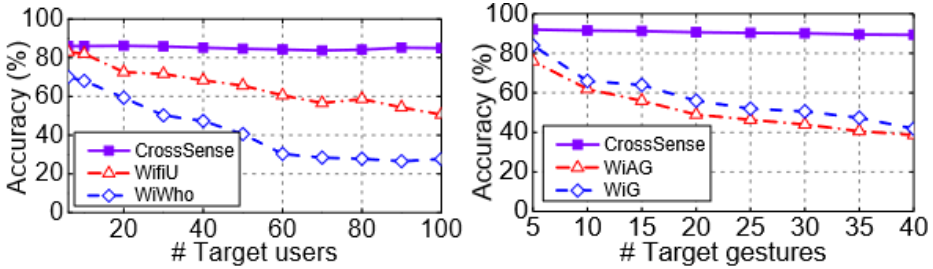
CrossSense: Selected Findings



- Evaluations using gait identification and gesture recognition as target problems
- Selected findings:
  1. Roaming model translates performance to new environments
  2. Performance without roaming drops by up to 80%
  3. MoE makes performance consistent as problem size increases
  4. MoE + roaming model overcome spatial variation



➔ CrossSense reduces need for **additional labels** and adapts model to new environments or larger problems



Data Diversity: Temporal Variation

- Real-world data often has temporal structures
  - Weather today helps to predict how it will be tomorrow
  - Standard machine learning training ignores temporal structure

- Temporal structures easily result in overfitting
  - Total error and ranking of models influenced by how training and test sets constructed
  - Training model with apples doesn't help detect oranges

➔ How we curate data affects AI performance

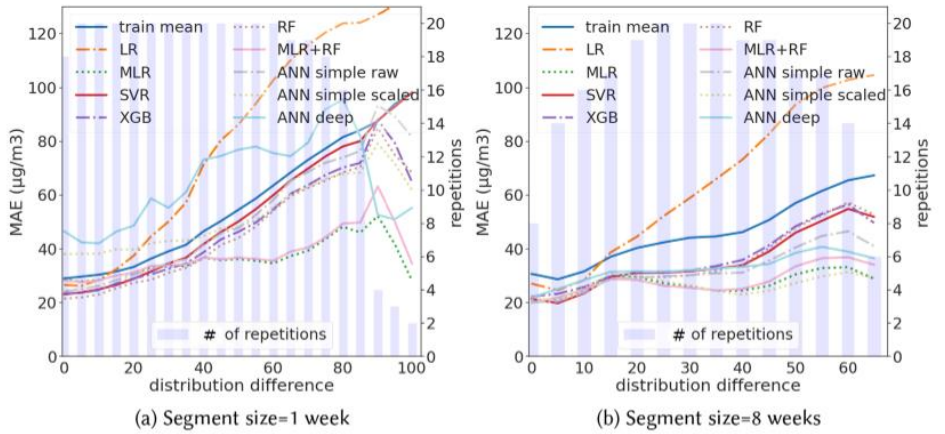
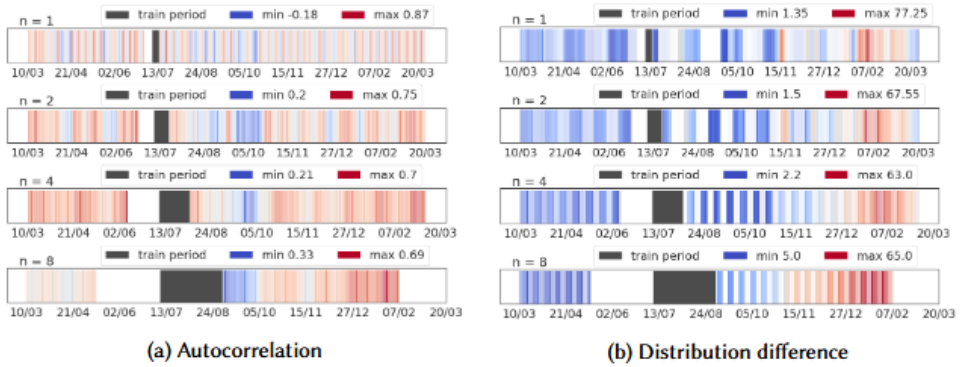


Fig. 4. The performance of calibration decreases as target variable's (NO_2) distribution in testing data differs from the distribution in training data. The error in the y-axis is mean absolute error.

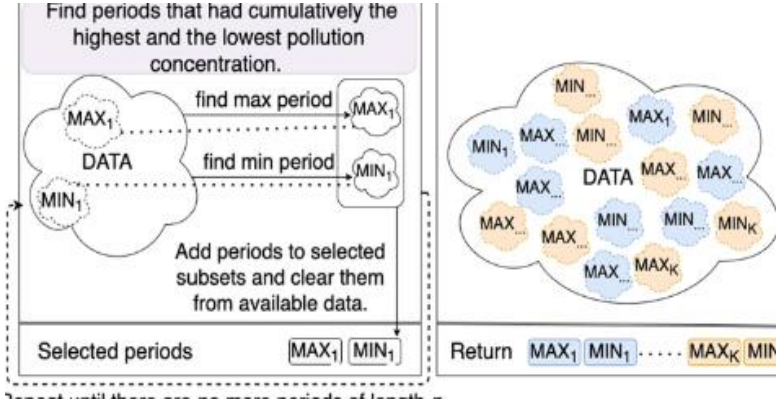
Source: Aula, K., Lagerspetz, E., **Nurmi, P.**, & Tarkoma, S. (2022). Evaluation of low-cost air quality sensor calibration models. *ACM Transactions on Sensor Networks*, *18*(4), 1-32.

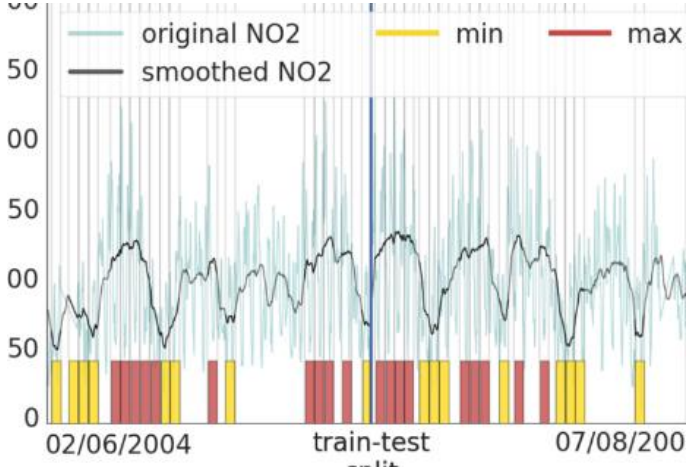


Training Data Diversification

- Diversification overcomes temporal diversity and improves consistency of AI models
- Training data constructed to ensure broader distributional coverage to ensure better performance

- Selected findings on diversification:
  - Helps test model generality
  - Creates splits that align with requirements for field testing
  - Improves training performance by removing redundant periods
  - Provides better guarantees on model performance





	cont.	low	high	diverse
Training mean	1.5	0.88 (-41%)	2.57 (+71%)	2.02 (+35%)
LR	0.97	**0.55 (-43%)**	1.34 (+38%)	**1.17 (+21%)**
MLR	**0.95**	0.72 (-24%)	1.3 (+37%)	1.25 (+32%)
SVR	**0.95**	0.63 (-34%)	1.44 (+52%)	1.3 (+37%)
XGB	1.07	0.61 (-43%)	1.57 (+47%)	1.29 (+21%)
RF	0.98	0.63 (-36%)	1.39 (+42%)	1.25 (+28%)
MLR+RF	0.97	0.84 (-13%)	**1.24 (+28%)**	1.21 (+25%)
ANN simple raw	0.96	0.88 (-8%)	1.47 (+53%)	1.45 (+51%)
ANN simple scaled	0.98	1.11 (+13%)	1.45 (+48%)	1.52 (+55%)
ANN deep	1.01	0.6 (-41%)	1.54 (+52%)	1.24 (+23%)
AVG	1.03	0.74 (-28%)	1.53 (+49%)	1.37 (+33%)
4 months (2880 data points w/ 75:25 training-testing split				

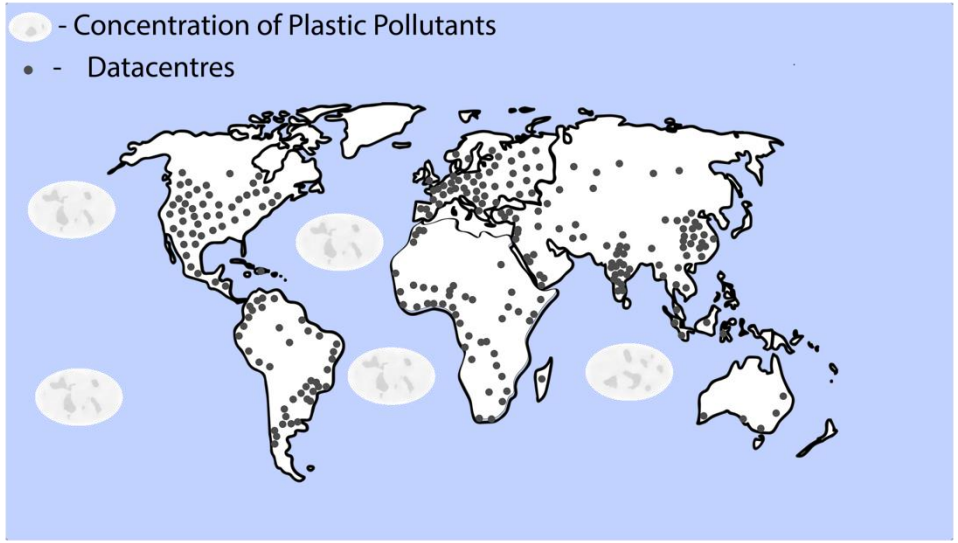
(a) CO (in mg/m^3)

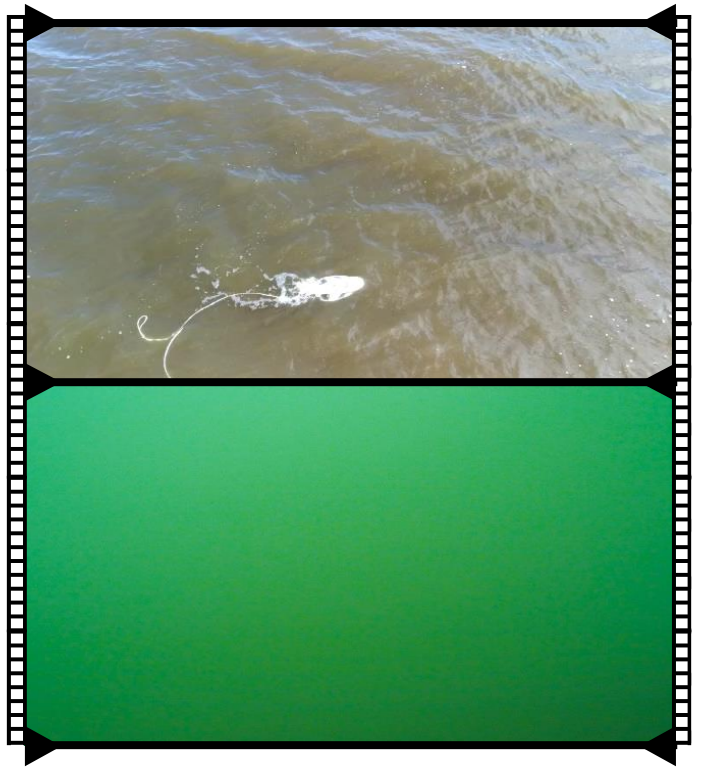
Environment Hostility

- Cloud data centres close to inhabited areas
- However, what if the cloud is not available?
  - 71% of Earth's surface is Ocean, cloud access available only at coastal areas
  - Counting underwater areas further reduces coverage

- Ocean not just difficult for connectivity
  - Requires careful device designs
  - Subject to biofouling
  - Highly varying environmental conditions

➔ Hostile environments such as Oceans present ultimate stress test for embedded / edge AI



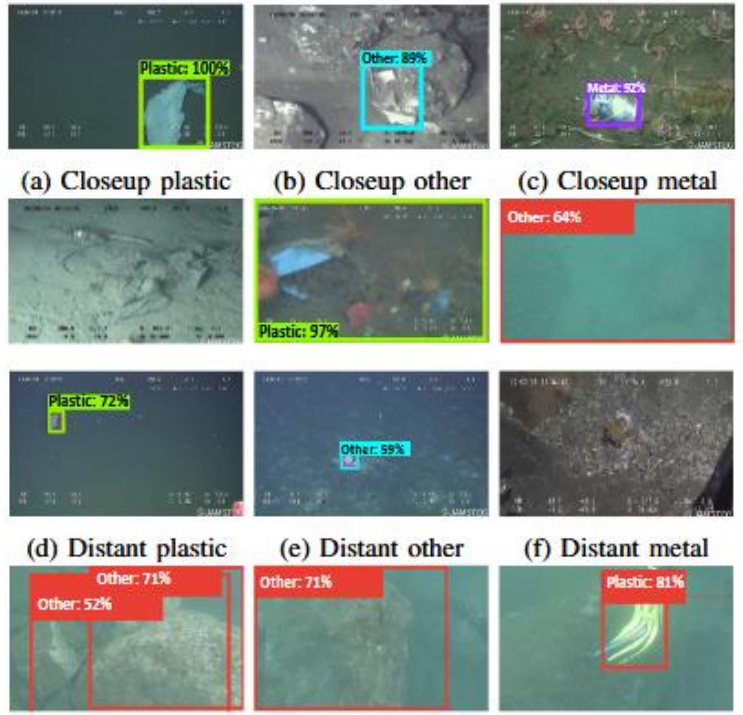


Bring Your AI

- Current loop: collect on device, analyse in lab

➔ Substantial data-to-insight latency

- Embedding intelligence reduces latency but introduces hurdles
  - Do-no-harm: gathering input cannot disturb ecosystem
  - Expectation gap: people assume pristine data, reality highly different
  - Dynamic noise: turbidity, turbulence, temperature, water characteristics actively corrupt signals

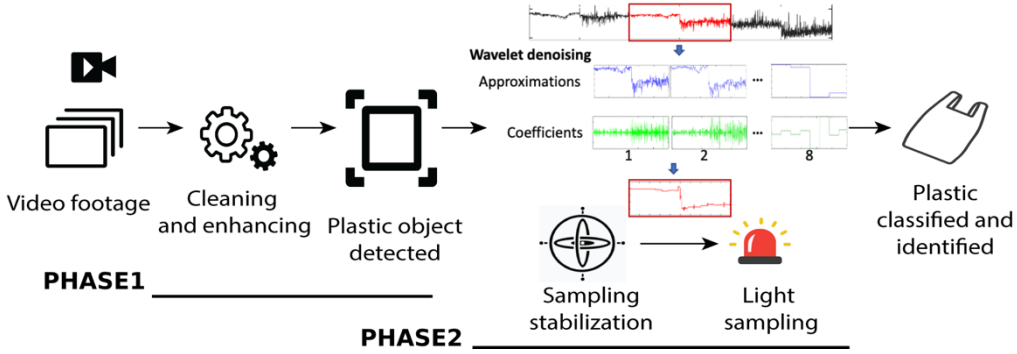


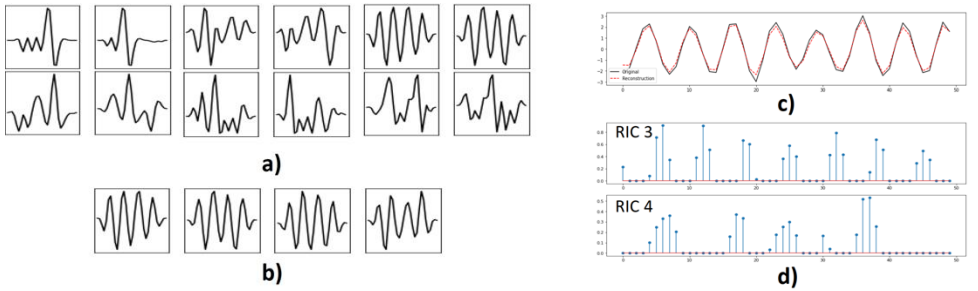
Radeta, M., Zuniga, A., Motlagh, N. H., Liyanage, M., Freitas, R., Youssef, M., Tarkoma., S., Flores, H., & Nurmi, P. (2022). Deep learning and the oceans. *Computer*, *55*(5), 39-50.

Two-stage underwater plastic classification

- SEAGULL research example of how to bridge these hurdles
- Vision and optical sensing combined to overcome visibility issues
- Convolutional sparse coding (CSC) used to represent signals to improve generality

- Selected findings
  - ~ 85% accuracy detecting plastic types
  - CSC improves generality by ~10 percentage points



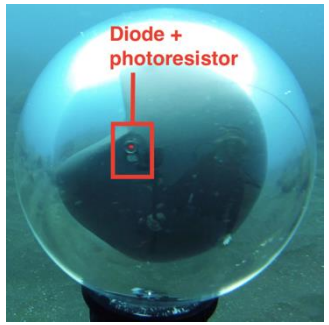


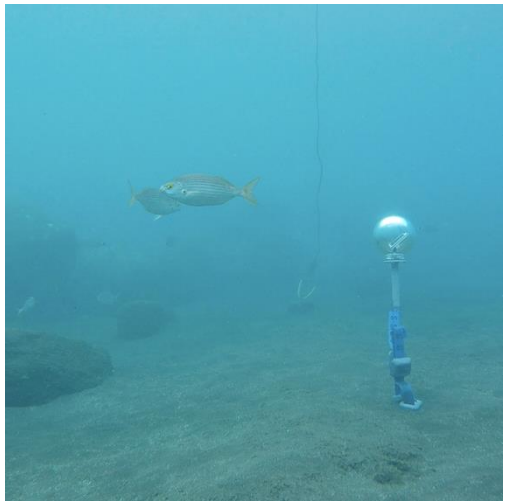
Test conditions		CSC RF	CSC GB	Full spectrum (non-CSC) RF	Full spectrum (non-CSC) GB
Direct	(r)	84.6	**87.0**	81.5	81.5
	(r,l)	85.2	**86.8**	83.3	83.3
Water	(r)	82.8	83.3	84.6	**85.1**
	(r,d)	83.0	83.1	**90.8**	90.4
Ice	(r)	**87.9**	87.5	74.2	76.3
	(r,l)	87.6	**87.8**	85.9	85.4
Turbulence	(r)	84.8	**86.2**	63.5	63.5
	(r,d)	85.3	**85.5**	71.9	71.4
Turbidity	(r)	**82.3**	81.7	76.7	80.7
	(r,l)	82.0	82.0	**87.0**	86.1
	(r,d)	82.4	81.9	**87.8**	86.7
	(r,t)	82.3	82.5	**85.1**	85.0
Average	(r)	84.5±2.2	**85.1±2.5**	76.1±8.1	77.4±8.4
	(r,l)	84.9±2.8	**85.5±3.1**	85.4±1.9	84.9±1.5
	(r,d)	**83.6±1.5**	83.5±1.8	83.5±10.2	82.8±10.2
	Overall	84.2±2.1	**84.6±2.4**	81.0±7.9	81.3±7.5

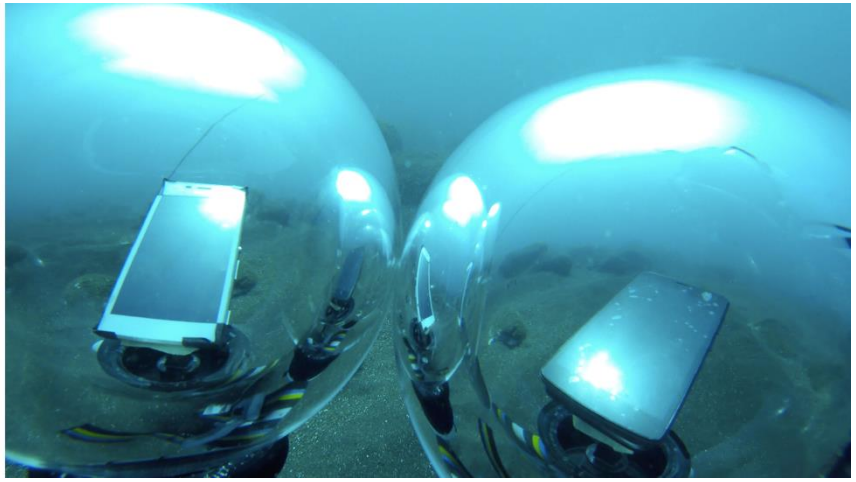
Flores, H., Zuniga, A., Radeta., M., Yin, Z., Liyanage, M., Motlagh, N. H., Nguyen, N. T., Tarkoma, S., Youssef, M., Nurmi, P. *ACM/IEEE IoTDI 2025*

Bringing compute to underwater

- Where to get compute power for AI?
  - Movement consumes significant power, difficult to integrate with drones or divers
- Solution: bring your own edge
  - Smartphone based opportunistic cluster
  - Can also use repurposed eWaste
  - Devices isolated from water ➔ can use WiFi for data transfer
- 2-4 smartphones can run 30+ fps object recognition from video



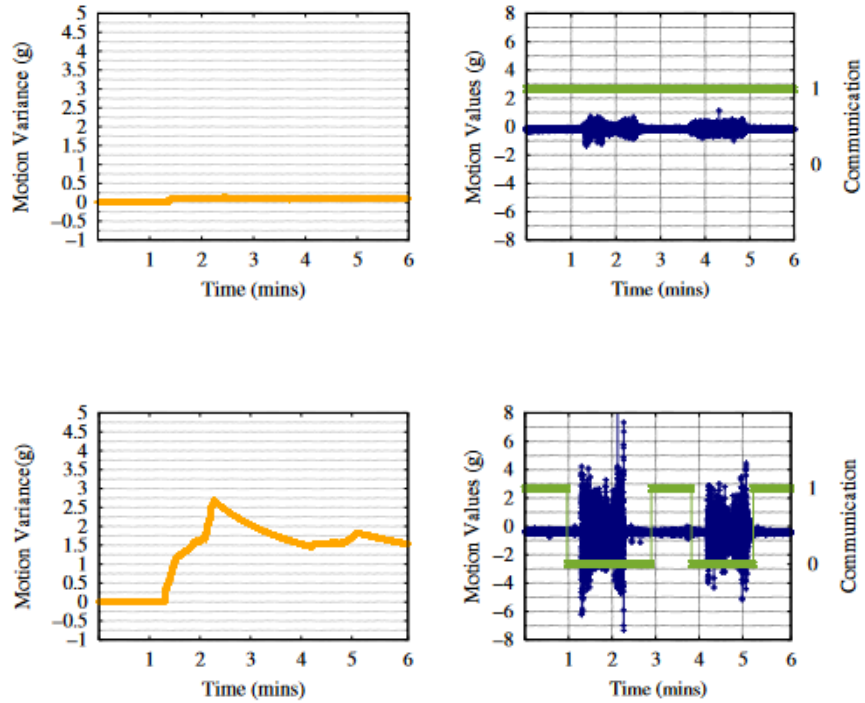




Dar, F., Liyanage, M., Radeta, M., Yin, Z., Zuniga, A., Kosta, S., S. Tarkoma, P. Nurmi, & Flores, H. (2023). Upscaling Fog Computing in Oceans for Underwater Pervasive Data Science using Low-Cost Micro-Clouds. *ACM Transactions on Internet of Things*, *4*(2), 1-29.

Bringing compute to underwater

- As with plastics, multiple sensors can improve robustness
  - Water motion disrupts WiFi exchange
- Accelerometers used to detect optimal transmission periods
  - Variance small ➔ motion stable
- Hostile environments require designing for **intermittent** connectivity
- Cloud often assumes “too perfect” communications



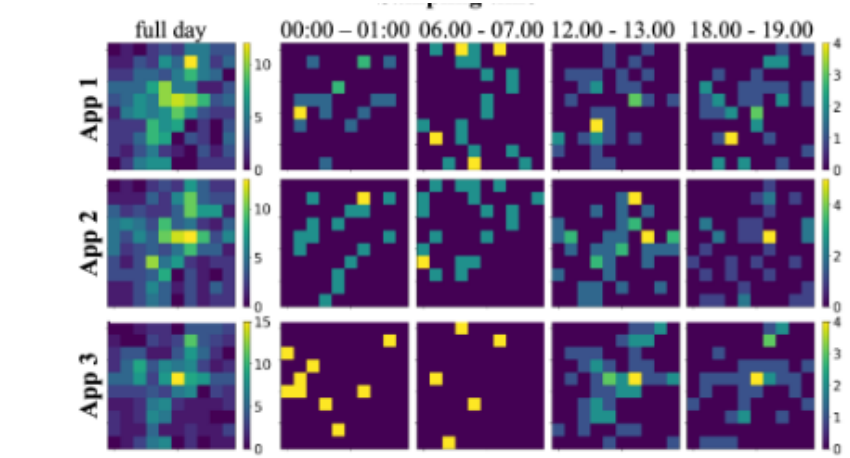
Dar, F., Liyanage, M., Radeta, M., Yin, Z., Zuniga, A., Kosta, S., S. Tarkoma, P. Nurmi, & Flores, H. (2023). Upscaling Fog Computing in Oceans for Underwater Pervasive Data Science using Low-Cost Micro-Clouds. *ACM Transactions on Internet of Things*, *4*(2), 1-29.

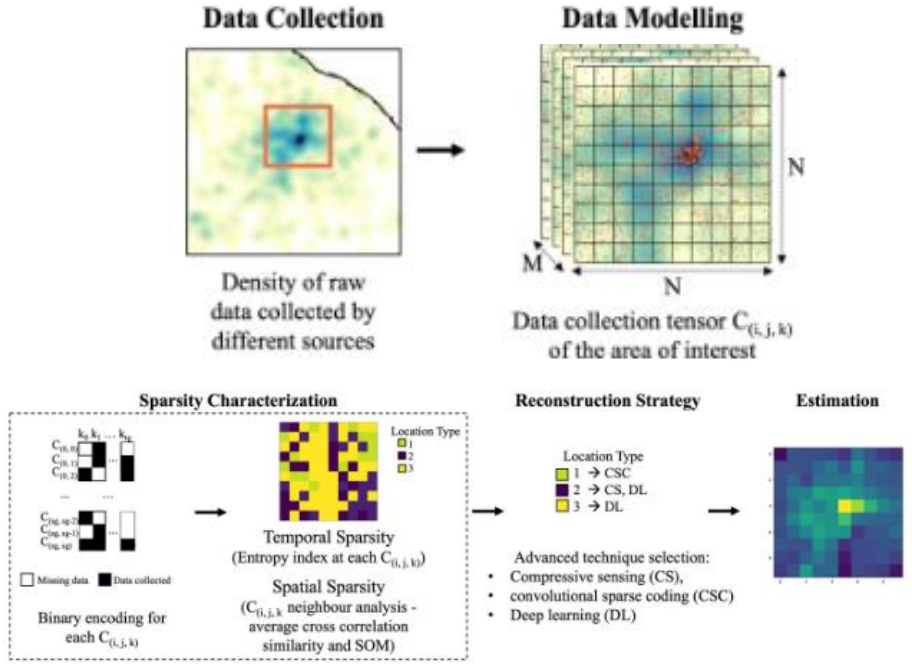
Missing Data Bottleneck

- Temporal and spatial sparsity challenge AI training
- Unintentional biases in how data gathered

- SARF addresses sparsity through reconstruction
  - Analyze sparsity of measurements
  - Dynamically adapt reconstruction based on sparsity

- Data collection design can affect which data is missing
- Missingness a key design criteria to consider





Zuniga, A., Flores, H., Nguyen, N. T., Hui, P., Tarkoma, S., & Nurmi, P. (2026). SARF: Sparsity-Aware Reconstruction Framework for Large-Scale Datasets. *IEEE Transactions on Big Data*, *12*(2), 612-624. https://doi.org/10.1109/TBDATA.2025.3639968

The Data Volume Bottleneck

- Sometimes the opposite is true: too much data, but signals of interest rare
- Event-based computing an emerging paradigm for addressing this challenge
  - “Finding needles in a haystack”
  - Spiking neural networks offer high volume reduction
  - But accuracy depends on nature of events
- Paves the way for neuromorphic computing
  - But can be deployed on commodity hardware

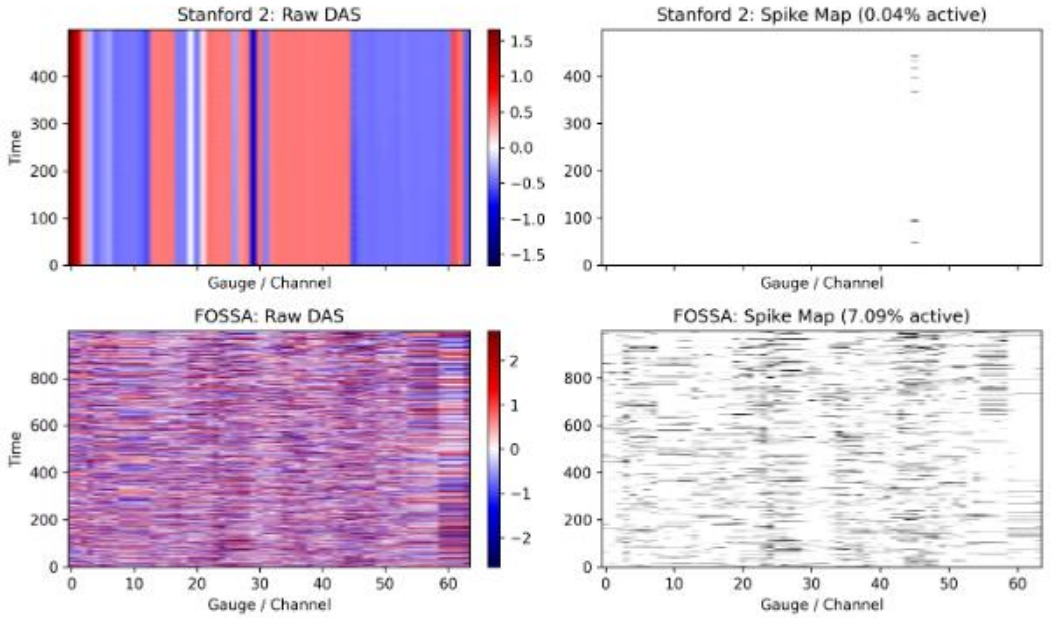


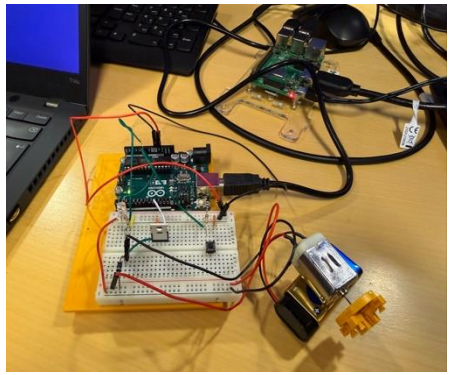
TABLE I
DETECTION PERFORMANCE COMPARISON BETWEEN CNN AND SNN ON THE COMBINED DAS DATASET.

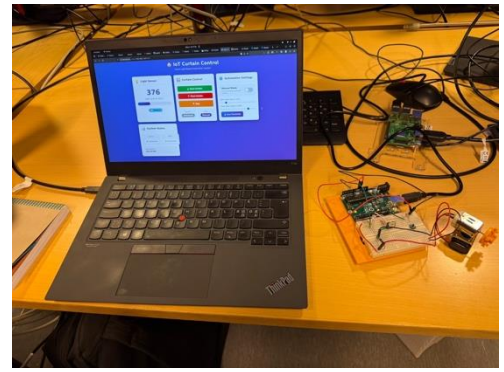
Model	Acc	Prec	Rec	F1	Spec	FPR	Bal Acc
CNN	0.8428	0.8145	0.7063	0.7566	0.9150	0.0850	0.8106
SNN	**0.8597**	**0.9167**	0.6538	**0.7633**	**0.9686**	**0.0314**	**0.8112**

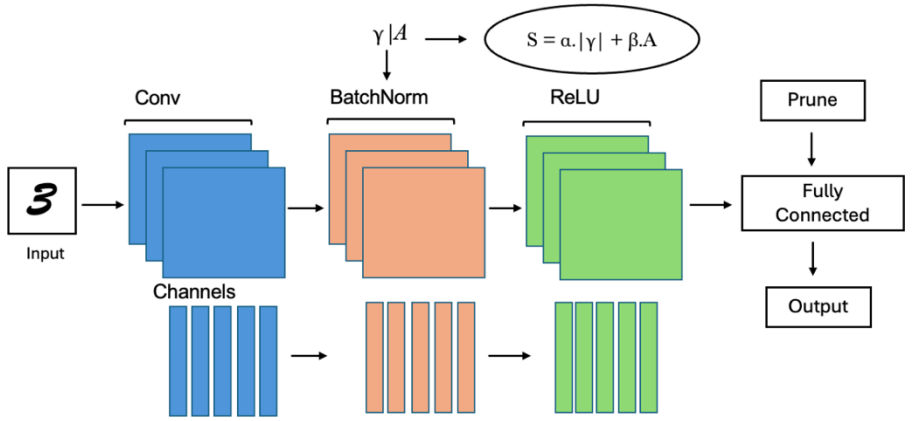
Khan, M., Sarhaddi, F., Zuniga, A., Tarkoma, S., Radeta, M., Flores, H., Rao, A., Heinonen, S., Kangasharju, J., & Nurmi, P. (2026). Through the Fiberglass: Scaling Global Fiber Networks into Event-Driven Observation Systems. Manuscript submitted for publication.

Hardware Ecosystem

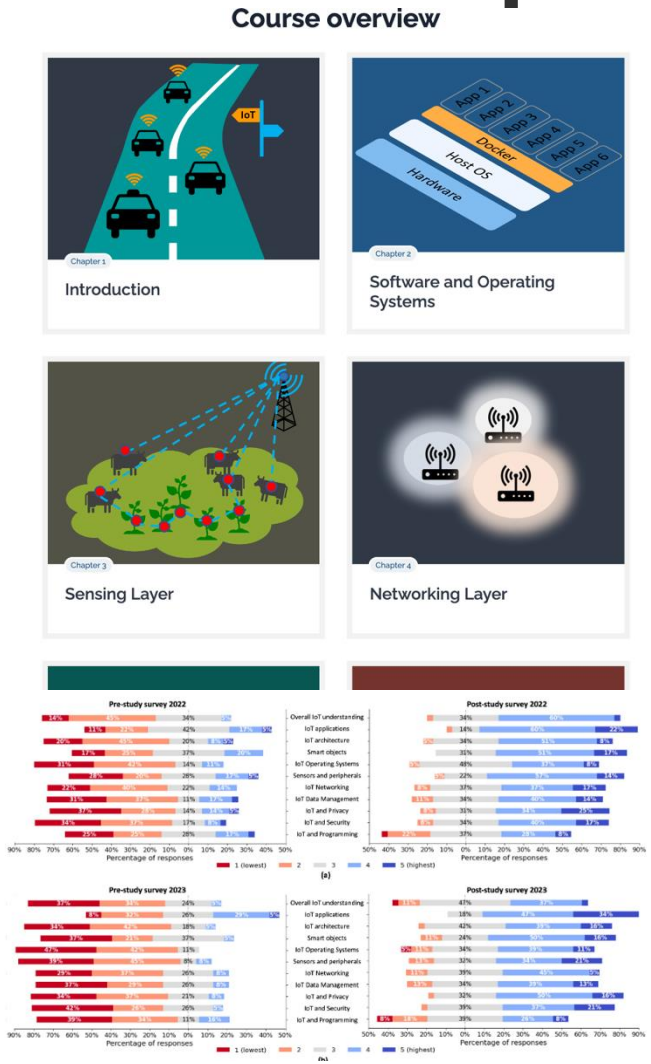
- Edge ecosystem not standardized
  - CPU, GPU, NPU, TPU, QPU, ….
  - Different software environments
  - Different performance optimizations
- Research and teaching experience has shown that translating code across platforms a bottleneck
  - Learning gap across platforms
- Current research looks into overcoming this
  - Hardware agnostic optimization frameworks
  - Software abstractions for emerging xPU families

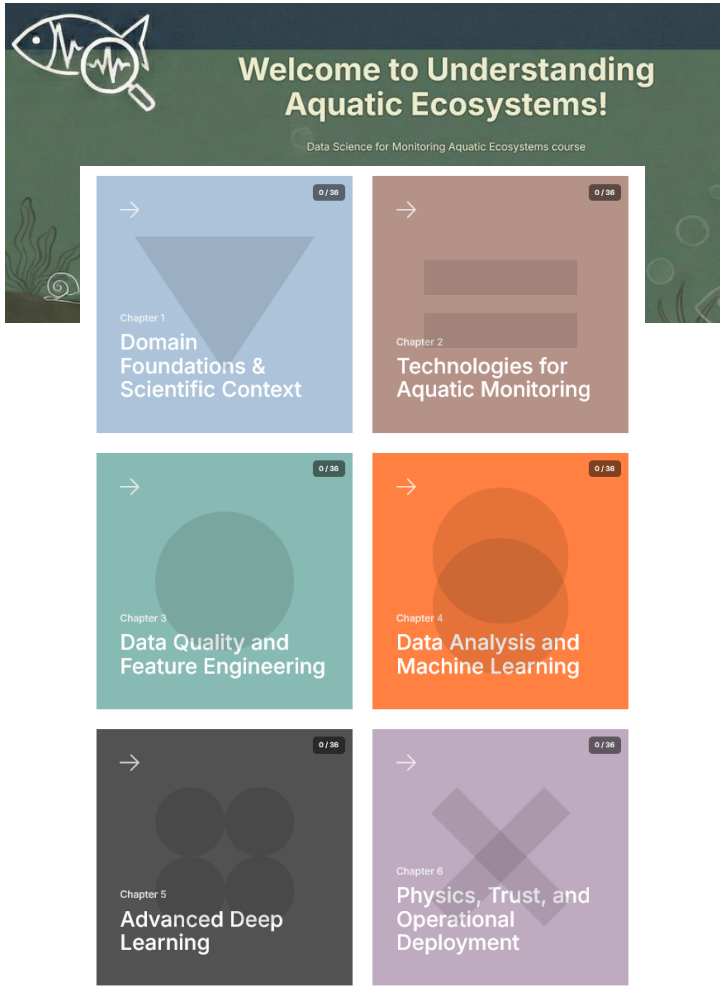


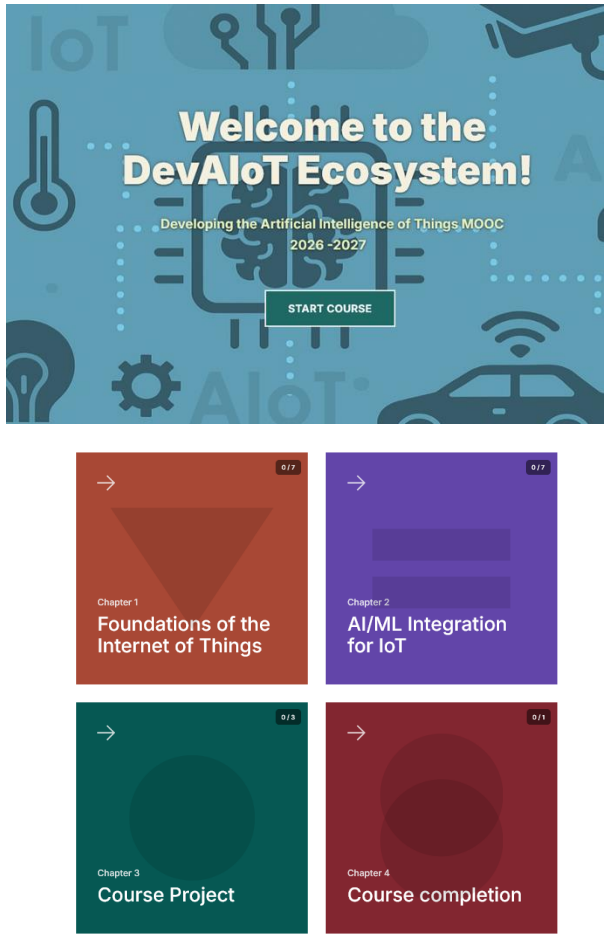




Education: Openly accessible resources

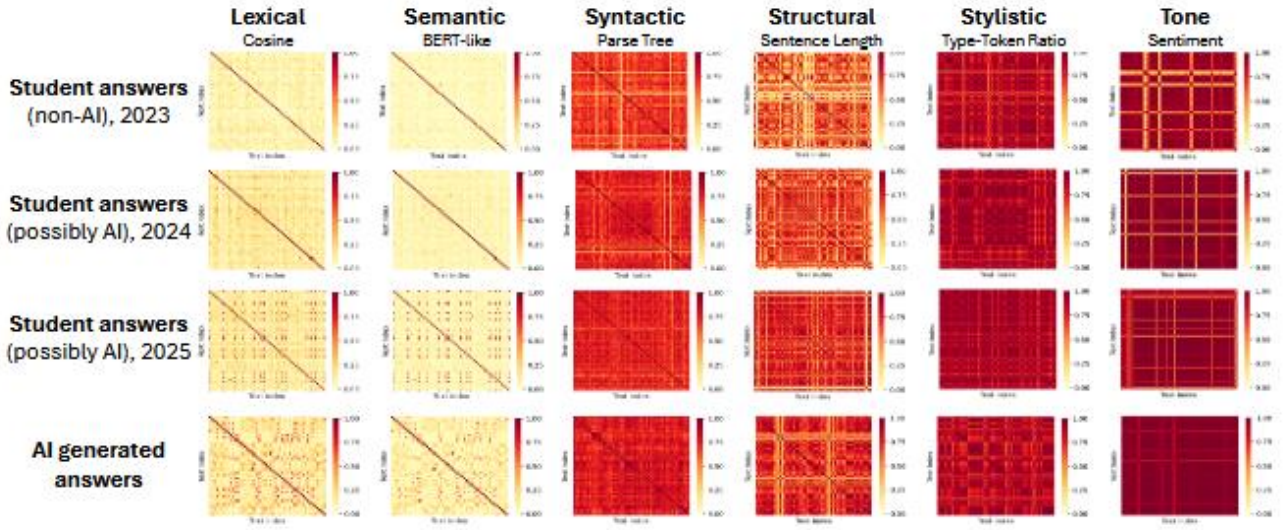






iot.mooc.fi aquaticds.mooc.fi aiot.mooc.fi

Education: Research



Similarity	Metric	Description
Lexical	Jaccard	Measures shared unique items over total unique items.
	Cosine	Measures angle between text vector representation directions.
Semantic	Word2vec	Learns word vectors from context in large text.
	BERT	Deep model generating context-aware sentence or word embeddings.
Syntactic	Levenshtein	Counts edits needed to change one string into another.
	Parse tree	Hierarchical structure showing grammatical relations in sentences.
Structural	Length ratio	Compares lengths between two text segments quantitatively.
	Sentence length	Number of words or tokens in a sentence.
Stylistic	Type–Token ratio	Unique words divided by total words.
	Punctuation frequency	Frequency of punctuation marks in text.
Tone	Sentiment	Overall positive, negative, or neutral feeling expressed.
	Emotion	Specific feelings like joy, anger, or sadness detected.

AI See What You Did There – The Prevalence of LLM-Generated Answers in MOOC Responses

Petteri Nurmi
University of Helsinki
Helsinki, Finland
petteri.nurmi@helsinki.fi

Musfira Khan
University of Helsinki
Helsinki, Finland
musfira.khan@helsinki.fi

Zahra Safaei
University of Helsinki
Helsinki, Finland
zahra.safaei@helsinki.fi

Ngoc Thi Nguyen
National University of Singapore
Singapore, Singapore
ngoc.nguyen@nus.edu.sg

Fatemeh Sarhaddi
University of Helsinki
Helsinki, Finland
fatemeh.sarhaddi@helsinki.fi

Mika Tompuri
University of Helsinki
Helsinki, Finland
mika.tompuri@helsinki.fi

Henrik Nygren
University of Helsinki
Helsinki, Finland
henrik.nygren@helsinki.fi

Päivi Kinnunen
University of Helsinki
Helsinki, Finland
paivi.kinnunen@helsinki.fi

Agustin Zuniga
University of Helsinki
Helsinki, Finland
agustin.zuniga@helsinki.fi

Abstract

Large language models (LLMs) are reshaping the educational landscape, particularly in online learning environments where student supervision is often limited. Early evidence and anecdotal reports suggest that the use of AI-generated content is highly prevalent among students. However, definitive statistics remain elusive, primarily due to the challenges associated with distinguishing between AI-generated and human-generated responses. Establishing clear evidence and effective mechanisms for identifying AI-generated responses is crucial for understanding the significance of this challenge and for developing policies to address it. To tackle these issues, we present a large-scale empirical study on the prevalence of AI-generated content in online education. Our study analyzes over 4045 student responses from an introductory MOOC on the Internet of Things, employing textual analysis techniques to evaluate various metrics for identifying AI-generated responses and understanding their characteristics. Our findings reveal that a significant majority of student responses (up to 90.1%) exhibit strong similarities to AI-generated content in both wording and contextual meaning, regardless of the specific LLM or similarity metric employed. In terms LLMs usage, DeepSeek, Gemini, and Grok are the three most popular LLMs used to generate responses.

CCS Concepts

• **General and reference** → **Cross-computing tools and techniques**; • **Information systems**; • **Applied computing** → **Education**;

Keywords

internet of things, massive open online course, llms, text similarity

SIGCSE TS 2026, St. Louis, MO, USA

ACM ISBN 979-8-4007-2256-1/26/02
https://doi.org/10.1145/3770762.3772602

ACM Reference Format:
Petteri Nurmi, Musfira Khan, Zahra Safaei, Ngoc Thi Nguyen, Fatemeh Sarhaddi, Mika Tompuri, Henrik Nygren, Päivi Kinnunen, and Agustin Zuniga. 2026. AI See What You Did There – The Prevalence of LLM-Generated Answers in MOOC Responses. In *Proceedings of the 57th ACM Technical Symposium on Computer Science Education V.1 (SIGCSE TS 2026), February 18–21, 2026, St. Louis, MO, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3770762.3772602

1 Introduction

The rise of online education, particularly through Massive Open Online Courses (MOOCs), has transformed the educational landscape, enhancing accessibility and flexibility for students worldwide [20, 21]. However, this shift has introduced challenges in maintaining academic integrity and ensuring the authenticity of student submissions. The emergence of large language models (LLMs) complicates these issues, as modern AI tools can generate highly human-like responses [18] that may not accurately reflect student's true understanding.

Despite the increasing presence of AI in education, empirical evidence on its prevalence in online settings remains scarce. While early studies and anecdotal reports suggest that many students utilize AI-generated content [4, 13], definitive statistics are lacking. This gap in reliable data hinders our understanding of the issue's significance and the development of effective strategies to promote responsible AI use and mitigate potential misuse. Moreover, research indicates that over-reliance on AI can diminish cognitive engagement among students [4]. Therefore, understanding the prevalence of AI-generated content is essential for fostering academic integrity in online learning environments.

This paper presents the first large-scale empirical study on the prevalence of AI-generated content in online education. Utilizing a dataset of over 1000 student responses from an introductory MOOC on the Internet of Things, we employ textual analysis techniques to evaluate various metrics for identifying AI-generated responses and understanding their characteristics. Our analysis sheds light on the extent of AI use and provides insights into the effectiveness

770

Education: Projects

Collaborative International Education Projects

OPETUSHALLITUS
UTBILDNINGSSTYRELSEN

UNIVERSITY OF JYVÄSKYLÄ
JYVÄSKYLÄN YLIOPISTO

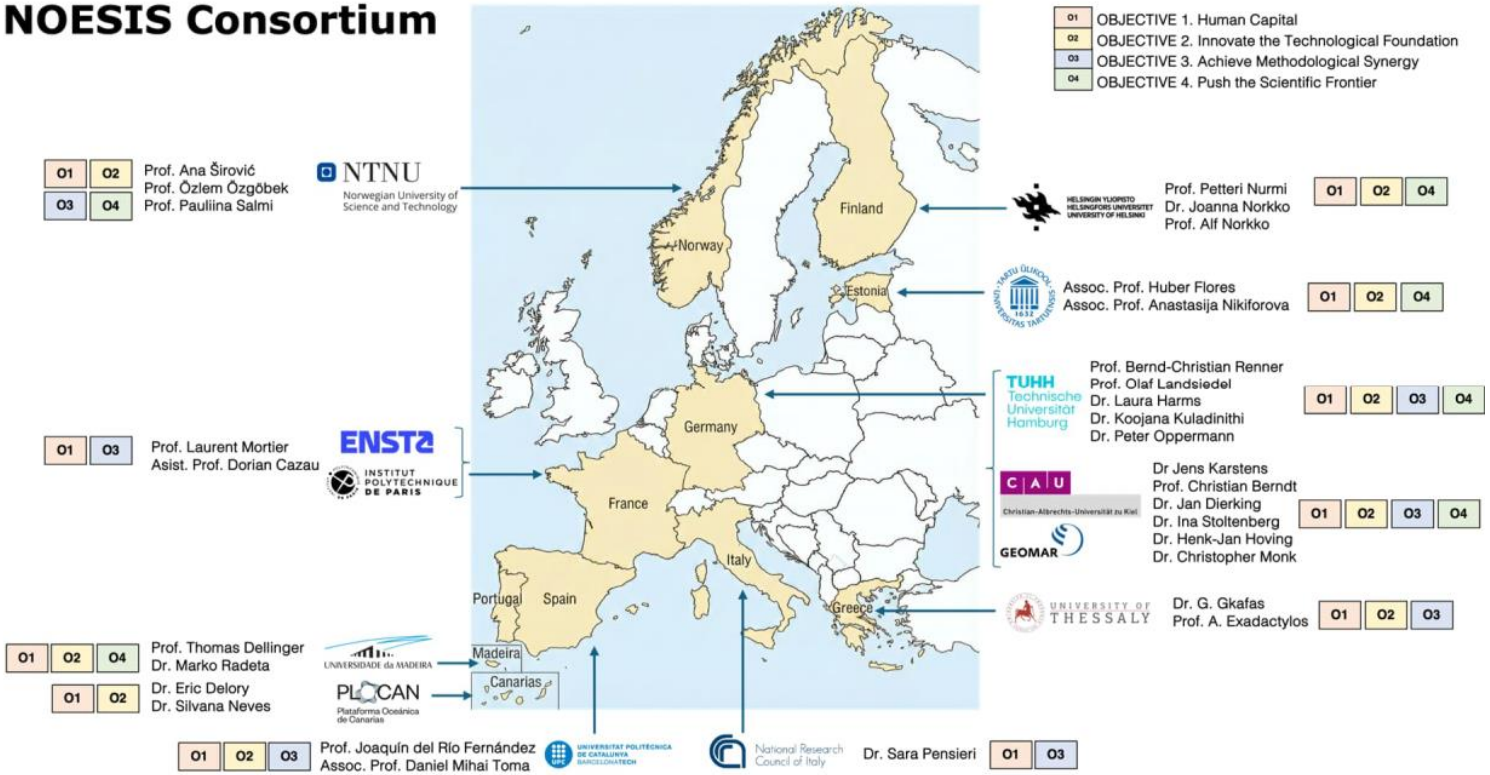
HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

SMU
SINGAPORE MANAGEMENT UNIVERSITY

CMU CHIANG MAI UNIVERSITY

MSCA-DN: Doctoral Network on Marine Sciences and AI

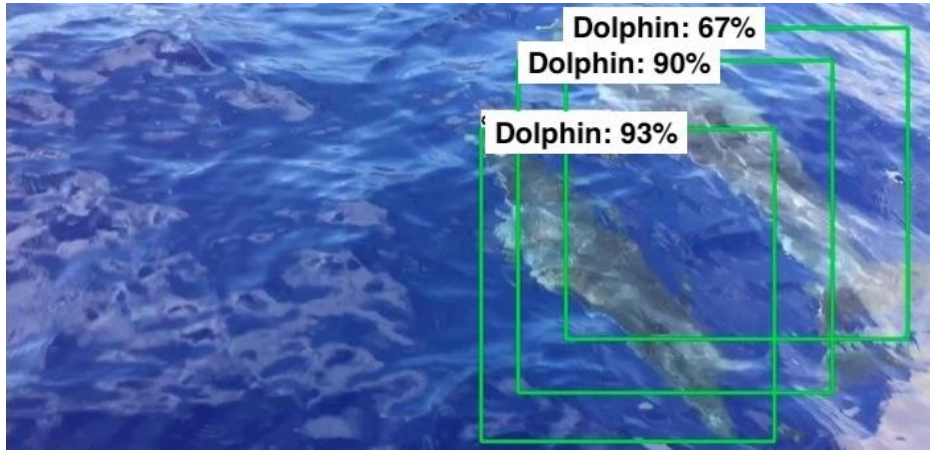
NOESIS Consortium

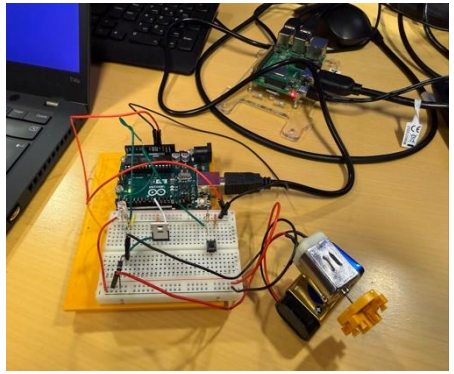


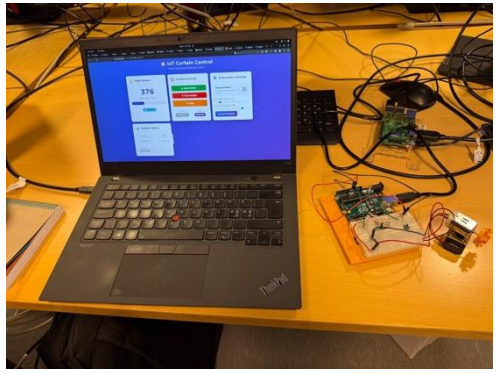
Summary

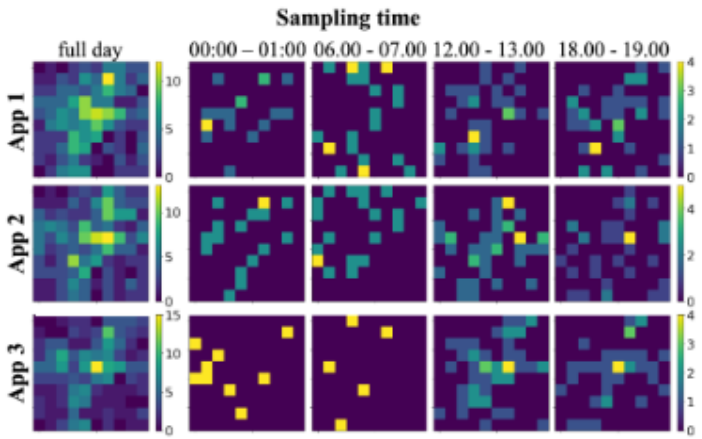
1. **The Edge is Inevitable**
   - Cloud-centric loops not possible for all real-world deployments.
   - Pushing intelligence directly to the sensor necessary, especially for intermittent / hostile environments.
2. **Reality is the Ultimate Adversary**
   - Human subjectivity, temporal drift, and extreme environmental hostility corrupt AI models.
   - Systems must be engineered to expect heavily degraded inputs and system unavailability.
3. **Isolated Solutions Cascade Issues**
   - Solutions targeting one part of the pipeline often cause issues elsewhere
   - Improvements in one task can create biases in another task
4. **Holistic Workflows are Mandatory**
   - Surviving the real-world requires optimizing the entire pipeline as a single, interconnected loop.

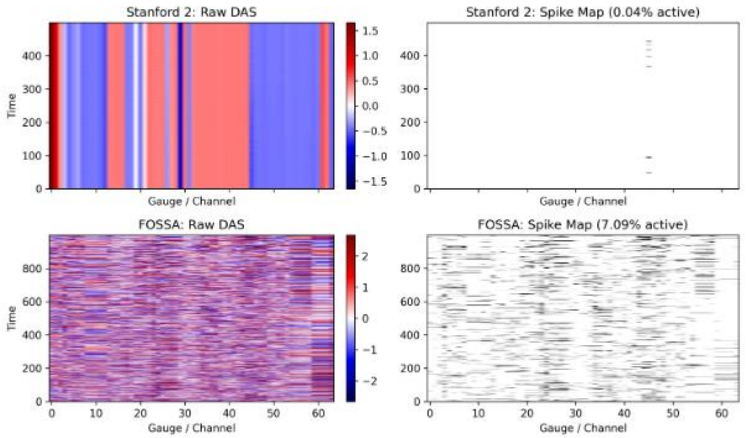
Thank you

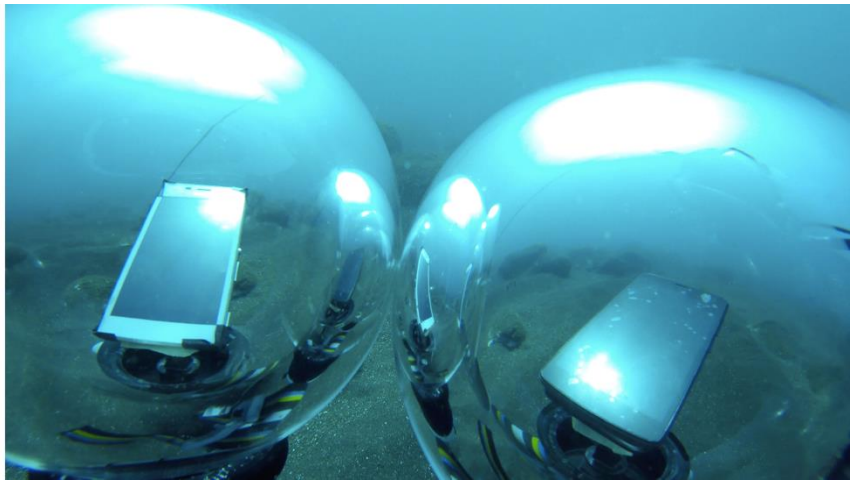












Petteri Nurmi

petteri.nurmi@helsinki.fi