Consiglio Nazionale delle Ricerche

Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo"

ISTI-Day 2025

18 November 2025 · Proceedings

ISTIday25

Chairs
Giulio Del Corso
Andrea Pedrotti

Contents

Consiglio Nazionale delle Ricerche

Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo"

ISTI DAY

18 NOVEMBRE 2025

AUDITORIUM + AULA A29

POSTER SESSION: H 11:30-13:00

MAIN SESSION: H 14:00-17:30

L'APPUNTAMENTO ANNUALE PER RACCONTARE E CONDIVIDERE LE ATTIVITÀ DELL'ISTI

AGENDA

11:30 - 13:00

Poster session

Oltre 20 poster esposti in Aula 29 con i loro autori a disposizione per raccontarvi dei loro lavori.

Per votare i migliori, il giorno dell'ISTI Day, usa il QR code

14:00

Saluti istituzionali

Emilio F. Campana - Direttore DIITET

14:20

Stato dell'Istituto

Roberto Scopigno - Direttore Isti

14:40

Comingio e premiazioni Isti-Award

Fabio Carrara - Isti

15:00

Invited talk

Diego Ceccarelli - Bloomberg

15:40

Coffee Break (Aula 29)

16:00

AI Generativa: opportunità e rischi nel rigore della ricerca

Tavola rotonda con

Marianna Bolognesi - UniBologna,
Andrea Esuli - Cnr-Isti, Dino Pedreschi - UniPisa

16:50

Premiazione poster, ringraziamenti e chiusura

ISTI-Day 2025: Overview

Introduction

ISTI-Day is an annual information and networking event organized by the Institute of Information Science and Technologies 'A. Faedo' (ISTI) of the Italian National Research Council (CNR). This event features an opening talk of the Director of the Dept. DIITET (*Emilio F. Campana*) as well as an overview of the Institute's activities presented by the ISTI Director (*Roberto Scopigno*). Those institutional segments are complemented by dedicated presentations and round tables featuring former staff members, as well as internal and external collaborators.

To foster a network of knowledge and collaboration among newcomers, the 2025 ISTI Day edition also includes a large poster session that provides a comprehensive overview of current research activities. Each of the 13 laboratories contributes 1–3 posters, highlighting the most innovative work and offering early-career researchers a platform for discussion. Thus these proceedings include the posters selected for *ISTI-Day* 2025, reflecting the diverse and innovative nature of the Institute's research.

Invited Talks

The invited talks of the 2025 edition were designed to offer a complementary view of generative AI across both industrial and academic fields. In the first talk, *Diego Ceccarelli* of Bloomberg showcased the company's generative AI approaches, detailing how they are being used to enrich financial news and make vast amounts of data more discoverable and liquid. Shifting to the academic domain, in the second talk *Marianna Bolognesi*, *Andrea Esuli* and *Dino Pedreschi* (moderated by *Katia Genovali*) discussed on the impact of generative AI tools on both the scientific production and the rigour of peer-reviewing processes.

Generative AI at Bloomberg

Diego Ceccarelli is an Engineering Manager at Bloomberg, leading a team developing search products that combine traditional information retrieval with large language models. Before joining Bloomberg, he was a researcher at ISTI-CNR and IMT Lucca. His work focuses on making information discovery faster, smarter, and more intuitive.

Generative AI: Opportunities and Risks in the Rigor of Research

Andrea Esuli is a Research Manager at the Institute of Information Science and Technologies of the Italian National Research Council in Pisa. His research interests are in the fields of machine learning, with a special focus on text, and information retrieval. In recent years he is working on large language models, with a focus on the automatic detection of AI-generated content and the benchmarking of generative AI systems.

Marianna Bolognesi is Full Professor of Linguistics at the University of Bologna and Principal Investigator of the ERC project ABSTRACTION. Her research combines cognitive and distributional semantics, focusing on how words shape semantic categories and abstraction in thought and communication. She previously held research positions at the Universities of Amsterdam (Marie Curie Fellow) and Oxford (AHRC-funded project on metaphor processing). She is also Vice-PI of the PRIN project WEMB, in collaboration with ISTI, and her work integrates linguistic theory, behavioural experiments, and computational modelling.

Dino Pedreschi is Full Professor of Computer Science at the University of Pisa and co-director of the Knowledge Discovery and Data Mining Laboratory (KDD), a joint initiative between the University of Pisa and ISTI. He is a founding member of several European initiatives on Artificial Intelligence, such as IRSoBigData and the HumanE-AI-Net network. He participated in the drafting of the AI Act, the first European document aimed at regulating the field of AI, and holds positions at a European level for studies on the impact of AI on online platforms.

Poster selection

The poster abstracts were evaluated by three ISTI researchers, members of different laboratories: *Danila Germanese* (SI), *Daniela Giorgi* (VC), and *Fabrizio Falchi* (AIMH). Authors of the selected abstracts subsequently developed their posters in accordance with the provided graphic guidelines.

Poster evaluation

The posters on display were evaluated blind by ISTI staff participating in the event. Using a recognition system based on institutional credentials (developed by *Giuseppe Lipari* and *Federico Volpini*), each participant could select up to three preferred posters, excluding those from their own laboratory. The five posters with the most votes were awarded.

Award-winning Posters

LoomNet: Enhancing Multi-View Image Generation via Latent Space Weaving

Giulio Federico*, Claudio Gennaro, Fabio Carrara, Giuseppe Amato, Marco Di Benedetto

AIMH

Abstract. Generating consistent multi-view images from a single image remains challenging. Lack of spatial consistency often degrades 3D mesh quality in surface reconstruction. To address this, we propose LoomNet, a novel multi-view diffusion architecture that produces coherent images by applying the same diffusion model multiple times in parallel to collaboratively build and leverage a shared latent space for view consistency. Each viewpoint-specific inference generates an encoding representing its own hypothesis of the novel view from a given camera pose, which is projected onto three orthogonal planes. For each plane, encodings from all views are fused into a single aggregated plane. These aggregated planes are then processed to propagate information and interpolate missing regions, combining the hypotheses into a unified, coherent interpretation. The final latent space is then used to render consistent multi-view images. LoomNet generates 16 high-quality and coherent views in just 15 seconds. In our experiments, LoomNet outperforms state-of-the-art methods on both image quality and reconstruction metrics, also showing creativity by producing diverse, plausible novel views from the same input.

#ImageTo3D, #MultiViewDiffusion, #NeuralRendering

AIMH ARTIFICIAL INTELLIGENCE FOR MEDIA AND HUMANITIES

Consiglio Nazionale delle Ricerche

Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo"

ISTIday25

LoomNet

Enhancing Multi-View Image Generation via Latent Space Weaving

Giulio Federico - Fabio Carrara - Claudio Gennaro - Giuseppe Amato - Marco Di Benedetto

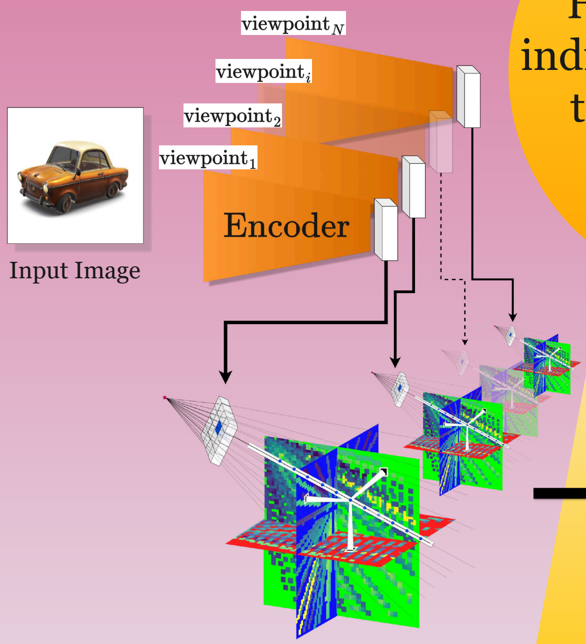
Input Image

16 Consistent Multiview Images in just 15 seconds!

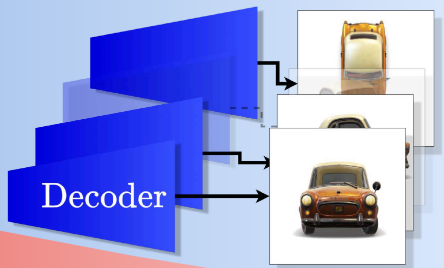
Scan to read our paper!

STEP 1: per-view splatting

Given an image and N desired viewpoints, each feature map is splatted onto three planes.

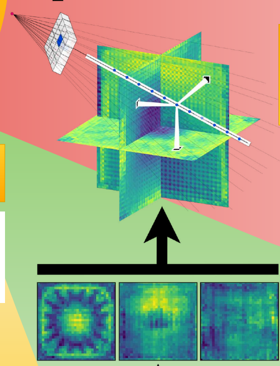


How can we align individual generations to reach a **visual compromise?**



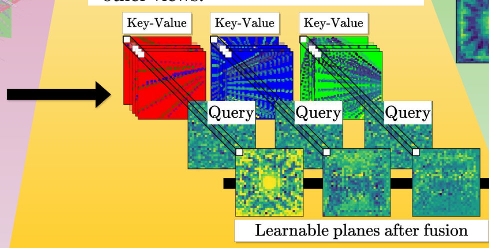
STEP 4: latent rendering

Re-render the feature map (hypothesis) using a common aggregation structure. This feature map accounts for all hypotheses.



STEP 2: fusion

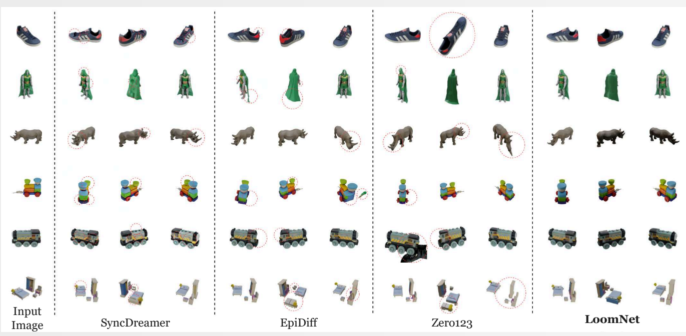
To reach a compromise among the hypotheses, each plane is fused with the corresponding ones from the other views.

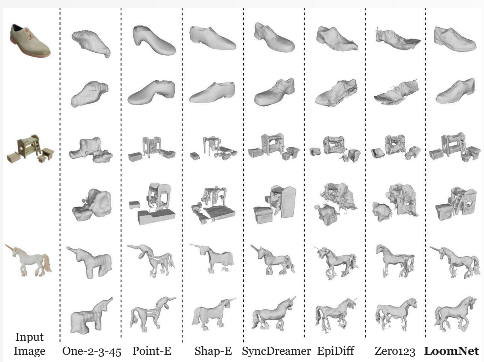


STEP 3: weaving

In regions where no hypothesis has been splatted, the space must be 'woven' using the surrounding information.

Experiments: we are **SOTA**!





Method	PSNR ↑	SSIM ↑	LPIPS ↓	Runtime ↓
Fixed Elevation				
Zero123 [22]	17.79	0.796	0.201	**7s**
SyncDreamer [23]	20.11	0.829	0.159	60s
EpiDiff [16]	20.49	0.855	0.128	17s
LoomNet (ours)	**21.60**	**0.901**	**0.070**	15s
Variable Elevation				
Zero123 [22]	15.91	0.772	0.231	**7s**
SyncDreamer [23]	15.90	0.773	0.246	60s
EpiDiff [16]	18.83	0.821	0.163	17s
LoomNet (ours)	**21.11**	**0.898**	**0.080**	15s

Table 1. **Multi-View Synthesis Results.** Quantitative comparison under two scenarios on the GSO dataset: **Fixed Elevation** (30° setting) and **Variable Elevation** (uniform setting). Metrics reported are PSNR (↑), SSIM (↑), LPIPS (↓) and Runtime (↓).

Method	Chamfer Dist. ↓	Volume IoU ↑
One-2-3-45 [21]	0.0768	0.2936
Point-E [32]	0.0570	0.2027
Shap-E [18]	0.0689	0.2473
Zero123 [22]	0.0543	0.3358
SyncDreamer [23]	0.0496	0.4149
EpiDiff [16]	0.0429	0.4518
LoomNet	**0.0260**	**0.5366**

Table 2. Quantitative comparison of surface reconstruction. We report Chamfer Distance (↓) and Volume IoU (↑) on the GSO dataset.

Securing Federated Learning against Extreme Model Poisoning Attacks via Multidimensional Time Series Anomaly Detection on Local Updates

Edoardo Gabrielli, Dimitri Belli*, Zoe Matrullo, Vittorio Miori, Gabriele Tolomei

WN

Abstract. Current defense mechanisms against model poisoning attacks in federated learning (FL) systems have proven effective up to a certain threshold of malicious clients (e.g., 25% to 50%). In this work, we introduce FLANDERS, a novel pre-aggregation filter for FL that is resilient to large-scale model poisoning attacks, i.e., when malicious clients far exceed legitimate participants. FLANDERS treats the sequence of local models sent by clients in each FL round as a matrix-valued time series. Then, it identifies malicious client updates as outliers in this time series by comparing actual observations with estimates generated by a matrix autoregressive forecasting model maintained by the server. Experiments conducted in several non-iid FL setups show that FLANDERS significantly improves robustness across a wide spectrum of attacks when paired with standard and robust aggregation methods.

#FederatedLearning, #ModelPoisoningAttacks, #AnomalyDetection

Securing Federated Learning against Extreme Model Poisoning Attacks via Multidimensional Time Series Anomaly Detection on Local Updates

Edoardo Gabrielli[a], Dimitri Belli[b], Zoe Matrullo[a], Vittorio Miori[b], Gabriele Tolomei[a,b]

[a]Department of Computer Science, Sapienza University of Rome, 00185, Rome (RM), Italy

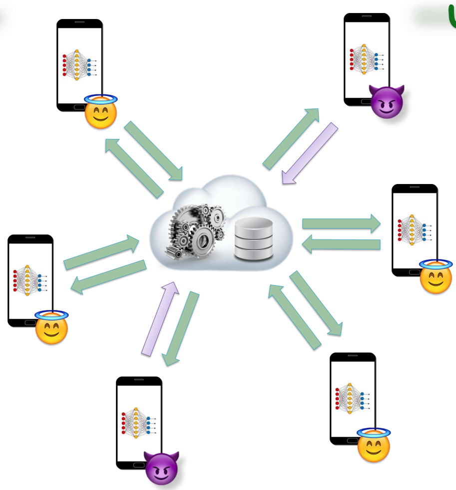
[b]Institute of Information Science and Technologies (ISTI), National Research Council (CNR), 56124, Pisa (PI), Italy

Core Idea in Federated Learning

Federated Learning (FL) allows to multiple edge clients to collaboratively train a shared, global model without disclosing their local private training data.

A typical FL round involves the following steps:

i. the server randomly picks some clients and sends them the current, global model;
ii. each selected client locally trains its model with its own private data; then, it sends the resulting local model to the server;
iii. the server updates the global model by computing an ***aggregation function***, on the local models received from clients.



Untargeted Model Poisoning Attacks

Adversaries attempt to tweak the global model weights by directly perturbing the local model's parameters of some infected clients 😈 before these are sent to the central server for aggregation. In doing so, the adversary aims to jeopardize the global model indiscriminately at inference time.

Problem

Existing defense mechanisms either rely on simple heuristics (e.g., Trimmed Mean and FedMedian) or need strong and unrealistic assumptions to work effectively (e.g., foreknowledge or estimation of the number of malicious clients, as for Krum/Multi-Krum, and Bulyan).

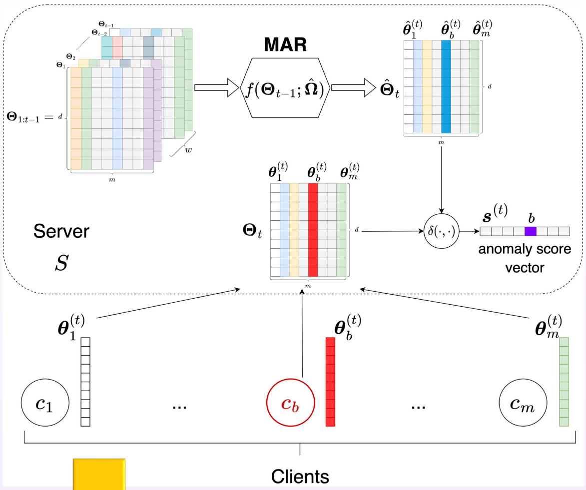
How can we overcome these limitations?

Our Solution

<u>F</u>ederated <u>L</u>earning meets <u>AN</u>omaly <u>DE</u>tection for a <u>R</u>obust and <u>S</u>ecure filter (FLANDERS)

FLANDERS is a pre-aggregation filter that the server can apply before any aggregation function. It does **not require knowing how many malicious clients there are,** and it **leverages the temporal dynamics** of local updates. In practice, it treats updates as a multidimensional time series and **flags as anomalies those that are less predictable**; legitimate updates tend to be far more regular than malicious ones.

Moreover, **FLANDERS handles cold start** by evaluating new clients against the global model until enough history is available.

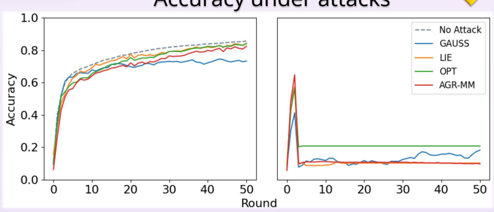


Step By Step

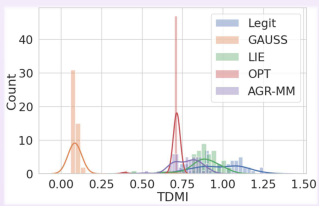
1. **Receive local updates**: Each device I sends its local model update u_i to the server
2. **Predict the expected update:** The server looks at the device's past updates and predicts what the next update should look like, called it $\hat{u}_i$
3. **Compute an anomaly score:** Compare the actual update u_i with the predicted update $\hat{u}_i$.
   - If they're similar → low score 👍
   - If they differ a lot → high score 👎
4. **Filter/select trustworthy updates:** Rank devices by their scores and keep (or weight higher) only the updates with low anomaly score. Discard or down-weight those with high scores.
5. **Aggregate to update the global model**: Use the selected, trusted updates to compute the new global model.
6. **Update history (learn for next round)**: Add this round's (cleaned) updates to the server's history so future predictions $\hat{u}_i$ get better.

Results

Accuracy under attacks

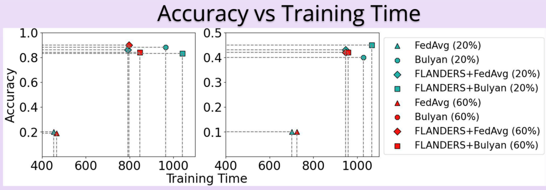


TDMI distributions



FLANDERS effectively detects and filters malicious updates, improving robustness of FL.
The plots show that under various poisoning attacks (GAUSS, LIE, OPT, AGR-MM) our filter restores the global model's accuracy close to the no-attack baseline (top-left) while suppressing abnormal behaviour of malicious clients (top-middle).
Distributions of Time-Delayed Mutual Information (TDMI) clearly separate legitimate from poisoned updates (top-right), enabling effective anomaly detection.
Finally, FLANDERS combined with FedAvg or Bulyan maintains high accuracy even with 60% malicious clients, with moderate additional training time (bottom).

Accuracy vs Training Time



Scan here to reach the paper

Scan here to reach the code

Improving Cybersecurity for Smart Home Systems

Tauheed Waheed*, Eda Marchetti, Antonello Calabrò

SEDC

Abstract. The Internet of Things (IoT) has been a game-changer for the industrial revolution, enabling more efficient and autonomous systems that have transformed industries. However, the recent challenges in IoT security and privacy benchmarks have had far-reaching consequences, impacting various industrial sectors and damaging user trust. IoT is revolutionizing our everyday lives by enabling the easy use of a wide range of smart devices, thereby improving our lifestyle, convenience, efficiency, and safety. It has been evident that people's and users' perspectives on technology have drastically changed over the years. Smart home systems consist of various interconnected devices that can be vulnerable to security risks, potentially compromising the integrity of the entire system. This research aims to address cybersecurity challenges and identify research gaps to enhance cybersecurity for smart home systems. The integration of a wide range of smart devices into a home network increases its overall complexity and the potential points of failure within the system. Managing and securing smart devices adequately is challenging because every IoT device has unique specifications and security protocols, resulting in a heterogeneous environment for smart home systems. We aim to highlight the significance and effects of testing on smart home systems, with a focus on the limitations of existing testing methodologies. The limitations in current testing methodologies have identified the lack of user involvement in the testing process. We have proposed our user-centric SCTM (Smart-home Cybersecurity Testing Methodology) and its behavioural model to leverage cybersecurity in smart home systems.

#Cybersecurity, #Smarthome, #Testing

IMPROVING CYBERSECURITY FOR SMART HOME SYSTEMS

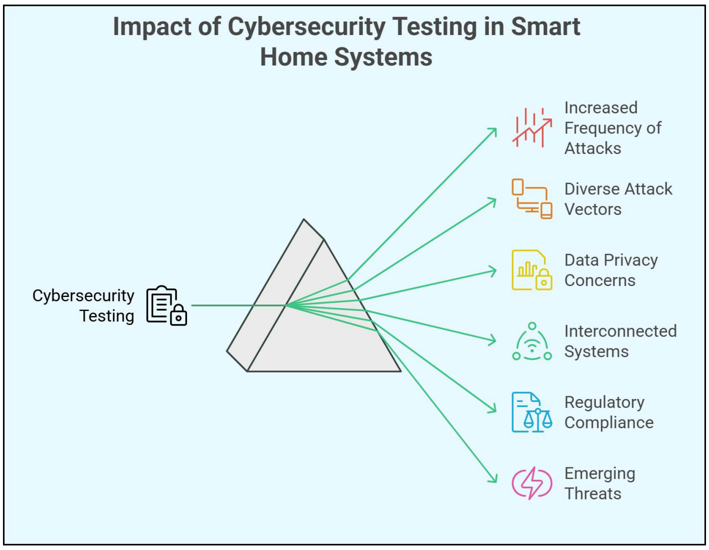
Tauheed Waheed
CNR-ISTI Pisa, Italy
tauheed.waheed@isti.cnr.it

Eda Marchetti
CNR-ISTI Pisa, Italy
eda.marchetti@isti.cnr.it

Antonello Calabro
CNR-ISTI Pisa, Italy
antonello.calabro@isti.cnr.it

Motivation

The primary objective is to encourage an environment where security is perceived as a shared responsibility among all participants from device manufacturers to end-users.



Impact

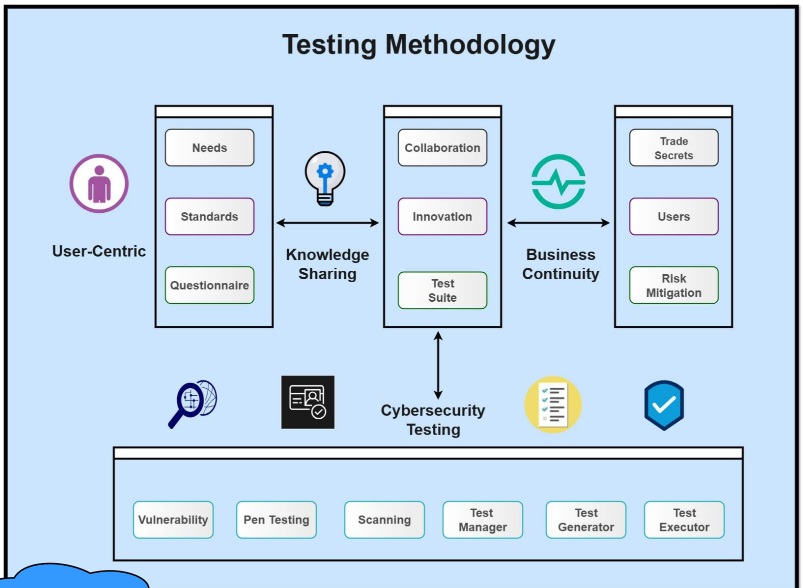
The approach not only leverages the security framework but also improves the overall usability and satisfaction of the users interacting with the system.

The work was partially supported by the project SERICS (PE00000014) under the PNRRP MUR program funded by the EU-NGEU.

Testing

Smart home systems consist of various interconnected devices that can be vulnerable to security risks.

Our holistic testing strategy strengthens defenses against cyber threats and drives continous improvement and adaptability in IoT infastructure of smart home systems.



Conclusion

Future work involves testing various IoT devices and anti-theft systems as it will broaden the scope of cybersecurity testing for smart home systems through SCTM (Smart-home Cybersecurity Testing Methodology).

YSocial: an AI-powered Social Media Virtual Twin

Giulio Rossetti, Massimo Stella, Rémy Cazabet, Katherine Abramski, Erica Cau, Salvatore Citraro, Andrea Failla*, Veronica Mesina, Virginia Morini, Valentina Pansanella

KDD

Abstract. This poster presents YSocial, a next-generation virtual twin of a social media platform designed to support the study of online interactions. Digital and virtual twins allow researchers to observe and manipulate complex dynamics in controlled settings. YSocial extends this concept to the domain of digital social networks, offering a high-fidelity, data-driven simulation environment capable of reproducing the social, communicative, and algorithmic processes typical of online platforms. At the core of YSocial is the integration of Large Language Models (LLMs), which make it possible to simulate agents with realistic behavior that adapt their actions to the context. This approach enables the generation of synthetic yet behaviorally coherent social data, useful for analyzing phenomena such as engagement, polarization, and the effects of recommendation systems. The system is organized into three main layers: (1) the agent layer, where LLM-based agents model individual behaviors; (2) the platform layer, which defines the operational characteristics of the system (the actions available to users, the logic of recommendation algorithms); and (3) the observation layer, which provides an interface for configuring, running, and analyzing simulations. By enabling controlled experiments in realistic social ecosystems without ethical risks, YSocial represents a bridge between computational modeling and the social sciences. It offers a flexible laboratory for studying emerging online phenomena without the privacy and accessibility limitations of real-world data. The poster present YSocial as a research infrastructure for computational social science, AI safety, and digital governance, illustrating how LLM-based digital twins can transform the way we understand and design online social systems. The YSocial project stems from a collaboration between ISTI-CNR, UniTn, and University Lyon 1.

#socialmedia, #llm, #socialsimulations

ISTIday25

YSocial: an AI-powered Social Media Virtual Twin

Giulio Rossetti[1], Massimo Stella[2], Rémy Cazabet[3], Katherine Abramski[1,4], Erica Cau[1,4], Salvatore Citraro[1], Andrea Failla[1,4], Veronica Mesina[1,4], Virginia Morini[1,4], Valentina Pansanella[1]

[1] KDDLab, ISTI-CNR, [2] University of Trento, [3] University of Lyon-1, [4] University of Pisa

Y Social in a nutshell

YSocial is an open-source platform designed to simulate **realistic social media environments** using **AI-powered** agents and support social science research.

With YSocial, researchers can design LLM-powered simulation scenarios where **agents interact with each other and with real content**. With it, scientists can study and analyze complex social and algorithmic phenomena occurring on online platforms (e.g., recommender systems effect, fake news diffusion).

Y Social: fine-grained customization

Name: Bob
Age: 54
Political leaning: Democrat
Toxicity: Low
Activity level: Medium

Name: Karen
Age: 34
Political leaning: Republican
Toxicity: High
Activity level: High

Media outlet (RSS feed)

CNN: New report reveals alarming rise in air pollution levels. Experts call for urgent measures to reduce emissions

Bob: Alarming but not surprising. We've ignored the warning signs for too long. Air pollution isn't just an environmental issue—it's a public health crisis.

Karen: Oh please, another "urgent" climate scare. These so-called experts have been crying wolf for decades while flying on private jets and cashing in on "green" deals. Air pollution isn't from hardworking Americans—it's from countries like China. Stop blaming us and start fixing *them.*

Content (e.g., reverse chronological) and **social** (e.g. common neighbors) **recommenders**

10+ user actions (like, share, search, follow, unfollow…)

Evolving agent interests based on content interaction

Simulation configuration, monitoring, and analytics

YSocial

Zero-code user interface
Easy simulation configuration and monitoring
Highly **customizable**

ySights

Open-source **Python** 🐍 library
Analysis and **visualization**
Advanced insights extraction

Smart Mirror for Cardiovascular Biomarkers Monitoring in Adolescents and Young Adult Cancer Survivors

Sara Colantonio*, Danila Germanese, Maria Antonietta Pascali

SI

Abstract. For a significant proportion of young cancer survivors, the end of therapeutic interventions marks the beginning of a lifelong journey toward recovery and self-management. Despite their motivation to return to normal life, Adolescents and Young Adults (AYA) often face long-term health challenges caused by late effects of cancer therapies. Cardiotoxicity, in particular, remains among the most frequent and severe late effects, increasing the risk of heart failure and premature mortality. It is therefore imperative that cardiovascular risk is detected early and proactively managed to enhance the long-term outcomes of survivors. There is a strong interest within the scientific community to develop and test innovative, holistic, and digital approaches that effectively support AYA cancer survivors in managing late cardiotoxic effects. Within the European project MAYA (Smart Mirrors Supporting Healthier Lives of Adolescents and Young Adults after Cancer)—funded under the Horizon Europe – Research and Innovation programme and coordinated by the University of Ioannina (Greece)—our team at the Signal and Image Laboratory (SILab) will bring complementary expertise to the development of the iCARE Health Hub—an integrated system combining a robust, trustworthy smart mirror and an AI-powered conversational agent—that integrates the data collected by multiple minimally-invasive sensors to monitor cardiovascular-related parameters continuously. Building on their experience from the FP7 SEMEOTICONS project, the SILab team will deploy real-time biomarker monitoring and advanced AI-driven analytics to evaluate cardiovascular health and overall well-being in participating AYA survivors, enabling effective management of modifiable risk factors such as hypertension, diabetes, and obesity. Special attention will be given to maintaining data integrity, building user trust, and guaranteeing long-term usability. The project will validate these innovative solutions in real-world conditions through clinical studies involving cancer survivors across several European countries. By providing accurate, real-time insights into modifiable risk factors such as hypertension, diabetes, and obesity, the smart mirror will contribute to reducing cardiovascular risk and improving overall well-being among AYA cancer survivors.

#SMARTMIRROR, #TrustworthyAI, #Personalizedmedicine

Consiglio Nazionale delle Ricerche
Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo"
ISTIday25

Smart Mirror for Cardiovascular Biomarkers Monitoring in Adolescents and Young Adult Cancer Survivors

Sara Colantonio, Danila Germanese, Maria Antonietta Pascali*

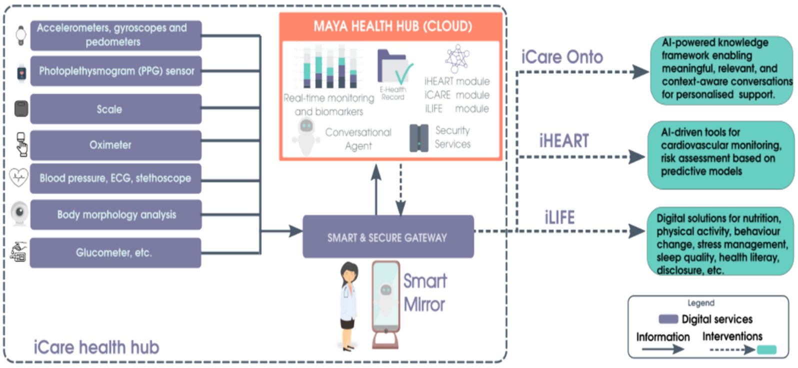
BACKGROUND

After treatment, many adolescent and young adult (AYA) cancer survivors face long-term health challenges, particularly **cardiotoxicity**, which increases the risk of heart failure and premature death.
Early detection and proactive management of cardiovascular risk are crucial.

There is growing interest in **innovative, holistic, and digital approaches** to help AYA survivors monitor and manage late cardiotoxic effects, improving long-term outcomes.

OBJECTIVES

Development of the **iCARE Health Hub**: an integrated system combining a robust, trustworthy **smart mirror** and an AI-powered conversational agent that integrates the data collected by multiple minimally invasive sensors to monitor cardiovascular biomarkers continuously.



METHODS

- Real-time biomarker monitoring & advanced AI analytics
- Focus on AYA cancer survivors' cardiovascular health and management of modifiable risk factors, such as hypertension, diabetes, and obesity
- Emphasis on data integrity, user trust, and long-term usability
- Validation through multi-country clinical studies across Europe

IMPACT

The smart mirror will provide real-time insights into modifiable risk factors such as hypertension, diabetes, and obesity.
By empowering AYA cancer survivors to monitor their health continuously, it will support the reduction of cardiovascular risk and promote overall well-being

Posters

Talking to DINO: Bridging Self-Supervised Vision Backbones with Language for Open-Vocabulary Segmentation

Lorenzo Bianchi*, Fabio Carrara, Nicola Messina, Fabrizio Falchi

AIMH

Abstract. Open-Vocabulary Segmentation (OVS) is about teaching AI to find and outline objects in images based on any text description—for example, segmenting "a red umbrella" or "a wooden chair" even if those specific categories weren't in its training data. Current models like CLIP can connect images with text, but they only have a rough sense of where things are in an image. On the other hand, visual models like DINO are great at understanding fine details and structure, but don't understand language. Our method, called Talk2DINO, combines the strengths of both. It uses DINOv2 for detailed visual understanding and CLIP for language grounding. We train a lightweight mapping that links CLIP's text representations with DINOv2's visual features—without retraining the original models. During training, we use DINOv2's attention maps to match parts of the image with relevant text concepts. The result is a model that produces cleaner, more accurate segmentations and better separates objects from the background. Across several benchmarks, Talk2DINO achieves state-of-the-art results in open-vocabulary segmentation.

#OpenVocabularySegmentation, #UnsupervisedLearning

ISTIday25

Consiglio Nazionale delle Ricerche

Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo"

AIMH Artificial Intelligence for Media and Humanities

Bridging Self-Supervised Vision Backbones with Language for Open-Vocabulary Segmentation

Code:

Luca Barsellotti[*1], Lorenzo Bianchi[*2,3], Nicola Messina[2], Fabio Carrara[2], Marcella Cornia[1], Lorenzo Baraldi[1], Fabrizio Falchi[2], Rita Cucchiara[1]

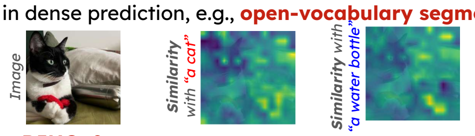
[1]University of Modena and Reggio Emilia, Italy [2] ISTI-CNR, Italy [3]University of Pisa, Italy

[1]{name.surname}@unimore.it [2]{name.surname}@isti.cnr.it

Problem Formulation

CLIP excels in global vision-language tasks, but:

- **its patch embeddings lack local semantics,**
- limited in dense prediction, e.g., **open-vocabulary segmentation.**

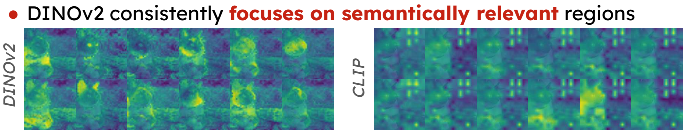


In contrast, **DINOv2**:

- has **spatially consistent** patches w/ **strong local semantics**, but
- lacks a **bridge to language**!

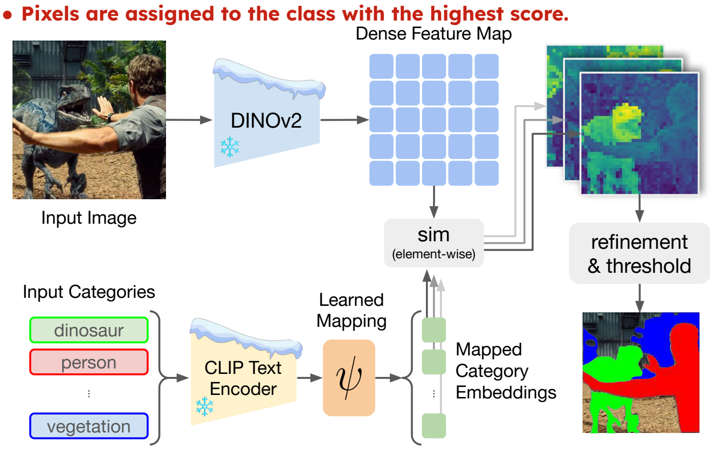
Attention maps per head further reveal the contrast:

- The maps of CLIP are diffuse and **harder to interpret**
- DINOv2 consistently **focuses on semantically relevant** regions



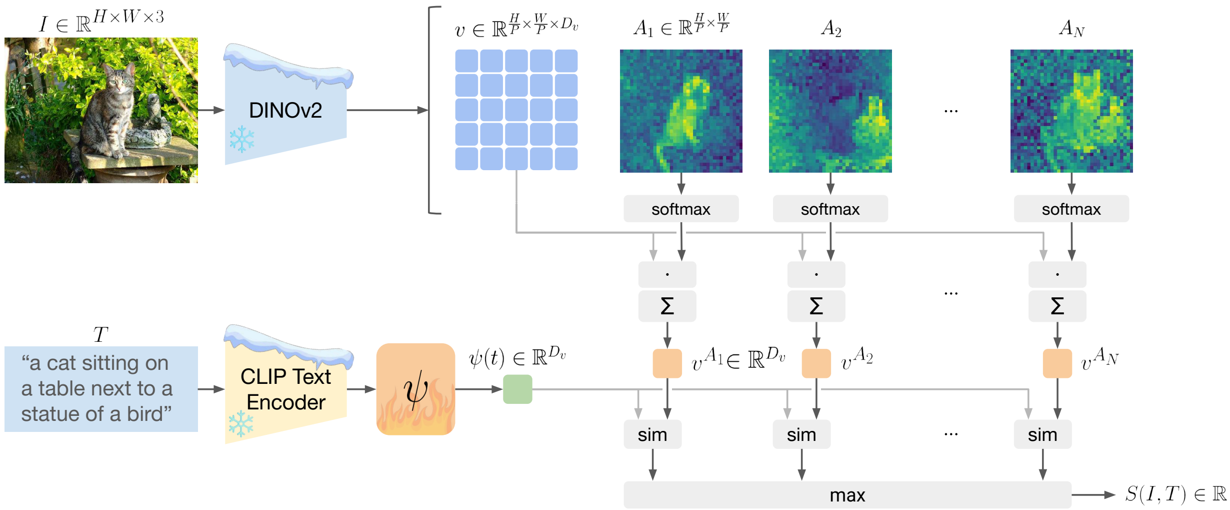
Our Proposal: Mapping CLIP into DINOv2 space

- We propose to learn a ***simple* mapping** between the text embeddings and the DINOv2 patch space.
- At inference, **we compute the similarity between each DINOv2 patch and each projected text embedding**, producing a similarity map per class.
- **Pixels are assigned to the class with the highest score.**

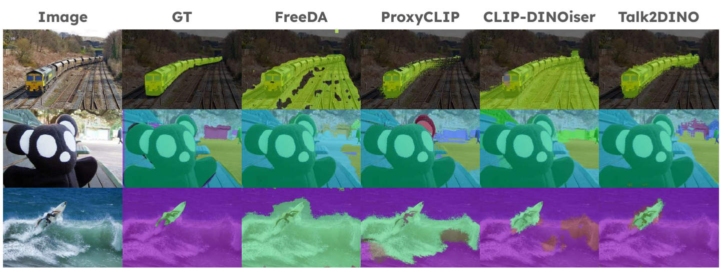


Training

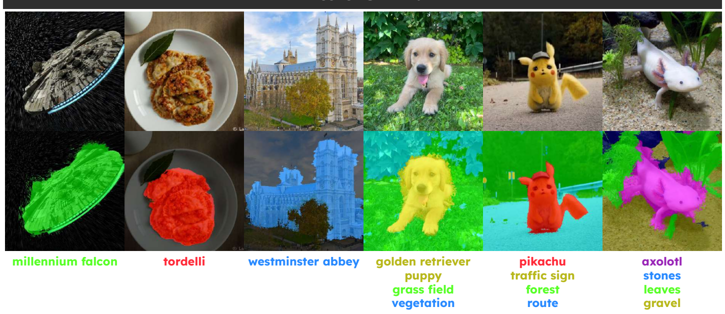
- **We train the mapping through a contrastive learning over a large set of image-caption pairs (COCO).**
- Differently from traditional contrastive learning, for each caption, **the projected text embedding is aligned with a weighted average of the image's patch embeddings**.
- Instead of a uniform average, we exploit DINOv2 attention maps: **the most relevant head for the caption is selected, and its weighted patch features are used in a contrastive loss**.



Qualitatives



Into the Wild



Comparison with State-of-the-Art

Method	Backbone	V20	C59	Stuff	City	ADE	V21	C60	Object	Avg
CLIP-DINOiser	CLIP	81.5	37.1	25.3	31.5	20.6	64.6	33.5	36.1	41.3
FreeDA	CLIP + DINOv2	85.2	42.1	27.0	33.8	21.8	51.8	37.4	38.6	42.2
ProxyCLIP	CLIP + DINO	80.3	39.4	26.9	**38.6**	20.2	60.8	35.3	37.2	42.3
Talk2DINO	DINOv2	**88.5**	**42.4**	**30.2**	38.1	**22.5**	**65.8**	**37.7**	**45.1**	**46.3**

Talk2DINO achieves SOTA results on 7 open-vocabulary semantic segmentation benchmarks!

One Patch to Caption Them All: A Unified Zero-Shot Captioning Framework

Lorenzo Bianchi, Giacomo Pacini*, Fabio Carrara, Nicola Messina, Giuseppe Amato, Fabrizio Falchi

AIMH

Abstract. Recent advances in artificial intelligence allow machines to describe images using natural language—a task known as image captioning. Most captioning models, however, need large datasets of images paired with text, and usually generate only one caption for the entire image. Patch-ioner is a new framework that breaks this limitation by teaching AI to describe any part of an image—from a small detail to a complex region—without requiring region-specific training data. Instead of seeing an image as a single whole, Patch-ioner views it as a collection of small "patches" and learns to describe them individually and in combination. This patch-based approach enables more flexible, fine-grained, and scalable caption generation, even in a zero-shot setting where no examples are provided. Experiments show that Patch-ioner can describe objects, regions, and even user-drawn traces more accurately than previous models, setting a new direction for open-world visual understanding and image-language reasoning.

#IMAGE-LANGUAGEREASONING

Code & Data

Code available on

Try the demo on

One Patch to Caption Them All

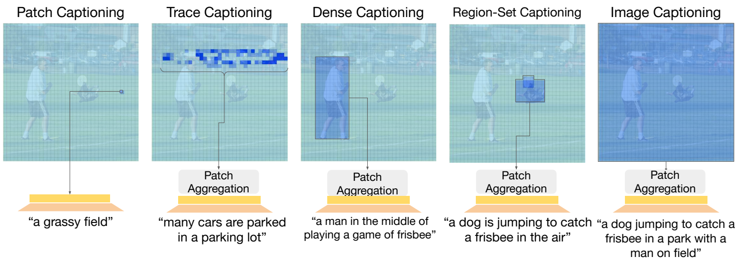
A Unified Zero-Shot Captioning Framework

Bianchi L.*[1,2], Pacini G.*[1,2], Carrara F.[1], Messina N.[1], Amato G.[1], Falchi F.[1]

[1]ISTI-CNR, Pisa [2]University of Pisa

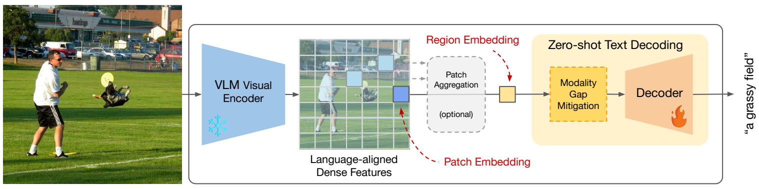
Abstract

Patch-ioner is a **zero-shot captioning** framework that shifts from an image-centric to a patch-centric approach, enabling flexible caption generation without region-level supervision.



- Treats individual patches as *atomic captioning units*, allowing descriptions at any granularity.
- Leverages language-aligned dense visual representations for adaptable zero-shot captioning.

Patch-ioner



- We train a text decoder to generate captions from textual dense representations.
- At inference phase, we get the patches embeddings of DINOv2
- We (optionally) aggregate the patches, depending on the task, then
- We project those embeddings in the textual space and use the decoder to caption them

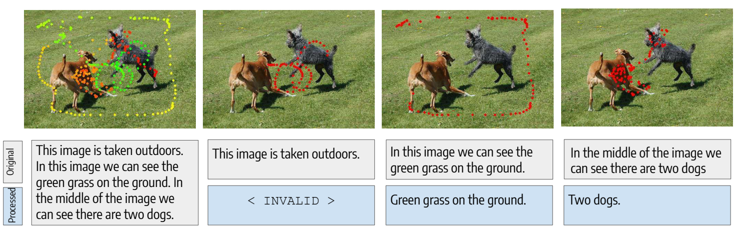
Vision-Language Backbones Analysis

The framework flexibility enables to adopt different combinations of backbones and captioning models.

Captioning Task: (Dataset)		Trace (COCO)		Dense (VG v1.2)		Region-Set (COCO Entities)		Image (COCO)		
Encoder	Backbone	C	P	C	P	C	P	C	P	CLIP-S
CLIP	CLIP B/16	10.9	75.0	10.9	74.2	41.6	78.8	42.1	84.0	66.2
DenseCLIP	CLIP B/16	18.6	75.3	19.9	75.2	51.0	77.6	28.0	77.0	57.3
INViTE	CLIP B/16	13.8	76.4	16.8	77.3	43.3	78.9	21.3	79.1	60.6
ProxyCLIP	DINO B/8 + CLIP B/16	16.7	75.7	15.7	76.0	41.2	78.4	28.7	79.0	61.7
ProxyCLIP	DINOv2 B/14 + CLIP B/16	16.5	75.7	15.5	76.0	40.6	78.5	27.4	78.6	61.0
DINO.txt	DINOv2 B/14	<u>23.2</u>	**78.8**	<u>23.4</u>	<u>78.5</u>	<u>91.8</u>	<u>86.3</u>	<u>67.8</u>	<u>87.2</u>	<u>70.8</u>
Talk2DINO	DINOv2 B/14	**27.9**	<u>78.7</u>	**31.9**	**78.8**	**109.1**	**87.5**	**69.2**	**87.4**	**72.8**

Tab.1 **Vision-Language Backbones.** CIDEr (C) and RefPAC-S (P) across four captioning tasks varying backbone.

Trace Captioning Datasets



Results

Model	Trace		Dense			Region-Set		Image		
	C	P	mAP	C	P	C	P	C	P	CLIP-S
Whole-image Zero-shot Captioners										
ZeroCap (Tewel et al., 2022) CVPR'22	-	-	-	-	-	-	-	14.6	-	-
MAGIC (Su et al., 2022) ArXiv'22	-	-	-	-	-	-	-	49.3	-	-
ViECap† (Fei et al., 2023) ICCV'23	24.3	74.4	14.90°	26.4°	74.3°	102.7	85.0	89.7	88.5	75.6
MeaCap† (Zeng et al., 2024) CVPR'24	22.5	74.4	15.01°	28.6°	75.1°	97.9	85.2	86.0	88.6	77.8
DeCap† (Li et al., 2023b) ICLR'23	20.5	75.3	17.75°	24.6°	77.8°	95.1	<u>87.4</u>	87.4	90.6	**79.3**
With Region-level Supervision										
RegionCLIP (Zhong et al., 2022) CVPR'22 + Mem. (≃ DeCap)	-	-	15.85	21.7	76.7	-	-	93.4	**91.2**	77.5
AlphaCLIP (Sun et al., 2024) CVPR'24 + Mem. (≃ DeCap)	21.3	75.4	14.63	19.1	73.9	95.1	<u>87.4</u>	89.7	<u>91.1</u>	<u>78.2</u>
Patch-ioner (Our Patch-based Framework)										
T2D + Mem. (≃ DeCap)	27.9	<u>78.7</u>	**21.31**	**31.9**	<u>78.8</u>	<u>109.1</u>	**87.5**	88.5◊	90.2◊	76.0◊
T2D + Noise (≃ CLOSE, CapDec)	**29.3**	78.1	<u>20.26</u>	26.3	77.0	97.5	85.6	65.5	86.2	70.9
T2D + Noise + External knowledge (≃ ViECap)	<u>28.2</u>	78.2	18.43	<u>30.3</u>	<u>77.8</u>	**109.3**	86.7	88.5◊	89.2◊	73.7◊
T2D + Noise + Filtered knowledge (≃ MeaCap)	27.4	**78.8**	18.66	**31.9**	**78.9**	104.4	86.9	83.0◊	89.6◊	74.8◊

Future Works

- Enforce an image-level captioning loss to DINO-based contrastively learned representations to obtain better patch-level features
- Explore different aggregation strategies

Tasks

PATCH New

-	-	GT
a tennis player is playing tennis on the court for a serve.	a group of people in a kitchen are cooking food	DeCap
a couple of people are in the middle of a tennis court.	a couple of people that are standing around each other	DeCap Patches
a street light in front of a large building.	**a forest with trees in the background.**	🔥**Our**

TRACE New

Two giraffes, rocks, and a fence	A sky.	GT
a giraffe in a zoo with a city in the background	a giraffe in a zoo with a city in the background.	DeCap
there are some people that are in a lot by a tree	there are some people that are out by a lot of trees.	DeCap Patches
two giraffes standing in a fenced area	**A view of a city with a sky in the background.**	🔥**Our**

DENSE

A clock at a train station.	a white ceiling fan hanging in the kitchen	GT
A train traveling along the platform of a public train.	a kitchen with a large refrigerator , cabinets and stove	DeCap
A train is on the tracks and going by.	a kitchen has a lot of fridge and a stove in it	DeCap Patches
A black cat is leaning on a black cat.	a bathroom sink with a variety of toilet above the wall.	DeCap Crops
A clock on a train station platform above a train.	**a ceiling fan is hanging in the kitchen**	🔥**Our**

REGION-SET

A brown-haired woman is pushing a baby stroller.	a man swinging a baseball bat as another looks on.	GT
a man and a child walking in the street while holding a stroller.	a baseball player at bat getting ready to hit the ball.	DeCap
there are some cars and a man about to go down the street.	some baseball players are on the field playing baseball.	DeCap Patches
a woman pushing a stroller with a child inside.	**a baseball player is swinging his bat as a crowd watches**	🔥**Our**

IMAGE

A man in a wetsuit rides a wave.	A woman with blond-hair is sitting in a booth with a drink working on her laptop	GT
A man on a surf board riding a wave in the water.	a woman sitting at a table using a laptop.	DeCap
A man surfing in the area 0.	a reader's writing on a laptop on desk-mounted computer	ZeroCap
A man on a surf board riding a wave in the ocean.	a woman sitting at a table with a laptop and a drink	CLOSE
A man on a surfboard riding a wave.	**a woman sitting at a cafe using her laptop**	🔥**Our**

Evaluating Bias on Encyclopedic Knowledge in Large Language Models

Maria Cassese*, Giovanni Puccetti, Andrea Esuli

AIMH

Abstract. Large Language Models (LLMs) acquire knowledge through exposure to vast and diverse textual corpora, where high-frequency occurrences influence the strength and accuracy of information retention. However, these models are often employed by users as a substitute for search engines, or even as oracles to query about the fundamental questions about the life, the universe, and everything. Therefore, it is essential to quantify their encyclopedic knowledge, namely the ability to provide single authoritative entries for relevant concepts, irrespective of their popularity or frequency of mention. In this study, we evaluate the extent to which LLMs acquire and report encyclopedic information and explore the forms of bias that can arise in the process. Specifically, we investigate whether LLMs can reliably retrieve factual data—such as the names of prime ministers of all countries—as recorded in Wikipedia, and assess potential biases related to web popularity, geography, gender, and economic status. Our evaluation spans multiple LLMs, including both text-only and multimodal architectures, across a broad set of topics, including both generic and culturally determined entities.

#LLM, #CulturalBias, #EncyclopedicKnowledge

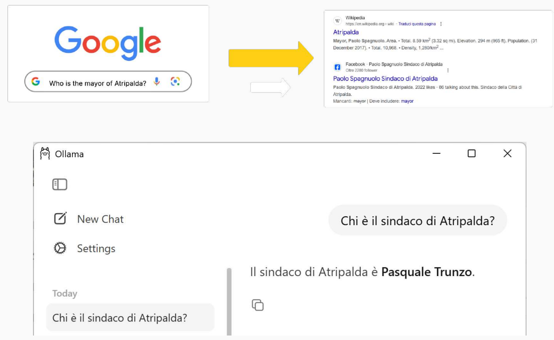
Evaluating Bias on Encyclopedic knowledge in Large Language Models

Maria Cassese, Giovanni Puccetti, Andrea Esuli
Istituto di Scienza e Tecnologie dell'Informazione, CNR

Motivation

- Large Language Models (LLMs) acquire knowledge through exposure to vast and diverse textual corpora, where **high-frequency occurrences** influence the strength and accuracy of information retention.
- This frequency-based learning contrasts with the nature of **encyclopedic knowledge**, which is curated to provide **single, authoritative entries for relevant concepts**, irrespective of their popularity or frequency of mention.

Example: retrieval vs generation



Research Questions

- **RQ1:** To which extent LLMs **acquire and report encyclopedic information?**
- **RQ2:** Can LLMs reliably recall factual data as recorded in Wikipedia equally **with respect of each country of the world?**
- **RQ3:** How much **web popularity, geography, gender, and economic status** affect the bias?

Dataset from Wikidata

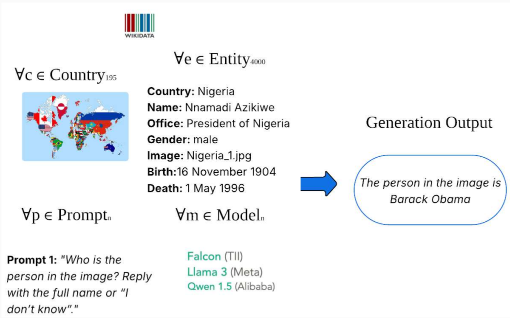
Fields:

1. Entity name
2. Country
3. Image
4. Gender
5. Date

Entity	Number	Countries
Prime Minister	2045	122
President	1814	56
Olympic Medalist	29073	170
Film director	25402	161

Methodology

- Extract entities for every country worldwide.
- Design progressively more granular prompts (e.g., name, image, name+image, name+image+birth/death date...)
- Generate responses for each prompt (and model, e.g., Llama, LLaVA, Falcon, Qwen).
- Compare the differences in accuracy across relevant dimensions (e.g., geography, gender...).

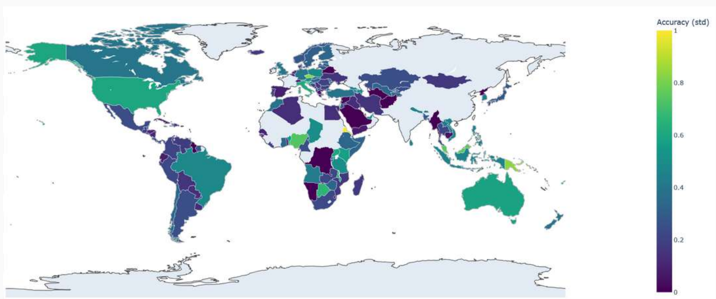


Prompt	Modality	Model	acc_std	ans_acc	coverage
1	text-only	llama3-llava-8b	0.02	0.05	0.44
2	text-only	llama3-llava-8b	0.05	0.34	0.17
3	VL	llama3-llava-8b	0.009	0.03	0.26
4	VL	llama3-llava-8b	0.04	0.066	0.66
5	VL	llama3-llava-8b	0.09	0.17	0.519
1	text-only	llava-1.5-7b	0.44	0.56	0.79
2	text-only	llava-1.5-7b	0.11	0.24	0.45
3	VL	llava-1.5-7b	0.01	0.02	0.56
4	VL	llava-1.5-7b	0.01	0.105	0.13
5	VL	llava-1.5-7b	0.05	0.27	0.21
1	text-only	Llama-3.1-8B-In	0.01	0.019	0.76
2	text-only	Llama-3.1-8B-In	0.24	0.49	0.50
1	text-only	Qwen2.5-7B-In	0.00	0.001	0.48
2	text-only	Qwen2.5-7B-In	0.07	0.36	0.195
1	text-only	Falcon3-7B-In	0.10	0.29	0.34
2	text-only	Falcon3-7B-In	0.09	0.217	0.428

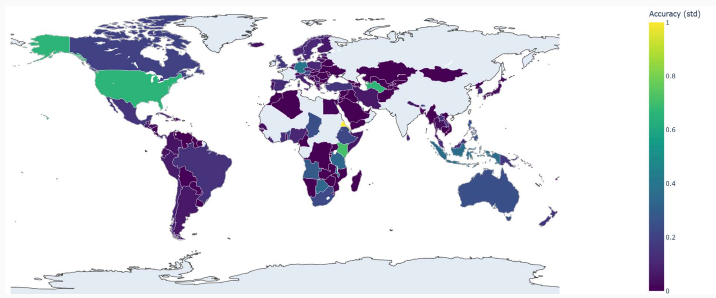
(A) Table of results for President and Prime Minister.

Initial Results

Llama-3 results on the President and Prime Minister queries



Qwen results on the President and Prime Minister queries



Cherry-picking made Legal: Transductive Model Selection under Prior Probability Shift

Lorenzo Volpi*, Alejandro Moreo, Fabrizio Sebastiani

AIMH

Abstract. Transductive learning is a supervised machine learning task in which, unlike in traditional inductive learning, the unlabelled data that require labelling are a finite set and are available at training time. Similarly to inductive learning contexts, transductive learning contexts may be affected by dataset shift, i.e., may be such that the assumption according to which the training data and the unlabelled data are independently and identically distributed (IID), does not hold. We propose a method, tailored to transductive classification contexts, for performing model selection (i.e., hyperparameter optimisation) when the data exhibit prior probability shift, an important type of dataset shift typical of anti-causal learning problems. In our proposed method the hyperparameters can be optimised directly on the unlabelled data to which the trained classifier must be applied; this is unlike traditional model selection methods, that are based on performing cross-validation on the labelled training data. By tailoring model selection to the actual test distribution, our approach contributes to the trustworthiness of AI systems, as it enables more reliable and robust classifier deployment under changed conditions.

#ProbabilityShift, #HyperparameterOptimization

Cherry-picking made Legal:
Transductive Model Selection under Prior Probability Shift

Lorenzo Volpi, Alejandro Moreo, Fabrizio Sebastiani
E-mail:{firstname.lastname}@isti.cnr.it

Context-aware model selection for robust AI models

- A key requirement for **trustworthy AI** is **robustness** under **dataset shift**, as models validated under the IID assumption often face reliability issues when this assumption is not verified (i.e., when dataset shift is at play).
- More specifically, we assume **prior probability shift** (anti-causal learning), in which $P(Y) \neq Q(Y)$ and $P(X|Y) = Q(X|Y)$.
- **Idea:** estimate accuracy on the specific test set using techniques from **Classifier Accuracy Prediction** (CAP).
- CAP is the task of **predicting the accuracy** $A(h, U)$ that a classifier h trained on $L \sim P$ will have **on an unlabelled set** $U \sim Q$, where P and Q are two unknown probability distributions related to each other via dataset shift.

Inductive vs. Transductive

Induction: Inferring **from particular to general** and then **from general to particular**

"*If you possess a restricted amount of information for solving some problem, try to solve the problem directly and never solve a more general problem as an intermediate step. It is possible that the available information is sufficient for a direct solution but is insufficient for solving a more general intermediate problem.*" (V. Vapnik)

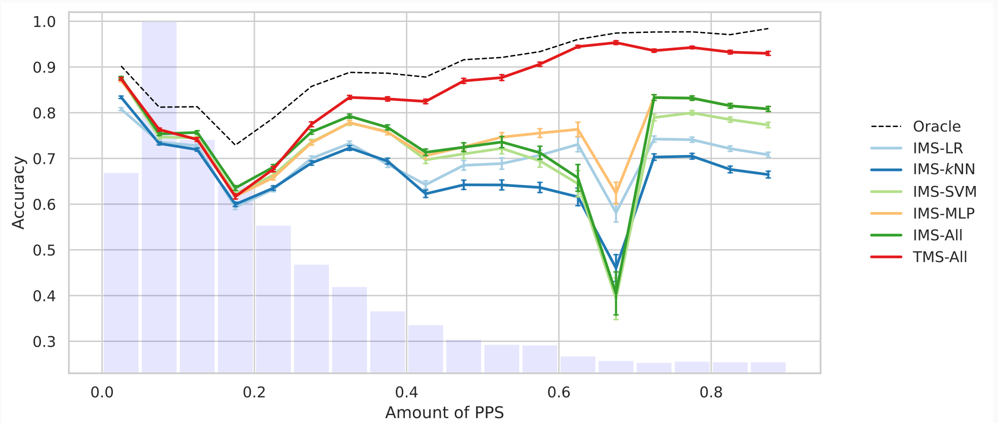
Transduction: Inferring **from particular to particular**

Model Selection

- Consider the class of hypotheses $\mathcal{H}$ for differently parameterized classifiers h_θ
- **Setup:** training data L and validation data V drawn from P; test data U drawn from Q
- **Goal:** choose the best performing $h_\theta \in \mathcal{H}$ for the test data U

Inductive Model Selection		Transductive Model Selection
Train all $h_\theta \in \mathcal{H}$ (1)	L	(1) Train all $h_\theta \in \mathcal{H}$
Compute acc. of all $h_\theta \in \mathcal{H}$ (2) Select h_θ^* (3)	V	(2) Train all CAP methods ψ_{h_θ}
Apply h_θ^* (4)	U	(3) **Estimate** acc. of all h_θ via ψ_{h_θ} (4) Select h_θ^* (5) Apply h_θ^*

Analysing accuracy of selected classifiers



Experimental Setting

- 4 classifiers: LR, kNN, SVM, and MLP
- 29 UCI ML datasets:
  - L (35%), V (35%), U (30%)
- **Artificial Prevalence Protocol**:
  - 1000 samples
  - 100 items
  - prior probability shift
- Evaluation measure: **Accuracy**
- CAP method: **LEAP** [1]

Conclusions

- TMS enables the selection of classifiers that perform **robustly under distribution shift**, a key requirement in **trustworthy AI**.
- **Replace** accuracy **computation** on training with accuracy **estimation** on unlabelled data.
- **Future work**:
  - Other types of datasets shift (e.g., **covariate shift**) supported by effective CAP methods.
  - TMS applied to TAR, online content moderation, etc.

References

1. Volpi, L., Moreo, A., and Sebastiani, F. (in press). LEAP: Linear equations for classifier accuracy prediction under prior probability shift. *Machine Learning*. Accepted for pubblication.

Code

PSET: a Phonetics-Semantics Evaluation Testbed

Gianluca Sperduti*, Dong Nguyen

AIMH

Abstract. We introduce the Phonetics-Semantics Evaluation Testbed (PSET), a new English-based testbed to evaluate phonetic embeddings. Our testbed is built on the assumption that phonetic embeddings should always prioritize phonetics over semantics, and it therefore leverages homophones and synonyms. We use PSET to test three phonetic embedding models: articulatory embeddings, Phoneme2Vec, and XPhoneBERT. The phonetic-based embeddings solve the task with varying degrees of success, with Phoneme2Vec performing the best. We also test five recent LLMs, GPT-4o, Gemini 2.5 Flash, Llama 3.1-8B, OLMo-7B and OLMo 2-7B. Gemini 2.5 Flash performs better than the other models. With this testbed, we hope to advance the development and evaluation of phonetic embedding models.

#PhoneticEmbeddings, #Phonetics, #Dataset

AIMH ARTIFICIAL INTELLIGENCE FOR MEDIA AND HUMANITIES

Consiglio Nazionale delle Ricerche

Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo"

ISTIday25

PSET: A Phonetics-Semantics Evaluation Testbed

Gianluca Sperduti[1 2], Dong Nguyen[3]

[1]CNR-ISTI [2]University of Pisa [3]Utrecht University

Introduction

PSET is a new testbed for evaluating phonetic embeddings with a focus on phonetics over semantics. It leverages homophones and synonyms to test whether embeddings preserve sound similarity even when semantics differ.

Word quintets: some examples

Overview: Our final testbed includes 250 word quintets.

1. Homophone Selection: Compared multiple online homophone lists and kept only pairs appearing in several sources.

2. Synonym Matching: Used Thesaurus.com to find synonyms for one word in each homophone pair; verified meanings using the Cambridge Dictionary.

3. Filtering: Removed pairs without clear synonyms (e.g., their/there) or with semantic overlap (e.g., spelling variants like archaeology/archeology).

4. Distractor Generation:

Created two types of edit-distance distractors:

- **Phonetic distractors:** IPA-transcribed words at edit-distance 1.
- **Grapheme distractors:** Alphabet-based edit-distance 1 words, manually checked to ensure no synonymy.

Using the Testbed:

For each quintet, we compute **cosine similarities** between the Anchor and: (i) its Homophone, (ii) Synonym, (iii) Phonetic Distractor, and (iv) Grapheme Distractor. A good phonetic model should assign the highest similarity to the Homophone.

Word quintets: some examples

Quintet Components: Anchor (reference word, e.g., bait), Homophone (same pronunciation, different meaning, e.g., bate), Synonym (similar meaning, different pronunciation, e.g., torment), Phonetic Distractor (edit-distance 1 phonetically, e.g., babe), Grapheme Distractor (edit-distance 1 in letters, e.g., ait)

Anchor	Homophone	Synonym	Phonetic Distractor	Graphemic Distractor
cell	sell	cage	bel	bell
cue	queue	hint	clew	due
earn	urn	acquire	an	arn
scene	seen	setting	bean	scent
rest	wrest	vacation	best	cest

Results - Phonetic Embeddings

Articulatory Embeddings
Represent each IPA symbol as a 21-dimensional vector of phonetic features.

Phoneme2Vec
Skip-gram Word2Vec trained on ARPAbet phoneme sequences from the CMU Pronouncing Dictionary (134k words).

XPhoneBERT
IPA-based BERT pretraining for cross-lingual phonetic representation. Static embeddings extracted by averaging the last hidden layer over 10 sentences per word.

BERT / Word2Vec (semantic)
Used as semantic baselines to test contrast with phonetic similarity.

Model	H	S	D_P	Dg
Art. Phonemes *(phon)*	0.748	0.020	0.108	0.120
Phoneme2Vec *(phon)*	**0.903**	**0.000**	0.056	0.036
XPhoneBERT *(phon)*	0.730	**0.000**	0.170	0.090
Word2Vec *(sem)*	0.072	0.744	0.052	0.132
BERT *(sem)*	0.050	0.750	0.040	0.140

Legend

H: Homophones **S:** Synonyms **D_P:** Phonetic distractors **Dg:** Grapheme distractors
Bold = best score for H

Results - LLMs

Prompting Setup Zero-shot prompting with two styles:
Technical (TL): *"Which word is more phonetically similar to [ANCHOR]: [WORD1], [WORD2], [WORD3], or [WORD4]? Only respond with the correct word."*
Layman (LL): As above, but phrased as "sounds more similar". Each style tested twice with shuffled word order → **1000 prompts** per model.

Model	H (mean ± stdev) – TL	H (mean ± stdev) – LL	Error Rate (TL)	Error Rate (LL)
GPT-4o	0.81 ± 0.02	0.78 ± 0.06	0.06	0.09
Gemini 2.5 Flash	**0.87 ± 0.00**	**0.87 ± 0.00**	0.07	0.06
Llama 3.1-8B	0.49 ± 0.12	0.34 ± 0.05	0.17	0.25
Olmo-7B	0.00 ± 0.00	0.00 ± 0.00	1.00	1.00
Olmo-2-7B	0.22 ± 0.04	0.17 ± 0.06	0.21	0.23

Legend

H: Homophone accuracy (higher = better)
Error Rate: proportion of extraction or parsing errors

Github

Summary and Limitations

Key Findings: Phoneme2Vec was the top performer, XPhoneBERT the worst, and Gemini 2.5 Flash achieved the best results among five LLMs.

Limitations: The study is constrained by a small, manually curated English dataset with low-frequency words, some ambiguous synonyms, and potential LLM training exposure.

LLMs for GUI Generation and Requirements Elicitation

Giovanna Broccia*, Maurice H. ter Beek, Alessio Ferrari

FMT

Abstract. The design of graphical user interfaces (GUIs) is a complex and time-consuming process that begins with identifying user roles and gathering requirements through interviews, surveys, or workshops. Designers then create low-fidelity sketches or digital wireframes, organising information into logical sections and selecting visual elements to enhance usability. This iterative process often demands extensive refinement based on stakeholder feedback, making mockup creation—especially for interactive prototypes—a time-consuming task. In particular, the mockup development process often entails spending significant effort on clerical activities, such as programming and debugging tasks, rather than concentrating on human interaction, quick feedback cycles with stakeholders, and creativity. Therefore, it is worth investigating whether large language models (LLMs) can assist GUI designers in streamlining the design process—reducing time and effort while maintaining design quality—enabling them to focus on the human aspects of user interaction by offloading technical programming tasks to the machine. We have done so in the context of the PNRR MOST National Center for Sustainable Mobility, Spoke 4: Rail Transportation, in collaboration with project partner Trenord, a railway transport company from Northern Italy managing a fleet of over 400 trains. We document our experience designing a dashboard for predictive maintenance in railways, illustrating how LLMs can support key tasks such as requirement analysis, information organisation, and mockup generation and refinement. We discuss insights and lessons learned, including the importance of clear requirements, the impact of LLM choice, and the benefits of iterative refinement in achieving stakeholder alignment. Our study shows that LLMs can support the GUI design process by automating specific tasks, thereby reducing design effort and enhancing the overall quality and satisfaction of the final product.

#Prompting, #LLM

Formal Methods and Tools Lab

Consiglio Nazionale delle Ricerche

Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo"

ISTIday25

Large Language Models as Design Partners

Automating GUI Mockups to Refine Requirements

G. Broccia, M. H. ter Beek, A. Ferrari

CONTEXT AND MOTIVATION

The creation of **graphical user interface (GUI) mockups** is a crucial step in the design process, supporting communication between designers, engineers, and end-users. In industrial settings, mockups serve not only to **visualise early design ideas** but also to **clarify and refine requirements** through iterative stakeholder feedback.

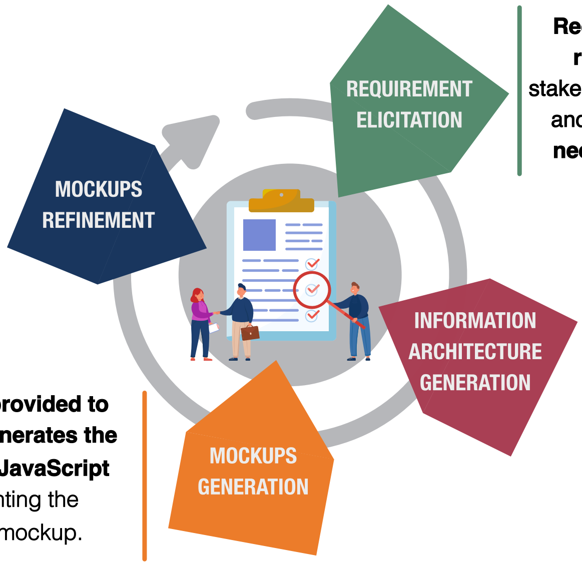
TIME-CONSUMING

Mockups creation and refinement often require **multiple design iterations and frequent stakeholder reviews**. This **makes the process slow and resource-intensive**, delaying validation and stakeholder alignment.

TECHNICALLY DEMANDING

Mockups creation and refinement frequently involve **manual coding and debugging**. These activities **divert attention from higher-level design reasoning and requirement elicitation**.

OUR HUMAN-LLM CO-DESIGN PROCESS



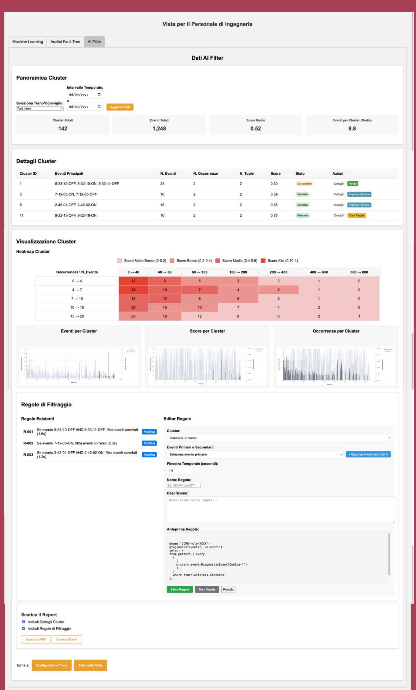
Requirements are elicited and refined in collaboration with stakeholders, using manual sketches and **refined mockups to clarify needs and align expectations**.

The elicited **requirements are provided to the LLM**, which generates the information architecture, **defining the main GUI sections**, their associated content, and the navigation links between them.

Each section is provided to the LLM, which **generates the HTML, CSS, and JavaScript code** implementing the corresponding mockup.

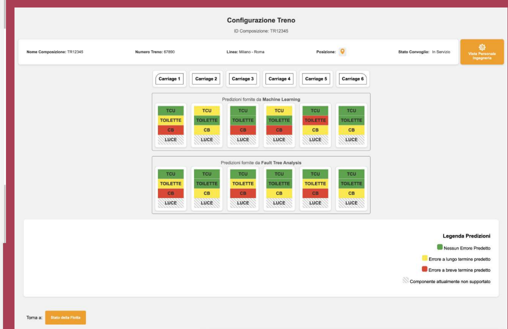
Each **mockup's code, along with stakeholder feedback** and requests, is provided to the LLM, which **refines the code** accordingly.

APPLICATION IN INDUSTRY



In spoke 4 "Rail Transportation" of **PNRR project MOST** (Sustainable Mobility Center) we collaborated with **Trenord**, a railway operator in Northern Italy, to **design a predictive maintenance dashboard** integrated into their existing diagnostic portal.

MOST
CENTRO NAZIONALE PER LA MOBILITÀ SOSTENIBILE



What we learned

- **REQUIREMENTS QUALITY MATTERS**
- **HUMAN IS CENTRAL**
- **LLM CHOICE MATTERS**
- **PROMPTING STYLE MATTERS**

Contact: giovanna.broccia@isti.cnr.it

Scan to read the full paper

Symbolic and Hybrid AI for Brain Tissue Segmentation using Spatial Model Checking

Mieke Massink*, Gina Belmonte, Vincenzo Ciancia

FMT

Abstract. This poster presents an innovative foundational approach for medical image analysis, leveraging spatial model checking techniques invented and developed recently by the FMT-lab@ISTI. The approach is supported by the highly efficient spatial model checking tool VoxLogicA. The tool comes with the ImgQL language, a logic-based image query language, rooted in the theory of topological spaces which can be used to describe imaging-related domain knowledge. The poster also shows the combination of VoxLogicA with the subsymbolic deep learning system nnU-Net. It forms the first-in-its-kind hybrid symbolic-subsymbolic approach to the determination of the Gross Tumour Volume in MRI images of glioblastoma (GBM), a form of malignant intracranial tumour, from the large public 2020 BraTS dataset for Brain Tumour Segmentation. First numeric results, testing and comparing purely symbolic, purely subsymbolic and hybrid approaches on 234 cases of the BraTS 2020 data set will be shown, indicating very promising outcomes for the proposed hybrid setting from the accuracy prespective, but, more importantly, also from the explainability and reliability perspective. The work has just been published in the international journal Artificial Intelligence in Medicine, Elsevier, September 2025, *https://doi.org/10.1016/j.artmed.2025.103154*

#SpatialModelChecking, #BrainTumourSegmentation, #NeuroSymbolicImageAnalysis

ISTIday25

Symbolic and Hybrid AI for Brain Tissue Segmentation using Spatial Model Checking

Gina Belmonte[a], Vincenzo Ciancia[b], Mieke Massink[b]

a) S. C. Fisica Sanitaria Nord, Azienda Toscana Nord Ovest, Lucca, Italy b) Istituto di Scienza e Tecnologie dell'Informazione `A. Faedo", Consiglio Nazionale delle Ricerche, Pisa, Italy

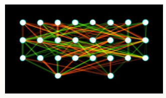
Full article published as Open Access in: Artificial Intelligence In Medicine, Elsevier, Vol. 167, Sept. 2025. DOI: https://doi.org/10.1016/j.artmed.2025.103154

Symbolic segmentation + Neuro-symbolic segmentation

Single threshold estimated by human expert

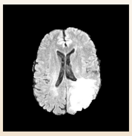
Logic specification with threshold

Per case thresholds estimated using nnU-Net segmentation

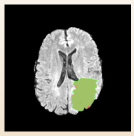


Gross Tumor Volume
Single threshold for dataset.
293 HGG cases BraTS 2020:

Mean Dice 0.82 (std 0.16)
- Naturally explainable
- Human readable segmentation
- Lower accuracy



VoxLogicA
Spatial model checker



Gross Tumor Volume per case thresholds for nnU-Net segmentation
236 HGG cases BraTS 2020:

Mean Dice 0.867 (std 0.087)
- Naturally explainable
- Human readable segmentation
- Higher accuracy

European Union –Next-GenerationEU – National Recovery and Resilience Plan (NRRP) – MISSION 4 COMPONENT 2, INVESTMENT N. 1.1, CALL PRIN 2022 D.D. 104 02-02-2022 – (Stendhal) CUP N. B53D23012850006.

European Union - Next Generation EU, in the context of The National Recovery and Resilience Plan, Investment 1.5 Ecosystems of Innovation, Project Tuscany Health Ecosystem (THE), CUP: CUP B83C22003930001

Symbolic Segmentation of brain tumour with VoxLogicA

ImgQL brain tumour segmentation

ImgQL **Specification 3:** Obtain image and ground truth

```
load imgFLAIR = "Brats17_2013_2_1_flair.nii.gz"
let flair = intensity(imgFLAIR)
load imgGrndTruth = "Brats17_2013_2_1_seg.nii.gz"
let grndTruthGTV = intensity(imgGrndTruth) >. 0
```

Loading images

ImgQL **Specification 4:** Identification of background and brain

```
let background = touch(flair <. 0.1,border)
let brain = !background
```

Identify brain area

ImgQL **Specification 5:** Tumour segmentation method

```
let pflair = percentiles(flair,brain,0)
let hI = pflair >. 0.95
let vI = pflair >. 0.88
let hyperIntense = smoothen(5.0,hI)
let veryIntense = smoothen(2.0,vI)
let growTum = grow(hyperIntense,veryIntense)
let tumSim = similarTo(5,growTum,flair,100)
let tumStatCC = smoothen(2.0,(tumSim >. 0.6))
let gtv= grow(growTum,tumStatCC)
let ctv = distleq(25,gtv) & brain
```

Thresholding
Smoothening
Finding similar tissue
Region growing

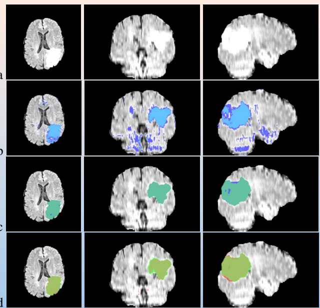


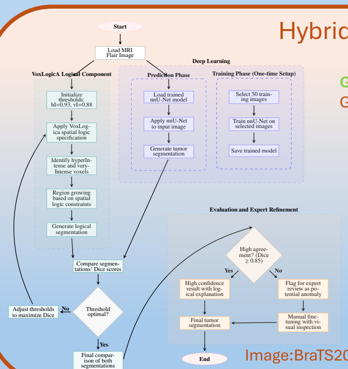
Figure 1: a) Cross section of Brats17_TCIA_335_1 MRI (fLTR: axial, coronal, sagittal view); b) hyperIntense (cyan) and veryIntense (blue); c) growTum (green) and gtv (blue); d) gtv (green) and GTV ground truth (red).

Similarity indexes symbolic segmentation

	Dice (276 cases)			**Sensitivity** (276 cases)			**Specificity** (276 cases)		
	Mean (Stdev)	Range	Median	Mean (Stdev)	Range	Median	Mean (Stdev)	Range	Median
GTV	0.85 (0.11)	0.0-0.97	0.88	0.859 (0.14)	0.0-1.0	0.91	0.999 (0.0)	0.99-1.0	1.0
CTV	0.91 (0.1)	0.0-0.99	0.94	0.937 (0.10)	0.0-1.0	0.97	0.993 (0.01)	0.93-1.0	1.0
	Dice (all 293 cases)			**Sensitivity** (all 293 cases)			**Specificity** (all 293 cases)		
GTV	0.82 (0.16)	0.0-0.97	0.87	0.843 (0.17)	0.0-1.0	0.91	0.998 (0.0)	0.99-1.0	1.0
CTV	0.889 (0.14)	0.0-0.99	0.94	0.923 (0.14)	0.0-1.0	0.97	0.992 (0.01)	0.91-1.0	1.0

Table 3: VoxLogicA evaluation on the HGG cases of the BraTS 2020 training data set (vI=88, hI=95).

Hybrid Neuro-symbolic Segmentation with VoxLogicA and nnU-Net



Green: Ground Truth

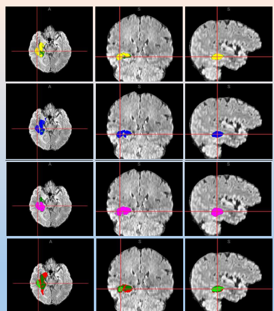


Image:BraTS20_Training_099

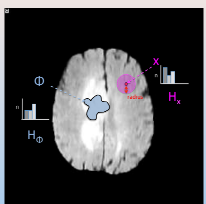
Comparison of similarity indexes neuro-symbolic segmentation

	Dice (234 cases)			**Sensitivity** (234 cases)			**Specificity** (234 cases)		
	Mean (Stdev)	Range	Median (IQR)	Mean (Stdev)	Range	Median (IQR)	Mean (Stdev)	Range	Median (IQR)
GTV-optimal	0.884* (0.075)	0.58-0.97	0.913 (0.07)	0.87* (0.10)	0.43-0.98	0.91 (0.09)	1.00* (0.00)	0.99-1.00	1.00 (0.00)
GTV-nnAsGT	0.867* (0.087)	0.46-0.97	0.895 (0.10)	0.85* (0.122)	0.42-1.00	0.89 (0.13)	1.00* (0.001)	0.99-1.00	1.00 (0.00)
GTV-nnU	0.891* (0.082)	0.42-0.98	0.914 (0.07)	0.92* (0.841)	0.27-1.00	0.94 (0.08)	0.99* (0.00)	0.99-1.00	1.00 (0.00)
GTV-symbolic	0.848 (0.11)	0.0-0.97	0.884 (0.103)	0.88 (0.126)	0.00-1.00	0.92 (0.13)	1.00 (0.00)	0.98-1.00	1.00 (0.00)

GTV-optimal: Optimal thresholds for individual images w.r.t Ground Truth (GT)
GTV-nnAsGT: Optimal thresholds for individual images w.r.t. nnU-Net segm.
GTV-nnU: nnU-Net segmentation (trained on 50 random images)
GTV-symbolic: Symbolic segmentation using the VoxLogicA specification
*: Wilcoxon signed rank statistical test with p-value < 0.01 w.r.t. GTV-symbolic

White and Grey Matter Segmentation with VoxLogicA

Texture Similarity

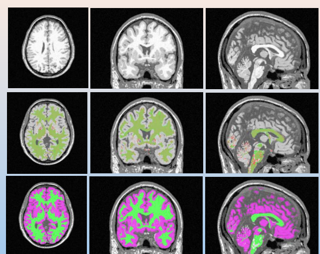


Histogram similarity through crosscorrelation

MRI FLAIR pat4 BrainWeb

Green: white matter
Red: Ground Truth

Green: white matter
Magenta: grey matter



Similarity indexes symbolic segmentation of white and grey matter

	Dice (20 cases)			**Sensitivity** (20 cases)			**Specificity** (20 cases)		
	Mean (Stdev)	Range	Median	Mean (Stdev)	Range	Median	Mean (Stdev)	Range	Median
white	0.93 (0.02)	0.88-0.95	0.93	0.97 (0.02)	0.92-0.99	0,97	0.99 (0.00)	0.98-1.0	0.99
grey	0.91 (0.16)	0.88-0.93	0.92	0.89 (0.03)	0.83-0.94	0.90	0.99 (0.00)	0.99-1.0	0.99

Table 5: VoxLogicA evaluation on the BrainWeb data set of 20 synthetic normal brains.

Symbolic segmentation takes several seconds for an image of 9M voxels

Complexity: VoxLogicA model checking algorithm is linear in the product of the number of tasks and the number of voxels of the image.

VoxLogicA is available as free software from:
https://github.com/vincenzoml/VoxLogicA

Copyright © 2025 Elsevier B.V. All rights are reserved, including those for text and data mining, AI training, and similar technologies according to the CC BY Open Access licence .

How Can an LLM Support Personalised Older Adults Cognitive Training Robot Games?

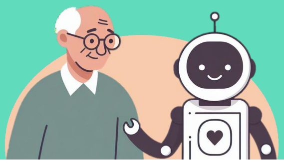
Giulio Canapa*, Benedetta Catricalà, Marco Manca, Fabio Paternò, Carmelina Santoro, Eleonora Zedda

HIIS

Abstract. As populations age, innovative solutions are needed to support cognitive health in older adults. This paper presents a novel platform that delivers personalised cognitive training games through a humanoid robot, leveraging a Large Language Model (LLM) to generate game content and users' profiles from autobiographical memories. Four serious games were designed and developed to stimulate cognitive resources, with content dynamically tailored using user profiles. A four-week study with 20 cognitively healthy older adults and interviews with 10 caregivers assessed system's usability, emotional impact, and perceived value. Results show that users found the robot-delivered games engaging and emotionally meaningful, with significant improvement in perceived user control and high usability ratings. Caregivers appreciated the potential of LLMs for scalable personalisation and provided constructive feedback. This work highlights the feasibility and promises of integrating LLMs and social robots for human-centred cognitive interventions in ageing populations, providing indications on how to achieve it.

#LLM, #CognitiveTraining

HIIS | Consiglio Nazionale delle Ricerche | ISTI Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo" | ISTIday25



How Can an LLM Support Personalised Older Adults Cognitive Training Robot Games?

G. Canapa, B. Catricalà, M. Manca, F. Paternò, C. Santoro, E. Zedda

Goal

Development of a **platform** consisting of a web application for collecting biographical memories and personalised serious games (SGs), delivered through a **humanoid robot** to enhance cognitive and social functions in older adults, with the aim of promoting independent living in an engaging and playful way.

The SERENI Approach

- **Requirement Collection:** psychologists are interviewed to identify effective ways of stimulating older adults' cognitive resources (e.g. attention, memory).
- **Biographical Personalisation:** SG's are **automatically personalised** according to autobiographical memories and life stories.
- **Benefits:** provides more meaningful and motivating experiences, with potential impact on cognitive stimulation and perceived usefulness in daily life.
- **Interactive multimodal HRI:** Playful scenarios mediated through human–robot interaction, using touch, speech, and body animations to enhance engagement.

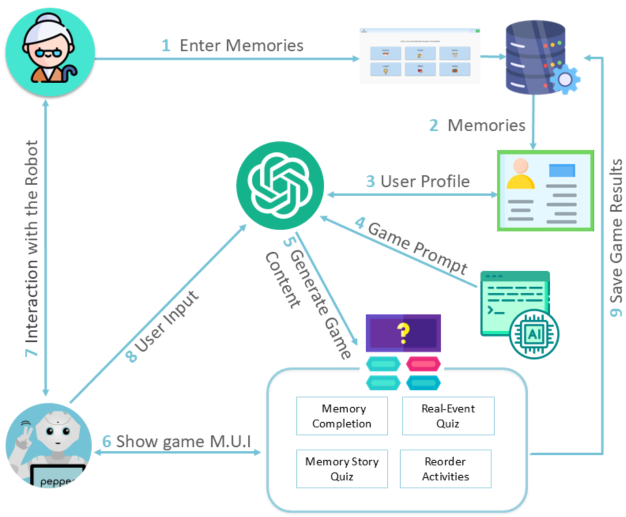
AI-powered Personalisation in SERENI Project

Our system uses an AI-LLM (GPT-4o) to transform users' memories into engaging SG content, adapting challenges and narratives to each individual for a truly personalised experience.

- AI-driven personalisation is useful because it creates **meaningful**, **adaptive**, and **stimulating experiences** that traditional static games and HRI cannot provide.

Here are some games included in the developed system:

- **Memory Story Quiz:** GPT-4o turns user memories into short narratives, adapting length and style to the difficulty levels and users' preferred genres.
- **Real-Event Quiz:** Connects historical events to users' personal memories, promoting temporal orientation and contextual recall through the use of challenging questions and distractors.



Trials in Real Settings

- **April–June 2025:** 12 participants, over 65, with MMSE scores [26-30] in a local clinic.
- **October 2024–January 2025:** 12 users, over 65, participated. Cognitive levels varied, with MMSE scores [18-30]. The trial took place in a nursing home "Residenza Sanitaria Anziani U. Viale - PAIM Social and Care".
- **May–June 2024:** Involved 15 MCI and SCD, over 65, older adults in a local clinic.
- **March–May 2023:** Conducted with 15 MCI users over 65 in a clinical setting.

Results

In summary, older adults responded **positively** to the Pepper memory-based games, with AI-generated, personalised content fostering **emotional engagement**, **autonomy**, and **cognitive stimulation**, while trials highlighted the importance of flexible, **user-centered designs** to address practical challenges such as scheduling and session length.

SERENI

CASPER: Contextual Assistant for Problems Explanation and Resolutions in IoT Automations

Simone Gallo*, Sara Maenza, Andrea Mattioli, Fabio Paternò

HIIS

Abstract. Automations are increasingly present in our everyday environments, enabled by the pervasive presence of connected objects and sensors. However, these automated environments can be opaque to users, particularly when multiple automated behaviours interact with each other, causing conflicts or chain activations. Furthermore, it is difficult to align automations with users' high-level goals. We present CASPER, a system for creating, managing, and explaining automations exploiting integrated visual and conversational interactions. CASPER can identify and explain problems that occur between automations, propose solutions, and help users achieve their goals through personalised suggestions.

#InternetOfThings, #Automations, #PersonalisedSuggestions

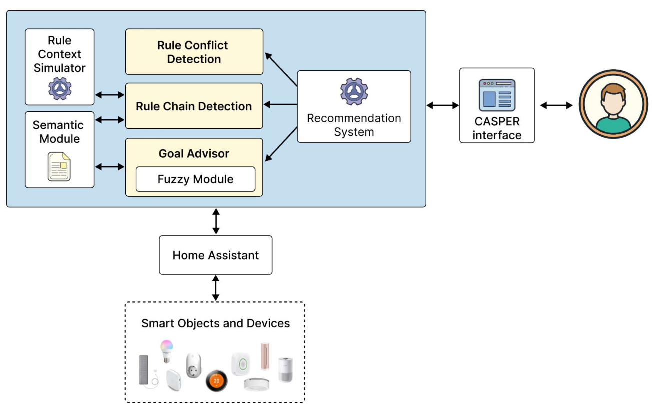
CASPER: Contextual Assistant for Problems Explanation and Resolutions in IoT Automations

Simone Gallo, Sara Maenza, Andrea Mattioli, Fabio Paternò

Motivations

Automations are increasingly present in our environments, enabled by the pervasive presence of connected objects and sensors. However, automated environments can be opaque to users, particularly when multiple automations are active. Furthermore, it is difficult to align automations with users' high-level goals. We present **CASPER**, a system for creating, managing, and explaining automations which can identify and explain problems that occur between them, propose solutions, and help users achieve their goals through personalised suggestions.

Architecture



Rule conflict detection: detects conflicts when two or more rules simultaneously act on the same device in contradictory ways.

Rule chain detection: identify direct and indirect activation chains

Goal advisor: warns whether automations can be in conflict with long term goals

Semantic module: specify the relationships between user goals and contextual variables and between automations' actions and contextual variables.

Fuzzy module: checks whether the current environment situation could activate the related fuzzy check rules

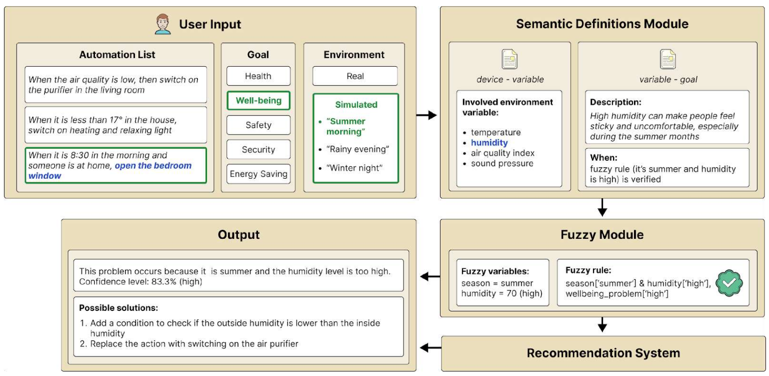
Recommender system: provides possible solutions for the problems in the context of use installation with the support of an LLM.

Goal Advisor

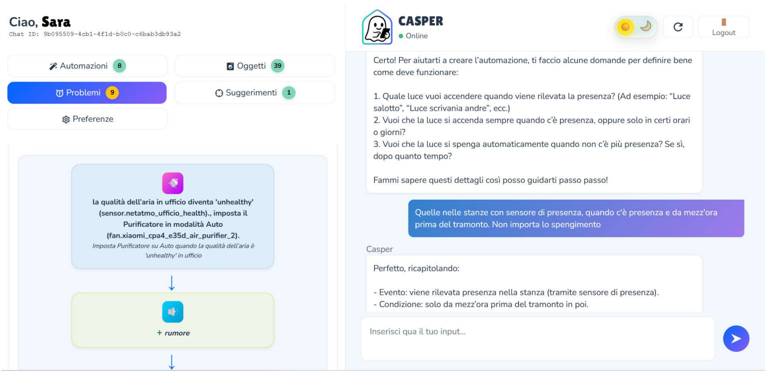
It constantly monitors the actual automations running in the environment. It receives as **input** the selected automations, the goal and data from the sensors in the environment. It provides as **output** the description of the problem, the confidence level (interpreted as its severity), and a list of possible solutions.

It also provides users with an index indicating how well the goals (e.g., well-being, energy saving, comfort) are met. When no specific problem is detected, it suggests improvements to automations based on the user's goal preferences. Preferences are inferred, but users can rearrange them as they wish.

Example Scenario



CASPER User Interface



Casper integrates visual and conversational user interfaces:

The **conversational UI** allows the creation of automations through dialogue. If some issues are detected, it warns the user inviting to explore details with the visual UI.

The **visual UI** provides a graphical representation of the issue, together with a form-based area to select solutions. It also contains the *goal satisfaction index* and the user-goal preferences sections.

PRIN PNRR 2022EUD4GSH project: https://giove.isti.cnr.it/eud4gsh

EUD4XR: Empowering End Users to Shape eXtended Reality Experiences

Marco Manca*, Carmelina Santoro, Ludovica Simeoli

HIIS

Abstract. EXtended Reality (XR) technologies have become increasingly mature and accessible through the availability of different devices and headsets. However, traditional development pipelines still require specialized skills in e.g. modelling 3D content and creating the associated interactions. This complexity hinders non-expert users from developing or adapting XR applications in IoT settings. The EUD4XR project addresses this challenge by introducing an End-User Development (EUD) methodology and platform that empower users to act as "end-users developers". They can participate actively in the definition of interactive XR behaviours through the specification of automations involving smart physical and virtual objects, devices, and services. The platform leverages contextual information and Large Language Models (LLMs) to be able to interpret user intentions (expressed through e.g. voice, gaze, and gestures) and generate appropriate Event-Condition-Action (ECA) rules. Users interact with the system through Tell-XR, a conversational agent supporting the orchestration of the joint behaviour of virtual and real objects in the user's environment, without requiring programming skills. A mixed-reality prototype is currently under development to validate this approach in a training scenario focused on room disinfection procedures. In this context, a sanitization manager wants to create a procedure for training operators to correctly clean a room. Thus, he could tell the system that, to clean a dirty surface, the operator must use a wet rag with a specific degreaser, where the rag is a real object, while both the degreaser and dirt are virtual elements. After generating the corresponding rule, the system updates the visual state of the surface (by changing its texture, to show that its state changes from dirty to clean), only when the user performs the correct actions. To show the effectiveness of the approach, we are planning to evaluate the system soon in a user study.

#ECA, #ConversationalAgent

EUD4XR: Empowering End Users to Shape eXtended Reality Experiences through AI Conversational Agents

Manca M., Santoro C., Simeoli L., Carcangiu A., Mereu J., Spano D.

Main Goal

EUD4XR introduces an End-User Development (EUD) methodology and platform that empower users to actively participate in the definition of interactive XR behaviour (thereby acting as "*end user developers*"), without requiring users' programming skills. To express their automation intentions, users interact (using e.g. voice, gaze, gestures) with Tell-XR, a **multimodal conversational AI assistant**, which, leveraging context information and LLMs, interprets user intentions and generates suitable automations expressed as Event-Condition-Action (ECA) rules, supporting the orchestration of virtual and real objects, devices, and services in user's environment, according to user intent.

Developing a Multimodal Conversational AI Assistant for EUD in VR and AR scenarios

Formative Study

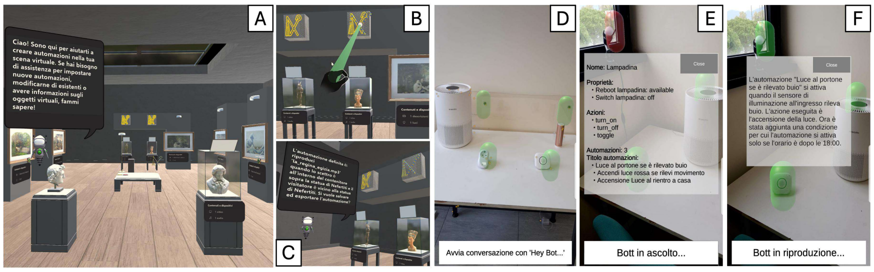
- Wizard of Oz study to understand how users naturally articulate automations through conversational interfaces in immersive environments. Two settings considered: a Virtual Reality museum, and an Augmented Reality smart home;
- *Main results*: VR users prioritized event-first temporal structures while AR users adopted action-first goal-oriented approaches; rule modification challenging in both VR and AR (users often redefined entire automation rules instead of making specific changes).

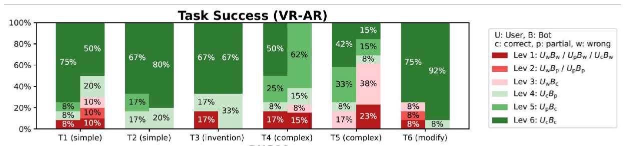
Implementation of the conversational assistant on two platforms (VR virtual museum: Meta Quest 3; AR smart home: smartphone)

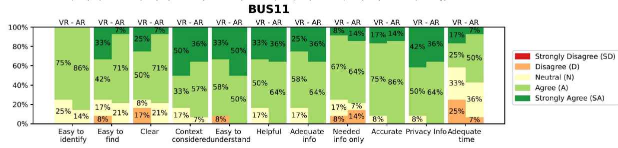
- Tell-XR implemented as a multimodal, multi-agent (define, explore, refine export) chatbot built using LangGraph framework and GPT-4o

Evaluation

- Mixed-method approach (VR: n=12; AR: n=13) combining quantitative metrics (e.g. NASA-TLX, BUS-11, UEQ-S) and qualitative data;
- *Main results:* Tell-XR shows generalisation potential across XR scenarios with good task success rates and positive UX ratings in both AR and VR, with quick user adaptation and low workload, in spite of some inaccuracies/hallucinations in the agent's response; still some difficulties with rule modifications (preferring complete rule restatements), but improved with agent guidance.

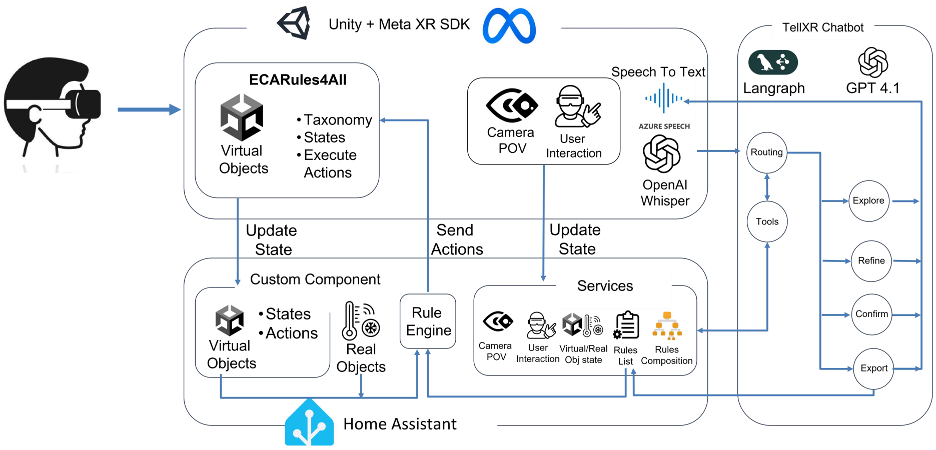


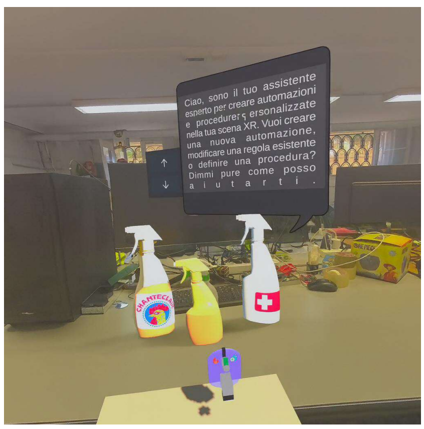




Developing a Working Platform exploiting Tell-XR in a Mixed Reality Prototype

- **Three main components:** i) Unity-based Mixed-Reality application prototype managing virtual objects and user interactions, ii) conversational AI chatbot that interprets the user's intentions and creates ECA automation rules by combining user input with contextual data provided by Home Assistant, iii) a Rule Engine that translates such rules into executable (via HomeAssistant) actions for virtual and real devices;
- Scenario of a hospital room disinfection procedure: a sanitization manager creates a procedure to train operators on correctly cleaning a room
- To create a rule, the manager could tell the system: "*When the surface is dirty, the operator must clean it using the real wet rag and specific degreaser*"**,** where the rag is a real object, while both the degreaser and the dirt are virtual elements. After generating the rule, the system updates the visual state of the surface (by changing its texture, to show that its state changes from dirty to clean), only when the user performs the correct actions.
- Composing different automation rules (via e.g. sequence, interleaving, choice) to support procedures





Next Work: Empirical Validation

- We plan an empirical validation of the **Mixed Reality prototype (Meta Quest 3)** to assess the platform in a working scenario
- **Interested to participate in the evaluation? Contact us: marco.manca@isti.cnr.it**

Self-adaptive and Decentralized Core Maintenance on Temporal Graphs

Davide Rucci*, Emanuele Carlini, Patrizio Dazzi, Hanna Kavalionak, Matteo Mordacchini

HPC

Abstract. Key graph-based problems play a central role in understanding network topology and uncovering patterns of similarity in homogeneous and temporal data. Such patterns can be revealed by analyzing communities formed by nodes, which in turn can be effectively modeled through temporal k-cores. This poster introduces a novel decentralized and incremental algorithm for computing the core decomposition of temporal networks. Decentralized solutions leverage the ability of network nodes to communicate and coordinate locally, addressing complex problems in a scalable, adaptive, and timely manner. By leveraging previously computed coreness values, our approach significantly reduces the activation of nodes and the volume of message exchanges when the network changes over time. This enables scalability with only a minimal trade-off in precision. Experimental evaluations on large real-world networks under varying levels of dynamism demonstrate the efficiency of our solution compared to a state-of-the-art approach, particularly in terms of active nodes, communication overhead, and convergence speed.

#Decentralized, #TemporalGraphs, #Algorithms

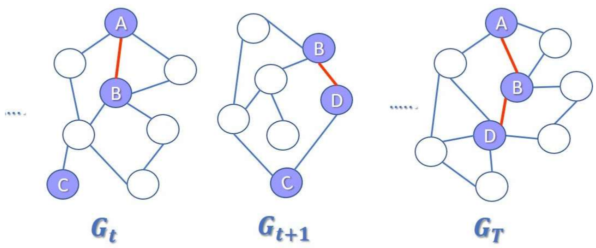
Self-Adaptive and Decentralized Core Maintenance in Temporal Graphs

Davide Rucci*, Emanuele Carlini, Patrizio Dazzi, Hanna Kavalionak, Matteo Mordacchini

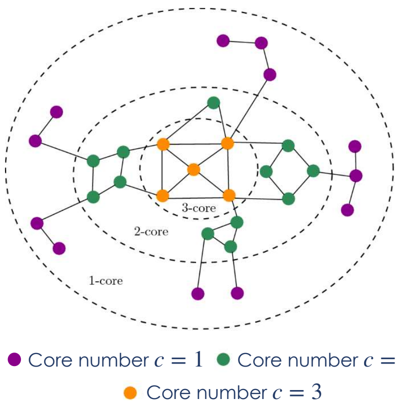
MOTIVATION

Graphs are everywhere, and *community detection* helps in understanding the dynamics and internal organization of networks. Useful for spotting trends, predicting changes or discovering anomalies. We adopt a decentralized strategy to keep track of the changes in communities through time. Implemented in Rust.

TEMPORAL GRAPHS AND K-CORES



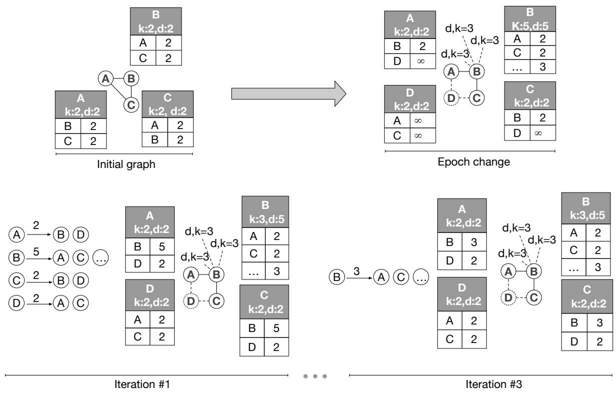
Edges are labelled with *timestamps t*: we can discover trends as they happen. Edges (and neighbors) may appear and disappear during the *lifespan* of the graph. Group edges according to a chosen *aggregation function*.



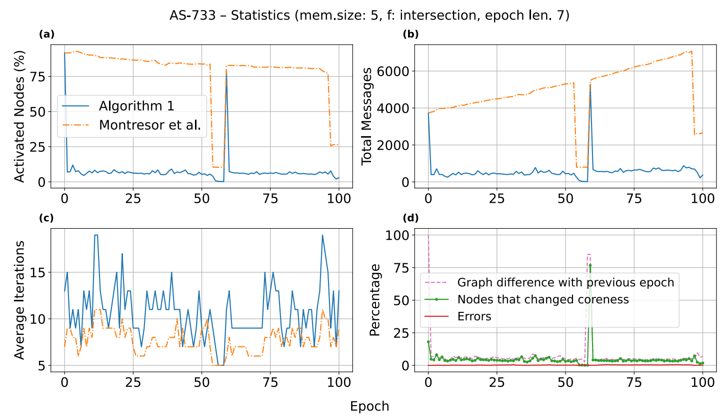
k-core: subset of vertices $C \subseteq V$ such that each vertex in $G[C]$ has degree *at least* k. **Coreness** or core number: highest k such that a vertex belongs to the corresponding k-core. For any given k, the corresponding k-core is unique.

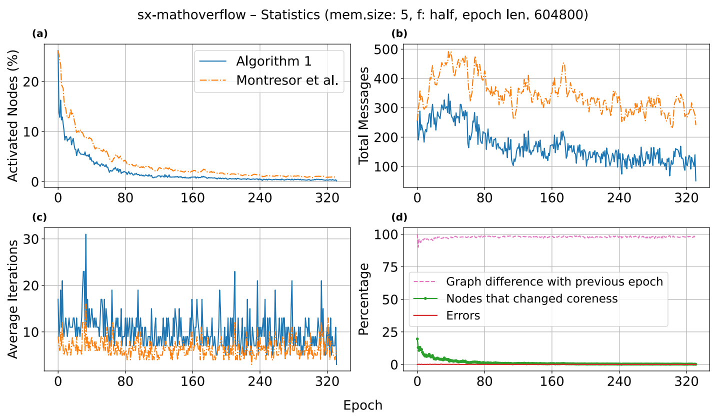
DECENTRALIZED ALGORITHM

Every node sends and receives coreness estimates to and from its neighbors, organized in rounds. Converges when no node changed estimate.



EXPERIMENTAL RESULTS





Rust implementation tested on 7 temporal graphs from different domains. Competitor: temporal adaptation of Montresor et al.'s algorithm [1].

REFERENCES

[1] Montresor, A., De Pellegrini, F., & Miorandi, D. (2011, June). Distributed k-core decomposition. In *Proceedings of the 30th annual ACM SIGACT-SIGOPS symposium on principles of distributed computing* (pp. 207-208).

Efficient Conversational Search via Topical Locality in Dense Retrieval

Cosimo Rulli*, Cristina Ioana Muntean, Franco Maria Nardini, Raffaele Perego, Guido Rocchietti

HPC

Abstract. Pre-trained language models have been widely exploited to learn dense representations of documents and queries for information retrieval. While previous efforts have primarily focused on improving effectiveness and user satisfaction, response time remains a critical bottleneck of conversational search systems. To address this, we exploit the topical locality inherent in conversational queries, i.e., the tendency of queries within a conversation to focus on related topics. By leveraging query embedding similarities, we dynamically restrict the search space to semantically relevant document clusters, reducing computational complexity without compromising retrieval quality. We evaluate our approach on the TREC CAsT, 2019 and 2020 datasets using multiple embedding models and vector indexes, achieving improvements in processing speed of up to 10.3X with little loss in performance (4.3X without any loss). Our results show that the proposed system effectively handles complex, multi-turn queries with high precision and efficiency, offering a practical solution for real-time conversational search.

#InformationRetrieval, #ConversationalSearch, #Efficiency

Cristina Ioana Muntean, Franco Maria Nardini, Raffaele Perego, Guido Rocchietti, Cosimo Rulli

EFFICIENT CONVERSATIONAL SEARCH VIA TOPICAL LOCALITY IN DENSE RETRIEVAL

Conversational systems enhance interaction through dynamic dialogue. While effectiveness is extensively studied, response speed remains crucial. Idea: Exploit topical locality in conversational queries to reduce retrieval time without compromising effectiveness

Methodology

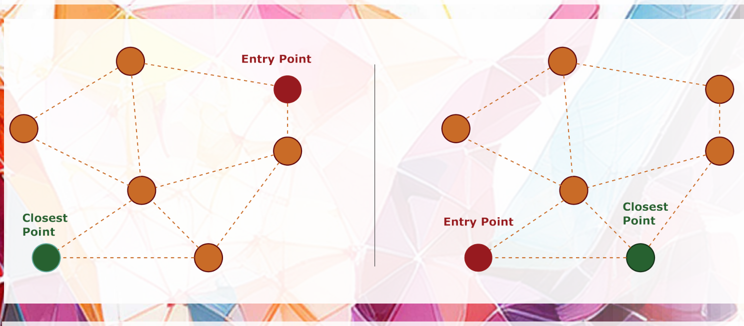
Topical locality. Queries within a conversation focus on related topics. We restrict the search space based on the topic set by the first query.

TopLoc IVF:

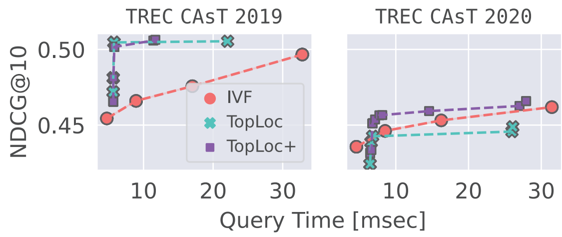
- Caches the set of centroids close to the initial query, and reuses them for the following, topic-related queries.
- Refreshes cache when topic shifts occur, ensuring maintained effectiveness.

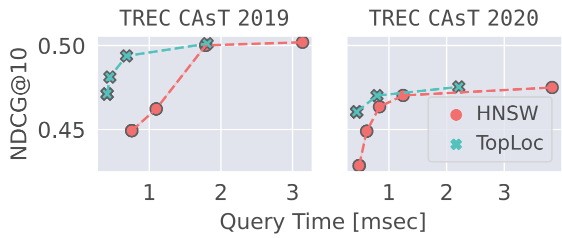
TopLoc HNSW:

- Establishes a conversation-wide privileged graph entry point based on the initial query.
- Reduces costly graph traversal for subsequent conversational turns.



Method	MRR @10	NDCG @3	NDCG @10	Time *msec* [speedup]
Dragon				
Exact	0.799	0.522	0.492	–
IVF	0.813	0.528	0.486	33.0 [–]
TopLoc IVF	0.789	0.517	0.479	6.5 [5.1×]
TopLoc IVF+	0.795	0.518	0.477	3.8 [8.7×]
HNSW	0.789	0.508	0.469	8.3 [–]
TopLoc HNSW	0.785	0.503	0.466	0.8 [10.4×]
SnowFlake				
Exact	0.817	0.550	0.502	–
IVF	0.815	0.544	0.497	24.9 [–]
TopLoc IVF	0.827	0.555	0.505	5.7 [4.4×]
TopLoc IVF+	0.827	0.554	0.501	5.7 [4.4×]
HNSW	0.814	0.548	0.500	1.8 [–]
TopLoc HNSW	0.808	0.549	0.493	0.7 [2.6×]





Experiments

- Datasets: TREC CAsT 2019 & 2020 (38M passages, multi-turn conversations)
- Models:
  - Dragon (BERT, 768-dim)
  - SnowFlake (RoBERTa, 1024-dim)

Effectiveness:

Retrieval performance remains comparable with baselines

Efficiency (Speedups):

IVF index improved by up to 8.7×
HNSW index improved by up to 10.4×

Conclusions

TopLoc significantly accelerates conversational search without accuracy loss.
Future work shall focus on advanced topic shift detection approaches.

SIGIR '25: Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval. https://doi.org/10.1145/3726302.3730186

Graphic Design: beatrice.rapisarda@isti.cnr.it

Auditing for movement fairness

Francesco Lettich, Chiara Renso, Chiara Pugliese*

HPC

Abstract. Fairness in machine learning is a highly active research area addressing concerns that predictive models, due to biases in data or training, can systematically disadvantage certain population subsets. Spatial fairness is a recent variant that focuses on ensuring that models do not penalize or favor individuals in specific geographic areas beyond chance. Auditing the spatial fairness of predictive models involves assessing whether these models systematically penalize (or advantage) individuals associated with a single, fixed geographical location (e.g., place of residence), beyond what would be expected by chance. Current research works on spatial fairness make this basic assumption; however, we argue that many real-world predictive tasks rely on models that take into account individuals' movements, such as trajectories, which are consistently associated with multiple geographical areas. This introduces new computational complexities, as individuals' movements are naturally associated with sets of regions, and not just single fixed locations. To address this, we propose a new notion of fairness, which we call movement fairness, and tackle the novel problem of auditing the movement fairness of predictive models. We argue that movement fairness is relevant in many real-world predictive tasks, e.g., insurance underwriting based on telematics data, gig-economy job allocation, or policing and surveillance guided by past movement patterns: in all these cases, predictive models may penalize (or advantage) systematically individuals whose movements tend to happen in certain sets of regions. Due to their limitations, we argue that existing approaches to spatial fairness auditing are not equipped to capture movement unfairness.

#Fairness, #ML

ISTIday25

Auditing for movement fairness

Francesco Lettich
francesco.lettich@isti.cnr.it

Chiara Renso
chiara.renso@isti.cnr.it

Chiara Pugliese
chiara.pugliese@iit.cnr.it

INTRODUCTION

Spatial fairness refers to ensuring that algorithmic decisions treat individuals equitably across space without systematically favoring or disadvantaging particular geographic areas. Location (e.g., neighborhood, district, or region) acts as a protected attribute to which an individual is tied.

When **auditing for spatial fairness,** predictive models, which take as input (also) such information, are checked to see if they might statistically (unfairly) penalize individuals whose location falls in a given area (e.g., high-crime area).

In this work, we introduce the new notion of **movement fairness**, where an individual's movements are considered instead of a single geographical location. Each trajectory typically spans multiple regions (home, work, gym, supermarket) in a set.

THE MOVEMENT FAIRNESS AUDITING PROBLEM

We want to find out if a predictive model disproportionately penalizes people whose movements are associated with specific sets of geographic regions.

We assume to have access to: (1) individuals movement traces and (2) the model's output label for each person (black-box audit).

We want to find out the subsets of geographical regions in which the associated individuals are **statistically penalized** beyond mere chance: we call this the problem of auditing a predictive model for movement fairness.

SKECTH OF OUR AUDITING APPROACH

1) Segment each individual's movements into **stop** and **move** segments.

2) Discretize space with a (uniform) grid; associate each individual to the subset of grid cells in which their stop segments consistently fall (need to consider and weigh stop **frequency** and **duration** in each cell).

3) Find out the subsets of cells for which the distribution of the labels of the associated individuals diverges from the global label distribution beyond a certain threshold.

4) Run **hypothesis tests** on these subsets, and report those that are statistically penalized by the predictive model (classification) beyond mere chance.

ONGOING WORK

Extensive experimental evaluation on a synthetic but realistic dataset

Add the regression algorithm case to the classification

Better study the impact in real-world sectors: telematics auto-insurance, gig-work assignment, predictive policing,etc

Reasoning in Defeasible Description Logics with System W and Lexicographic Inference

Giovanni Casini*, Jonas Haldimann, Thomas Meyer

InfraScience

Abstract. Description Logics (DLs) are widely applied in AI and database systems. However, like other classical logics, they cannot adequately handle defeasible information. Building on the notion of rational closure—a form of defeasible reasoning originally developed for the propositional setting and later adapted to DLs—we extend this approach by incorporating two further forms of defeasible reasoning: System W and lexicographic closure. Both are well-established entailment relations in the propositional case and are known to satisfy several desirable properties. In this paper, we provide model-theoretic definitions of these extensions for DLs, analyze their behaviour by relating them to their propositional counterparts, and present algorithms for their computation.

#KnowledgeRepresentation&Reasoning, #UncertainReasoning, #Ontologies

REASONING IN DEFEASIBLE DESCRIPTION LOGICS WITH SYSTEM W AND LEXICOGRAPHIC INFERENCE

Giovanni Casini[1,2] , **Jonas Haldimann**[2,3] , **Thomas Meyer**[2]
[1]ISTI-CNR, Italy [2]University of Cape Town and CAIR, South Africa [3]TU Wien, Vienna, Austria

KR 2025 - Int. Conf. on Princ. of KRR
November 11-17, 2025
Melbourne, Australia

Keywords

Knowledge Representation and Reasoning; **Formal Logic**; **Formal Ontologies**; **Uncertain Reasoning**

Motivations & Goals

Motivation:
Description Logics (DLs):

- logical foundation of **formal ontology languages** (**OWL**);
- each language a tradeoff: expressivity vs. computational tractability;
- cannot adequately handle **defeasible reasoning**.

Goal:

- Express **defeasible information** and reason with it.

Example

Description Logic Knowledge Base:

- **TBox**: inclusion axioms $A \sqsubseteq B$ (Every A is a B).
- **ABox**: assertions $a{:}B$ (a is a B) and $(a,b){:}R$ (a and b are in a relation R).

Classic DLs do not admit exceptional cases:

- students do not pay taxes ($St \sqsubseteq \neg T$);
- employed students do ($St \sqcap Em \sqsubseteq T$);
- employed students with children do not ($St \sqcap Em \sqcap Par \sqsubseteq \neg T$).
- **Conclusion:** employed students cannot exist.

Previous Work

DBox: generalizations that admit exceptions $A \sqsubsim B$ (*typically*, A's are B's).

Model appropriate **reasoning systems**:

- **Guiding principles**, like
  Presumption of Typicality: compatibly with the available information, everything behaves as typically as possible.
- **Correspondent consequence relations**, like
  Rational Closure: introduced for Propositional Logic [4], already modeled for Description Logics [1].

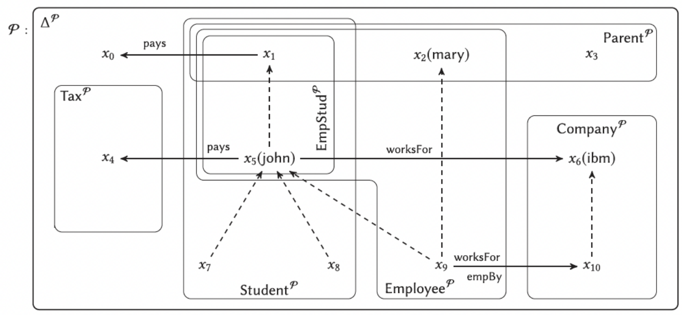
Rational Closure's limit: **inferentially weak**.
Here we consider reasoning **systems extending Rational Closure**, originally formulated for Propositional Logic:

- **Lexicographic Closure** [3] and **System W** [2].

Semantics

Preferential model $\mathcal{P}$: preference relation $\prec$ among individuals.

$a \prec b$ means that a is considered **more typical than** b.



A **defeasible inclusion** $A \sqsubsim B$ is satisfies by a model if the **most typical objects in** A **are also in** B.

$$\mathcal{P} \Vdash A \sqsubsim B \text{ iff } \min_{\prec}(A^{\mathcal{P}}) \subseteq B^{\mathcal{P}}$$

Given a KB, each entailment relation

- **System W**
- **Lexicographic Closure**

is characterised by one specific preferential model.
We define the DL models for System W and Lexicographic Closure.

Reasoning Procedures

Whether an inclusion $A \sqsubsim B$ is a consequence of a KB (according to System W or Lexicographic Closure) is decided implementing **decision procedures** which are

- **correct** and **complete** w.r.t. the respective semantic constructions;
- **easy to implement**: we can build on top of classical DL reasoners.
- **Inferentially stronger** than Rational Closure.

Future Work

- **Implement** the procedures.
- Develop **defeasible ontologies**.
- **Optimize** the procedures **for specific DLs**.
- Elaborate **ABox reasoning**.

References

[1] K. Britz, G. Casini, T. Meyer, K. Moodley, U Sattler, and I. Varzinczak. Principles of KLM-style defeasible description logics. *ACM Trans. on Computational Logic*, 22(1):1:1–1:46, 2021.
[2] C. Komo and C. Beierle. Nonmonotonic reasoning from conditional knowledge bases with system W. *Ann. Math. Artif. Intell.*, 90(1):107–144, 2022.
[3] D. Lehmann. Another perspective on default reasoning. *Annals of Mathematics and Artificial Intelligence*, 15(1):61–82, 1995.
[4] D. Lehmann and M. Magidor. What does a conditional knowledge base entail? *Artificial Intelligence*, 55:1–60, 1992.

Deploying Conversational Agents in Virtual Research Environments: Approaches and Lessons Learned

Massimiliano Assante*, Leonardo Candela, Andrea Dell'Amico, Luca Frosini, Francesco Mangiacrapa, Alfredo Oliviero, Pasquale Pagano, Giancarlo Panichi, Biagio Peccerillo, Marco Procaccini

InfraScience

Abstract. The rapid progress of conversational artificial intelligence and Large Language Models (LLMs) has opened new opportunities to enhance user interaction, support, and accessibility in Virtual Research Environments (VREs). This poster presents the approaches, challenges, and lessons learned from a multi-year effort to design, develop, and deploy conversational agents within the D4Science infrastructure. Through three successive implementation cycles—Janet, D4Science AI Agent, and DAVE—the poster traces a process of iterative refinement aimed at improving flexibility, extensibility, usability, and integration with existing VRE services. Janet, the first prototype, explored modular NLP components but revealed limitations in adaptability and feedback integration. The second approach, based on the Cheshire Cat framework, improved modularity and LLM interoperability but remained constrained by a single-agent design. The latest solution, DAVE (D4Science Assistant for Virtual research Environments), introduces a multi-agent architecture built with Google's Agent Development Kit, enabling secure and context-aware interaction with multiple D4Science services. DAVE combines specialized agents for tasks such as document analysis, catalogue navigation, social interaction summarization, and algorithm deployment within D4Science's computational platform. Integrated feedback mechanisms and a Retrieval-Augmented Generation (RAG) knowledge base further enhance its learning and personalization capabilities. The findings demonstrate that conversational agents can lower barriers to VRE adoption, streamline workflows, and foster user engagement by offering intuitive, natural language interfaces. Lessons learned from this evolution suggest key design principles for future research infrastructure agents, emphasizing modularity, interoperability, and data security. Future work will involve usability evaluations, the integration of user-driven feedback, and experimentation with locally-hosted LLMs to strengthen privacy and operational sustainability.

#LLM, #NLP, #RAG

Deploying Conversational Agents in Virtual Research Environments: Approaches and Lessons Learned

M. Assante[1], L. Candela[1], A. Dell'Amico[1], L. Frosini[1], F. Mangiacrapa[1], A. Oliviero[1], P. Pagano[1], G. Panichi[1], B. Peccerillo[1], M. Procaccini[2]

[1] Consiglio Nazionale delle Ricerche - Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo", Pisa, Italy

[2] Consiglio Nazionale delle Ricerche - Istituto di Geoscienze e Georisorse, Pisa, Italy

The Challenge

The emergence and rapid progress of conversational artificial intelligence and Large Language Models (LLMs) have opened new opportunities to enhance user interaction, support, and accessibility across various digital contexts, including Research Infrastructures, Science Gateways, and Virtual Research Environments.

Embedding conversational agents in an infrastructure like D4Science presents unique challenges including (a) *complexity and diversity*–VREs combine heterogeneous services and data; (b) *flexibility and extensibility*–AI technologies evolve; (c) *user-centricity*–the system must assist users without adding friction and biases; (d) *security and trust*–respect data privacy.

The InfraScience laboratory is actively addressing these challenges by exploring a range of innovative conversational agent designs. Our approach focuses on seamless integration with the D4Science infrastructure, leveraging modular architectures for scalability and adaptability. We prioritize the adoption of the D4Science authentication system across all layers to ensure robust security, while maintaining user-centricity and minimizing potential biases or friction.

The D4Science Journey: from Janet to DAVE

JANET: the modular NLP Prototype (2023)

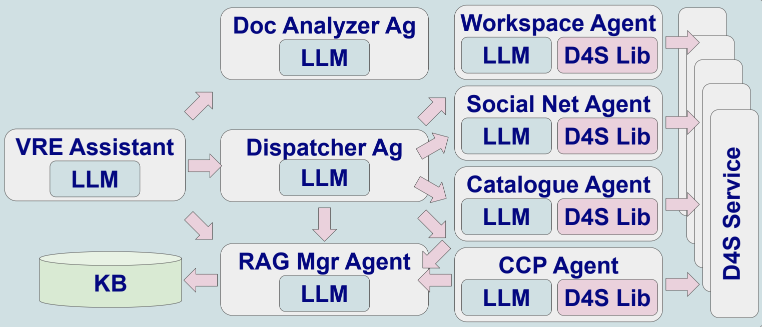
- Designed as a *pipeline of fine-tuned NLP components* (intent classifier, entity extractor, retriever, generator).
- Explored automated support for question answering and content retrieval.
- *Limitations*: Rigid pipeline, limited adaptability, and no real-time feedback integration.

D4Science AI Agent: the Single-Agent Framework (2024)

- Built using the Cheshire Cat framework for modularity and LLM flexibility.
- Integrated with D4Science APIs through Python plugins; supported RAG-based memory.
- *Lessons learned*: Improved flexibility but constrained by a single-agent architecture, with limited support for multi-user and collaborative settings.

DAVE: the Multi-Agent evolution (2025)

- Implemented with Google's Agent Development Kit (ADK).
- Features a central orchestrator and specialized sub-agents (for document analysis, social summarization, catalogue navigation, and algorithm deployment).
- Includes feedback collection, RAG-based knowledge base, and configurable LLM bindings per agent.
- *Outcome*: Enhanced modularity, interoperability, security, and user experience.



WHAT'S NEXT: (a) adopt a co-creation approach by engaging users in refining the agent; (b) assess usability and effectiveness through both quantitative metrics and qualitative feedback from real scenarios; (c) explore the possibility of including a locally-hosted LLM, so to keep all processing on-site and strengthen data protection and confidentiality.

REFERENCES

Ibrahim, A.S.T., Candela, L.: ***Introducing Janet: Early Findings on a Conversational Agent for Virtual Research Environments***. In: Pagè, C., Gesing, S. (eds.) Proceedings of the 16th International Workshop on Science Gateways (2024). doi:10.5281/zenodo.13862920

Oliviero, A., Peccerillo, B., Procaccini, M.: ***An LLM-powered Agent for the D4Science Digital Infrastructure***. Technical Report ISTI-TR-2025/010, Istituto di Scienza e Tecnologie dell'Informazione, Consiglio Nazionale delle Ricerche (2025). doi:10.32079/ISTI-TR-2025/010

L. Candela, D. Castelli and P. Pagano. ***The D4Science Experience on Virtual Research Environment Development***. *Computing in Science & Engineering*, vol. 25, no. 2, pp. 12-19, March-April 2023, doi:10.1109/MCSE.2023.3290433

M. Assante, L. Candela, D. Castelli et al. ***Enacting open science by D4Science.*** Future Generation Computer Systems, Vol 101, 2019, pp. 555-563, doi:10.1016/j.future.2019.05.063

Towards an Infrastructure for Responsible Research Assessment Data Management

Andrea Mannocci*, Leonardo Candela, Paolo Manghi

InfraScience

Abstract. Research evaluation is undergoing a profound transformation, and it is now widely recognised that the true value of a researcher's contribution extends far beyond the sheer volume of papers published in scientific outlets. Yet, despite the growing adoption of revised CV templates and assessment frameworks across many organisations participating in the research ecosystem, a critical gap remains: the lack of structured, interoperable metadata to represent the full spectrum of scholarly contributions. Essential contributions—such as organising conferences, mentoring, teaching, serving on scientific boards, or engaging in collaborative projects—are often undocumented or scattered across ephemeral sources, e.g. emails, web pages, or printouts. Without a robust system for capturing and preserving this information, much of the valuable scholarly record risks being lost as digital content is deleted, websites are updated or decommissioned, or institutional memory fades. To address this challenge, we propose piloting a suite of tools and services that harness the power of Scientific Knowledge Graphs (SKGs), Semantic Web technologies, and Artificial Intelligence. These tools will empower researchers applying for evaluation to capture, persist, and reference their diverse contributions in a CV-ready, machine-readable, and compelling format—on demand and with minimal friction. AI can complement this picture by assisting evaluands in generating narrative sections and impact stories, drafting text and retrieving supporting evidence online. Even more so, by aligning with SKG interoperability standards, this approach will enable the cross-institutional and transnational exchange of evaluation data, paving the way for a more streamlined, verifiable, and up-to-date research assessment process, which will reduce reliance on manual data entry, enhance transparency, and support the principles of Open Science and responsible research evaluation. This research endeavour—posing challenges including dynamic data collection and collation, data provenance and quality, data certification and reliability, generative AI—is not just a technical development; rather, it lays the foundations for a more inclusive, accurate, and future-proof evaluation ecosystem.

#ResearchEvaluation, #Scientometrics, #OpenScience

infrascience | Consiglio Nazionale delle Ricerche | Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo" | ISTIday25

Towards an Infrastructure for Responsible Research Assessment Data Management

A. Mannocci, L. Candela, P. Manghi
Consiglio Nazionale delle Ricerche - Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo", Pisa, Italy

The problem

- Reforms of Research Assessment are on the way
- CVs formats and assessment designs are catching up
- **Structured metadata of diverse contributions to science are lagging behind**
  - Literature, research data and software are usually covered, but...
  - Other kinds of contributions are **mostly undocumented**, or supported by **scattered and ephemeral evidence** (emails, websites, printouts)

What about CNR?

- In 2023: 4 calls for career progression (ex art.15)
- 45 Career Assessment Fields (CAFs)
- 44 different research products and qualifications types
- **Only 11 products (25%) are supported in IRIS**
- **The rest is documented manually**
- **3.864 applicants**
- **Three weeks** (15 full working days, optimistically)
- **~158 years "lost" preparing the application**

One-shot! All the work is **gone at the end!**

The vision

Provide an **infrastructure** to **elevate CVs and applications from mere paperwork**

For assessment designers (e.g., RPOs/RFOs)

- Instantiate an **on-premise complementary system** to support job application
- Provide tools to **easily define metadata** for the assessment items
- **Exchange assessment items** with external RRA infrastructures from other research institutions
- Provide **open access to assessment results**

For evaluators

- **Streamlined evaluation** supported by indicators, diversity of evidence and narratives
- **Born-digital, dematerialised** assessment procedure

For evaluands

- **Document on-the-go** or at application time the untracked contribution to science and narratives
- **Persistency of assessment items** with PIDs
- **Composition of the CV** by importing data from existing sources
- **Reuse, update and adapt** previously deposited assessment items at need
- **Accountability-by-design** with open access to the evaluation outcome and possibility to scrutinise or challenge the results

The research challenges

- **Knowledge engineering and semantic representation** of diverse research contributions
- **Research Information management and interoperability** leveraging open standards (e.g. SKG-IF)
- **Integration with Scholarly Knowledge Graphs (SKGs) and CRIS**
- Development and implementation of **novel metrics and indicators** to support RRA
- **Implementation and alignment with available evaluation frameworks** (e.g., RAF, OSCAM2, NOR-CAM)
- Use of **Artificial Intelligence and LLMs** to
  - Fetch and parse unstructured evidence and map it into ontology-ready knowledge
  - Draft narrative sections and impact stories
  - Reasoning and validation of evaluation consistency

This project has received funding from the European Union's Horizon Europe framework programme under grant agreement No. 101095129. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Executive Agency. Neither the European Union nor

Reconfiguration approach for resilient Systems of Systems

Francesca Lonetti*, DongJae Kang, Felicita Di Giandomenico, Eunkyoung Jee

SEDC

Abstract. Systems of systems (SoS) have become increasingly complex and frequently used in highly distributed, dynamic, and open environments. SoS are complex networks of constituent systems, each capable of operating independently, that contribute towards achieving, when interconnected, some common goals, called missions, surpassing the capabilities of each single constituent system operating in isolation. Since SoS are very dynamic and evolving systems, they must be capable of reconfiguration at runtime, i.e, able to adapt to dynamic changes of the constituent systems, as well as failures or unpredictable misbehaviors of the system. Managing the reconfiguration of a SoS is a very challenging task due to all the potential SoS configurations that could accomplish the functional and non-functional requirements of a pre-defined SoS mission as well as to the evolving properties of the constituent systems. Our main idea is to provide a framework for resilient SoS, i.e., a framework ensuring that the SoS continues to accomplish its mission even in the presence of misbehaviors of CSs, or events that could prevent the correct behavior of the SoS. Our proposal relies on the formal definition of the architectural configuration template. This template allows to express, among others features of the SoS, the list of functional and non-functional requirements of the constituent systems involved in the SoS configuration, the operational mode of the SoS, the violation condition of the SoS configuration, and the priority associated to each SoS. Finally, we aim to assess the effectiveness of the proposed approach on real-world application scenarios.

#SystemsofSystems, #FaultTolerance, #DynamicSystemsReconfiguration

ISTIday25

Reconfiguration approach for resilient Systems of Systems

Francesca Lonetti[1], DongJae Kang[2], Felicita Di Giandomenico[1], Eunkyoung Jee[2]

[1]ISTI-CNR, Pisa [2]School of Computing, KAIST

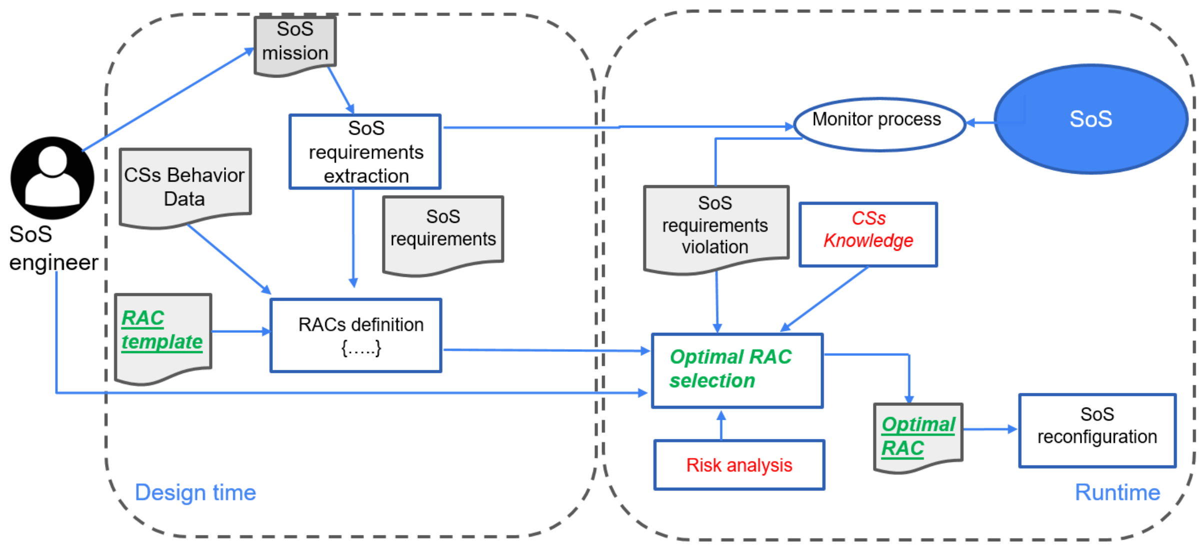
- Systems of Systems (SoS) have become increasingly complex and frequently used in highly distributed, dynamic, and open environments
- Since SoS are very dynamic and evolving systems, they must be capable of reconfiguration at runtime
- Managing the reconfiguration of a SoS is a very challenging task due to all the potential SoS configurations that could accomplish the functional and non-functional requirements of a pre-defined SoS mission

Our Idea

- Provide a formal definition of *Replacing Architectural Configuration (RAC)*
- Leveraging RAC, define a *Reconfiguration framework* for resilient SoS, i.e., a framework ensuring that the SoS continues to accomplish its mission even in the presence of misbehaviors of CSs, or events that could prevent the correct behavior of the SoS

RAC = [{*CS_k*}, *Op*, *V*, *P*]

- {*CS_k*} The complete list of constituent systems in the SoS
- *Op* The operational mode of the SoS [*low, medium, high*]
- *V* The violation condition of the current SoS configuration
- *P* The priority-related information includes various properties and estimated values that assist in reconfiguration decision-making



Reconfiguration framework

Raman Spectroscopy, Signal Processing and Machine Learning in biomedical diagnostics

Gianmarco Lazzini*, Francesco Conti, Mario D'Acunto, Davide Moroni,
Maria Antonietta Pascali

SI

Abstract. Raman spectroscopy (RS) has proven to be an effective and readily engineerable approach for characterizing the molecular composition of complex materials. Its versatility and its ability to collect information without damaging the measured sample make this technique particularly suitable for applications on biological materials. This has spurred interest in RS as a diagnostic tool in the biomedical field. Although the advantages mentioned above are significant, the complexity and high information density of a single Raman spectrum complicate the interpretation of the acquired data. This necessitates the use of accurate and computationally efficient algorithms capable of rapidly processing information, discarding irrelevant and/or redundant features, and integrating useful ones, thereby highlighting molecular differences among biological samples of diverse nature. In this context, combining Signal Processing (SP) with Machine Learning (ML) provides a powerful approach to overcome the inherent challenges of Raman data analysis. Machine Learning pipelines are specifically designed to identify and integrate diagnostically relevant biomolecular variance, highlighting subtle yet significant differences between diverse biological samples. We also explore future perspectives offered by emerging AI paradigms, such as Topological Data Analysis (TDA). By capturing the underlying shape and structure of spectral data, TDA presents a novel method for elucidating complex, non-linear relationships that conventional ML models may overlook. This approach promises the development of more robust and interpretable biomarkers, as suggested in recent literature. The research summarized here, conducted at the SI-LAB of ISTI-CNR, demonstrates the application of these integrated RS and ML approaches to disease diagnosis, proving their potential for creating computationally efficient and highly accurate diagnostic systems across three distinct oncological challenges

---

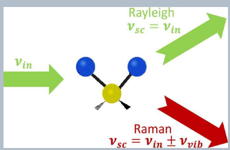
#RamanSpectroscopy, #MachineLearning, #TopologicalDataAnalysis

Raman Spectroscopy, Signal Processing, and Machine Learning in Biomedical Diagnostics

Gianmarco Lazzini[1], Francesco Conti[1,2], Mario D'Acunto[3], Davide Moroni[1], Maria Antonietta Pascali[1]

[1]Institute of Information Science and Technologies (ISTI), National Research Council (CNR), Pisa, Italy
[2]Institut national de recherche en sciences et technologies du numérique (Inria), Université Côte d'Azur, France.
[3]Institute of Biophysics (IBF), National Research Council (CNR), Pisa, Italy

Raman Spectroscopy (RS)

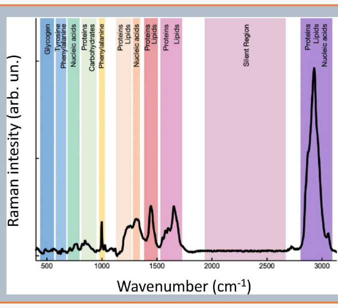


Raman Spectroscopy (RS) measures the so-called **Raman Effect (RE)**, i.e. the variation of photon wavelength ("color") induced by the collision with a molecule.

RS offers a **label-free**, **non-destructive** and **highly engineerable** way to obtain a **detailed molecular characterization** of a complex material.

RE is strongly **molecule-dependent.**

The spectral analysis of Raman photons provides information about the molecular composition of a complex material



Case Study

Diagnosis & Grading of the pancreatic cancer

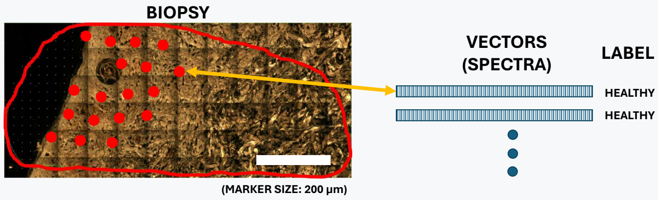
- Biopsies of different nature often exhibit only subtle differences in the Raman signal
- Different portions of the same biopsy often exhibit significant differences in the Raman signal
- The Raman signal is notoriously weak in comparison with concomitant contributions, e.g., autofluorescence

Difficulties in the qualitative interpretation of experimental data

Strategy

SIGNAL PROCESSING (SP)
+
MACHINE LEARNING (ML)

Dataset



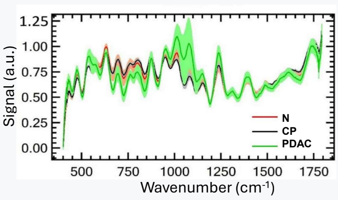
2592 spectra, from 17 biopsies:

4 Normal pancreas (N)
3 Chronic Pancreatitis (CP)
6 Pancreatic Ductal Adenocarcinoma (PDAC) of Grade 2 (PDAC-G2)
4 Pancreatic Ductal Adenocarcinoma (PDAC) of Grade 3 (PDAC-G3)

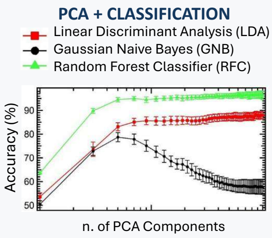
Traditional Pipeline

DATA → PRE-PROCESSING → SUPERVISED LEARNING → DIAGNOSIS

Data preprocessing:
- Baseline subtraction (Improved Multi-polynomial Fitting, degree 3)
- Savitzky-Golay high-frequency smoothing (window: 20, order: 3)
- Minimum subtraction
- Normalization with integral

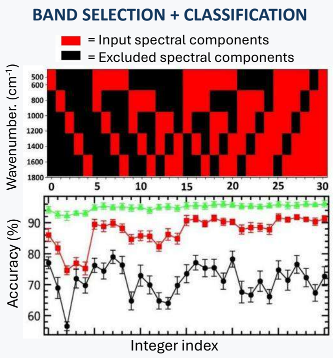


ML task: Three class classification problem (N, CP, PDAC)



Performances assessed through a 5-fold Cross-Validation (not sample-sensitive)

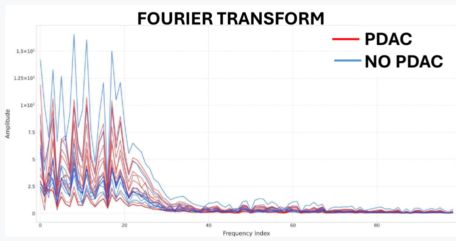
Potential issue: DATA LEAKAGE



Topological Machine Learning Pipeline

DATA → PRE-PROCESSING → PERSISTENT HOMOLOGY → SUPERVISED LEARNING → DIAGNOSIS

Data preprocessing:
- Baseline subtraction (Asymmetrically Reweighted Partial Least Squares)
- Savitzky-Golay high-frequency smoothing (window: 9, order: 2)
- Fourier Transform



ML task: Binary classification (PDAC, NO PDAC), Three-class classification (N, CP, PDAC)

Performances (Leave-One-Patient-Out Cross-Validation)

Method	Three-class	Binary	Classifier
Preprocessing + TML	0.82	0.88	Ridge Classifier
Preprocessing	0.58	0.76	SVM
Preprocessing	0.47	0.65	CNN

[1] Conti, Moroni, Pascali. "A Topological Machine Learning Pipeline for Classification". Mathematics *10* (17), 2022.
[2] Lazzini, Gaeta, et al. "Raman spectroscopy based diagnosis of pancreatic ductal adenocarcinoma". Scientific Reports *15* (1), 2025.
[3] Conti, Lazzini, et al. "Topological machine learning for Raman spectroscopy: perspectives for pancreatic diseases". Proceedings *129* (1), 2025.

Partially carried out at **Bio-ICT Lab**, a joint research initiative by: Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo", IBF Istituto di Biofisica

Unveiling Criticality in Neural Networks Through Temporal Complexity

Marco Cafiso*, Paolo Paradisi

SI

Abstract. Artificial Neural Networks (ANNs) have become a central focus of research in recent times, but their substantial power consumption and lack of interpretability continue to pose a significant obstacle. Looking to biological neural systems for guidance presents a hopeful path toward solving this problem. Spiking Neural Networks (SNNs) constitute a specific category of ANNs that rely on spike-based neurons and function as dynamical systems. SNNs have garnered significant interest due to their promise regarding power efficiency and their exceptional responsiveness to the timing aspects of input data. Over the past 20 years, complex network theory has gained momentum in computational neuroscience, with the brain being the perfect example of a complex system. Recently, principles and methods derived from complex network theory have also been employed in artificial neural networks and machine learning, emphasizing structural connectivity patterns. Nevertheless, temporal dynamics also represent an essential feature exhibited by biological neural systems and can be studied within the context of systems that demonstrate complex intermittent behavior, due to the intrinsic spiking dynamics of biological neural networks. The Intermittency-Driven Complexity (IDC), alternatively referred to as Temporal Complexity (TC), captures these intermittent dynamics by analyzing the metastability of neural states emerging from nonlinear neuronal interactions. IDC is distinguished by power-law decay in inter-event times or by scaling properties within an event-driven diffusion process. Central to understanding neural efficiency is the concept of criticality that describes a system existing in a "critical state" positioned between order and disorder. Originally developed by statistical physicists, criticality has found widespread application across diverse fields. In bio-inspired learning research, criticality has become particularly important since evidence shows that information transmission in the human brain is optimized when operating near critical transitions. The general purpose of this project is to study SNNs models during learning processes, leveraging the TC framework to study the self-organization and complexity of the temporal dynamics, also during critical transition phases. In this poster, we show some interesting results obtained by studying the criticality and complexity of a Modern Hopfield Network model by systematically varying the number of stored patterns, revealing insights into the relationship between network capacity, critical dynamics, and learning efficiency in bio-inspired neural architectures.

#BioInspiredLearning, #ComplexIntermittency, #Criticality

Unveiling Criticality in Neural Networks Through Temporal Complexity

Marco Cafiso[1,2] and Paolo Paradisi[2,3]
[1]Department of Physics 'E. Fermi', University of Pisa, Pisa, Italy
[2]ISTI-CNR, Institute of Information Science and Technologies 'A. Faedo', Pisa, Italy
[3]BCAM-Basque Center for Applied Mathematics, Bilbao, Spain
marco.cafiso@isti.cnr.it and paolo.paradisi@isti.cnr.it

Problem Statement:

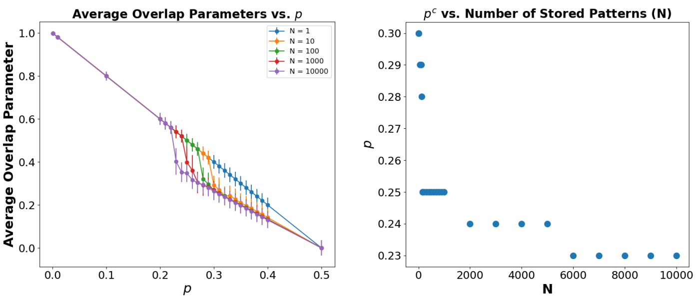
Over the past two decades, complex network theory has become central in computational neuroscience, as the **brain is a prime example of a complex system**. A key concept is **criticality**—the balance between order and disorder. Originating in statistical physics, criticality now spans multiple disciplines and is particularly relevant to **bio-inspired learning**, since the human brain appears to optimize information processing near critical states [1]. Despite various existing methods to detect critical transitions, their identification remains a subject of ongoing debate.

Classical Criticality Assessment:

To assess whether increasing the **number of stored patterns** (**N**) induces a phase transition, we used the **average overlap parameter** as the order parameter [2]:

$$\bar{Q} = \left\langle \frac{1}{L}\sum_{l=1}^{L} S_t(l)\,x_0(l) \right\rangle_t$$

And analyzed its behavior as a function of the **noise level** (***p***).



Model:

We applied analyses derived from the **Temporal Complexity** framework to assess the phase transition in a **Stochastic Modern Hopfield Model** [2] by progressively increasing the number of stored patterns (**N**):

$$\boldsymbol{S_t}[l] = \xi[l]\left(\mathrm{sgn}\left[-\sum_{i=1}^{N}\exp\left(\boldsymbol{x}_i^T\boldsymbol{S}_{t-1}^{(l+)}\right) + \sum_{i=1}^{N}\exp\left(\boldsymbol{x}_i^T\boldsymbol{S}_{t-1}^{(l-)}\right)\right]\right)$$

Where $\xi[t]$ is a binary random process in the time t, taking values in $\{-1,1\}$ with probabilities $P(-1) = p$ and $P(1) = 1 - p$.

Temporal Complexity for Criticality Assessment:

We used the **EDDiS algorithm** [3] to assess the critical phase transition. This method starting from the extraction of sequences of 'Global Events':

- **Coincidence Events**
- **Neural Avalanche Events**

After that an event-driven diffusion model is computed as Continuos Time Random Walk (CTRW) Model on these global events by applying the **Asymmetic Jump** (AJ) walking rule. Finally, we estimate two main diffusion scaling exponent:

1. **Second-moment scaling** (Hurst) H by using the **Detrended Fluctuation Analysis** (DFA):

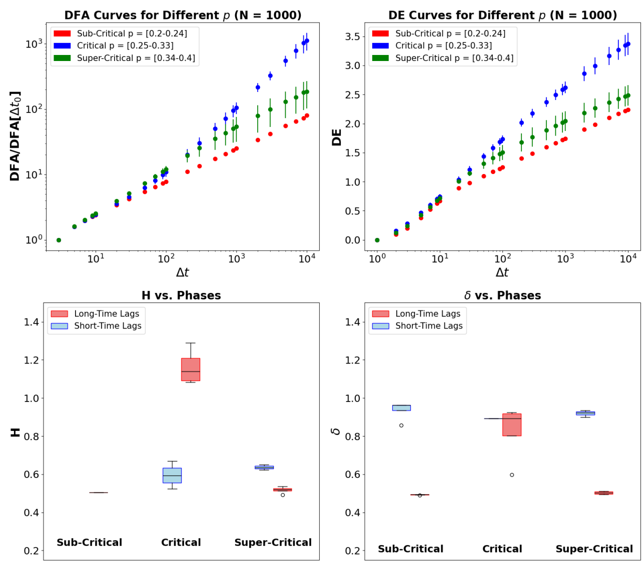
$$F^2(t) = \left\langle \left(X(\Delta t) - \bar{X}(\Delta t)\right)^2 \right\rangle \sim t^{2H}$$

2. **PDF self-similarity index** δ by using the **Diffusion Entropy Analysis** (DE):

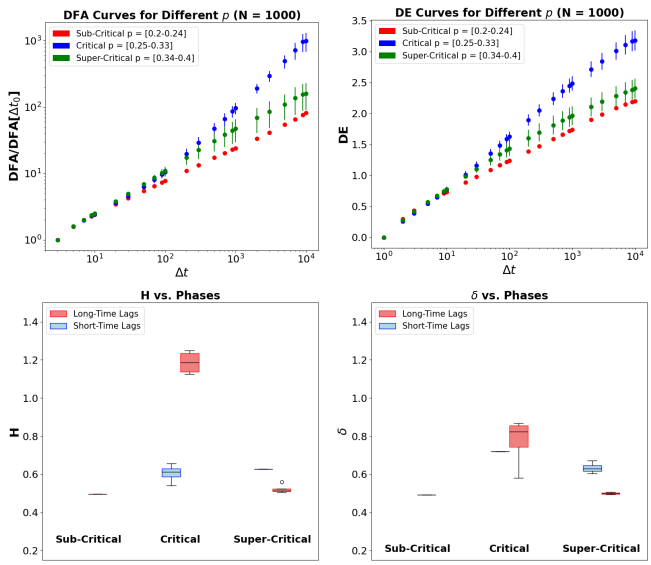
$$P(X,t) = \frac{1}{t^{\delta}} F\left(\frac{X}{t^{\delta}}\right) \rightarrow E(t) \sim \delta \cdot log(t)$$

Results

Coincidences



Avalanches



Conclusions:

Temporal Complexity indexes offer a **statistically robust framework for detecting phase transitions in neural networks**. They enable precise identification of the "**catastrophic forgetting**" point in associative models and quantify **self-organization and complexity**. Our results show that **self-organization rises sharply during critical phases**.

References:

[1] C. Haldeman and J. M. Beggs. "Critical Branching Captures Activity in Living Neural Networks and Maximizes the Number of Metastable States". In: Phys. Rev. Lett. (2005)
[2] M. Cafiso and P. Paradisi. "Criticality of a stochastic modern Hopfield network model with exponential interaction function". In: arXiv (2025).
[3] P. Paradisi and P. Allegrini. "Intermittency-Driven Complexity in Signal Processing". In: Complexity and nonlinearity in cardiovascular signals (2017).

The writing's on the wall

Marco Callieri*, Gaia Pavoni, Massimiliano Corsini, Anna De Falco, Filippo Sala, Quirino Saraceni, Gabriele Gattiglia

VC

Abstract. OPUS is a PRIN 2022 project, led by ISTI-CNR, that combines cutting edge AI tools for the study of historical masonry, both from an archeological and an engineering point of view. Understanding historical masonries in cultural heritage assets is a challenging task that involves different disciplinary fields and the work of experts. The annotation of historical masonry is a process of association between the graphically represented element and any relevant knowledge-based information. This autoptic analysis is vital for supporting the documentation, analysis and diagnosis of heritage structures. The conventional approach is based on the manual drawing of annotated regions over photos or orthographic images of the structure. This makes the process prohibitively time consuming and extremely costly, limiting the ability to manage large-scale monitoring. The OPUS project aims to develop and test an integrated AI-based system for the assisted and automatic annotation of historical masonries. The core of the project is the TagLab tool, an open source tool for the semantic segmentation of ortho-images developed by ISTI-CNR, that enables the researcher to easily and speedily segment survey maps and extract metric and statistical information for their study. The tool, initially developed to work on tasks related to marine biology and ecological mapping, has been expanded with both assisted and automated AI-based tools, to work on historical masonry at different levels of granularity: the mapping of construction features, namely techniques and materials, and the mapping of individual constituent elements.

#SemanticSegmentation, #AIforCH, #HistoricalMasonry

The Writing's on the Wall

Using AI, semantic segmentation and digital tools for the study of historical masonry

Marco Callieri, Gaia Pavoni, Massimiliano Corsini, Anna De Falco, Filippo Sala, Quirino Saraceni, Gabriele Gattiglia

OPUS bando PRIN 2022, ERC SH6_3, codice 202277KJ8F, Decreto Direttoriale n. 104 del 02-02-2022 e n. 1434 del 13-09-2023, finanziato dall'Unione Europea - Next Generation EU, Missione 4, Componente 1, CUP B53D23034170006
Unità: ISTI-CNR (leader), Università di Pisa (Dipartimento di Ingegneria Civile e Industriale, Dipartimento di Civiltà e Forme del Sapere)

Finanziato dall'Unione europea NextGenerationEU — Ministero dell'Università e della Ricerca — Italiadomani Piano Nazionale di Ripresa e Resilienza

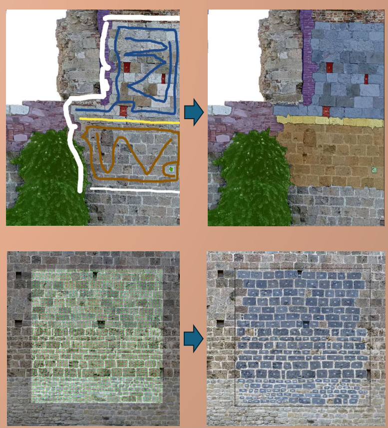
Understanding built heritage is a challenging task that involves different disciplinary fields and the work of experts. The annotation of **historical masonry** is a process of association between the graphically represented element and any relevant knowledge-based information. This autoptic analysis is vital for supporting the documentation, analysis and diagnosis of heritage structures.

The conventional approach is based on the manual drawing of annotated regions over photos or orthographic images of the structure. This makes the process prohibitively time consuming and extremely costly, limiting the ability to manage large-scale monitoring. The **OPUS** project aims to develop and test an integrated AI-based system for the assisted and automatic annotation of historical masonries.

The core of the project is **TagLab**, an opensource tool for the AI-assisted semantic segmentation of ortho-images developed by ISTI-CNR for marine biology and ecological mapping.

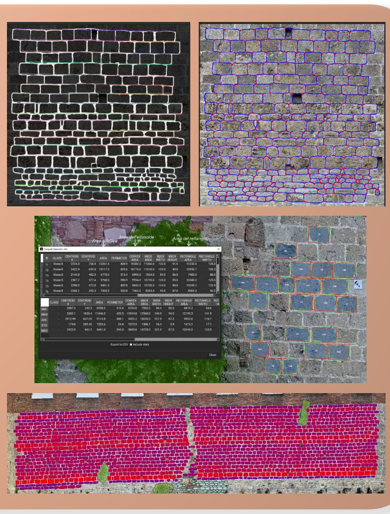
Segmenting

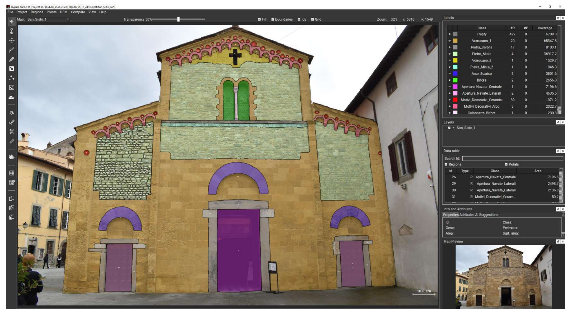
We expanded **Taglab** to cope with the segmentation of historical masonry, working at different levels of granularity: the construction features (e.g., wide areas with uniform techniques / materials), and the individual constituent elements.

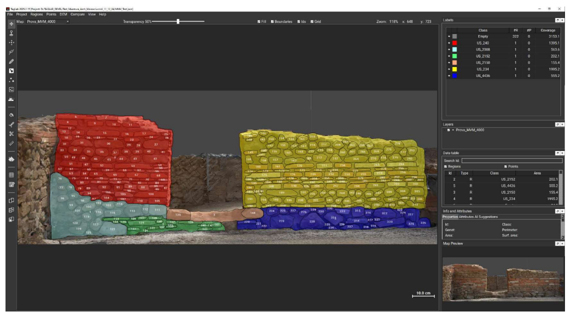


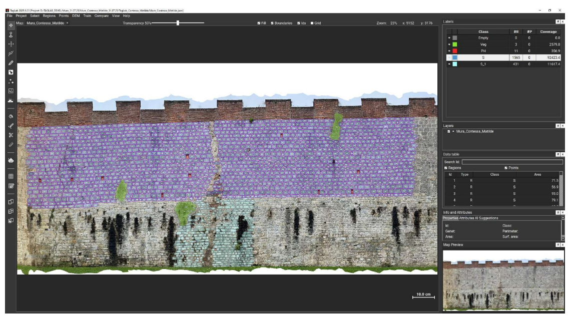
Analyzing

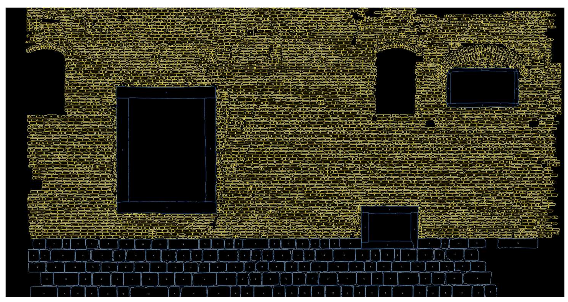
To study and understand the masonry, we addded analysis tools to extract and visualize both metric and structural informations, such as rows arrangement and regularity, staggering, inclination, fitting of primitives, measurement statistics.





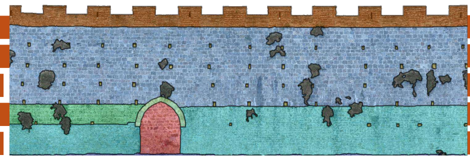






TAGLAB

https://taglab.isti.cnr.it



Acknowledgements

The success of this event and the preparation of these proceedings result from the combined efforts of numerous individuals, who are acknowledged here for their valuable contributions.

Federico Bartalini	*Technical support*
Dimitri Belli	*Advice and recommendations*
Marianna Bolognesi	*Invited Talk*
Antonello Calabrò	*Advice and recommendations*
Diego Ceccarelli	*Invited talk*
Giulio Del Corso	*Chair*
Andrea Esuli	*Invited talk*
Fabrizio Falchi	*Posters selection*
Manuela Ferrante	*Administration*
Danila Germanese	*Posters selection*
Katia Genovali	*Round table moderator, Social media communications*
Daniela Giorgi	*Posters selection*
Giuseppe Lipari	*Technical support, Software implementation*
Oscar Papini	*Proofreading and editing*
Lucia Pavoni	*Administration and Technical support*
Andrea Pedrotti	*Chair*
Beatrice Rapisarda	*Social media communications, Flyer preparation*
Vittorio Romano	*Technical support*
Gian Mario Scanu	*Technical support*
Roberto Scopigno	*Event supervision and organisation*
Federico Volpini	*Technical support, Software implementation*