

ERCIM



NEWS

Special theme:

LARGE LANGUAGE MODELS

Also in this issue's Research and Innovation section:

From Cultural Heritage Assets to Cultural Linked Data
– the Case of the Archaeological Museum of Messara

Editorial Information

ERCIM News is the magazine of ERCIM. Published quarterly, it reports on joint actions of the ERCIM partners, and aims to reflect the contribution made by ERCIM to the European Community in Information Technology and Applied Mathematics. Through short articles and news items, it provides a forum for the exchange of information between the institutes and also with the wider scientific community. This issue has a circulation of about 2,000 printed copies and is also available online, at <https://ercim-news@ercim.eu>.

ERCIM News is published by ERCIM EEIG
BP 93, F-06902 Sophia Antipolis Cedex, France
+33 4 9238 5010, contact@ercim.eu
Director: Dominique Hazaël-Massieux, ISSN 0926-4981

Contributions

Contributions should be submitted to the local editor of your country

Copyright notice

All authors, as identified in each article, retain copyright of their work. ERCIM News is licensed under a Creative Commons Attribution 4.0 International License (CC-BY).

Advertising

For current advertising rates and conditions, see <https://ercim-news.ercim.eu/> or contact peter.kunz@ercim.eu

ERCIM News online edition: <https://ercim-news.ercim.eu/>

Next issue:

January 2024: Large Language Models

Subscription

Subscribe to ERCIM News by sending an email to en-subscriptions@ercim.eu

Editorial Board:

Central editor: Peter Kunz, ERCIM office (peter.kunz@ercim.eu)

Local Editors:

- Ferran Argelaguet, Inria, France (ferran.argelaguet@inria.fr)
- Andras Benczur, SZTAKI, Hungary (benczur@info.ilab.sztaki.hu)
- Cecilia Hyrén, RISE, Sweden (cecilia.hyren@ri.se)
- José Borbinha, Univ. of Technology Lisboa, Portugal (jlb@ist.utl.pt)
- Are Magnus Bruaset, SIMULA, Norway (arem@simula.no)
- Monica Divitini, NTNU, Norway (divitini@ntnu.no)
- Marie-Claire Forgue, ERCIM/W3C (mcf@w3.org)
- Lida Harami, ICS-FORTH, Greece (lida@ics.forth.gr)
- Athanasios Kalogeras, ISI, Greece (kalogeras@isi.gr)
- Georgia Kapitsaki, Univ. of Cyprus, Cyprus (gkapi@cs.ucy.ac.cy)
- Annette Kik, CWI, The Netherlands (Annette.Kik@cwi.nl)
- Hung Son Nguyen, Univ. of Warsaw, Poland (son@mimuw.edu.pl)
- Alexander Nouak, Fraunhofer-Gesellschaft, Germany (alexander.nouak@iuk.fraunhofer.de)
- Laura Panizo, University of Malaga (laurapanizo@uma.es)
- Erwin Schoitsch, AIT, Austria (erwin.schoitsch@ait.ac.at)
- Thomas Tamisier, LIST, Luxembourg (thomas.tamisier@list.lu)
- Maurice ter Beek, CNR-ISTI, Italy (maurice.terbeek@isti.cnr.it)

Cover illustration by DALL-E with the prompt: "Two robots discussing a book while a human looks on."

JOINT ERCIM ACTIONS

- 4 Diversity Matters - INESC TEC's Diversity and Inclusion Commission**
- 6 Report on the 4th Joint JST/ERCIM Workshop** by Katsumi Emura (Fukushima Institute for Research, Education and Innovation) and Dimitris Plexousakis (FORTH-ICS)
- 11 ERCIM "Alain Bensoussan" Fellowship Programme**

SPECIAL THEME

Introduction to the Special Theme

- 12 Large Language Models** by the guest editors Diego Collarana Vargas (Fraunhofer FIT) and Nassos Katsamanis (Athena RC)
- 14 Role-Playing Simulation Games using ChatGPT** by Rita Stampfl (University of Applied Sciences Burgenland), Igor Ivkić (University of Applied Sciences Burgenland and Lancaster University) and Barbara Geyer (University of Applied Sciences Burgenland)
- 15 ChatGPT as a Learning Assistant in Distance Learning** by Michael Prodingler, Rita Stampfl and Marie Deissl-O'Meara (University of Applied Sciences Burgenland)
- 17 Using LLMs in a Language Teaching and Learning Application** by Ioanna Antoniou-Kritikou, Voula Giouli, George Tsoulouhas and Constandina Economou (ATHENA Research Centre)
- 18 Chatbots & Socrates: Dialogues in Learning** by Barbara Geyer, Rita Stampfl and Elisabeth Hauser (University of Applied Sciences Burgenland)
- 20 A Conversational AI Assistant for Teaching and Learning** by Spyridoula Stamouli, George Paraskevopoulos and Nassos Katsamanis (Athena Research Center)
- 22 Using LLMs in Professional Training: Criminal Investigators Improve their Skills in Children's Interrogations** by Don Tuggener (Zurich University of Applied Sciences) and Susanna Niehaus (Lucerne University of Applied Sciences and Arts)
- 23 Unveiling Ethical Biases in Generative AI** by Sergio Morales (Universitat Oberta de Catalunya), Robert Clarisó (Universitat Oberta de Catalunya) and Jordi Cabot (Luxembourg Institute of Science and Technology)

RESEARCH AND INNOVATION

- 25 Exciting Opportunities and Necessary Safeguards for Large Language Models in the Public Sector**
by Peter Biegelbauer, Alexander Schindler, Rodrigo Conde-Jimenez, and Pia Weinlinger (AIT Austrian Institute of Technology)
- 26 What Do LLMs Know about Deception in Text?**
by Katerina Papantoniou, Panagiotis Papadakis and Dimitris Plexousakis (ICS-FORTH)
- 28 Towards AI-Assisted Data Storytelling**
by Angelica Lo Duca (CNR-IIT)
- 30 VOTRACTOR: AI-Powered Vocabulary Design and Keyword Extraction Tool**
by Aradina Chettakattu and Denis Havlik (AIT Austrian Institute of Technology GmbH)
- 32 From Chats to Models: Assessing the Potential of LLMs in the World of Software Modeling**
by Javier Cámara, Javier Troya and Lola Burgueño (ITIS Software / Universidad de Málaga)
- 33 Enhancing Conversational Search with Large Language Models**
by Guido Rocchietti, Cristina Ioana Muntean, Franco Maria Nardini (CNR-ISTI)
- 35 Knowledge Graph Treatments for Hallucinating Large Language Models**
by Diego Collarana, Moritz Busch, and Christoph Lange (Fraunhofer FIT)
- 37 Revolutionising Theatre Archives: Using Large Language Models to Interact with Structured Archival Content**
by Chara Tsoukala, Georgios Paraskevopoulos and Athanasios Katsamanis (Athena Research Center)
- 38 Improving the Sample Efficiency of Pre-training Language Models**
by Gábor Berend (University of Szeged)
- 40 ChatGPT Multilingual Querying Consistency – A Test Case**
by George Tambouratzis (Athena Research Centre)
- 42 A Pipeline for Validating ChatGPT Responses Using Knowledge Graphs and Embeddings**
by Michalis Mountantonakis and Yannis Tzitzikas (FORTH-ICS and University of Crete)
- 44 A Unified Model for Automated Evaluation of Text Generation Systems**
by Jan Deriu and Mark Cieliebak (Zurich University of Applied Sciences)
- 46 Controllable Artificial Intelligence**
by Peter Kieseberg, Simon Tjoa (St. Pölten UAS) and Andreas Holzinger (University of Natural Resources and Life Sciences)
- 48 From Cultural Heritage Assets to Cultural Linked Data – the Case of the Archaeological Museum of Messara**
by Dimitris Angelakis, Lida Charami, Pavlos Fafalios and Christos Georgis (FORTH-ICS)
- 50 Taranis AI: Applying Natural Language Processing for Advanced Open-Source Intelligence Analysis**
by Florian Skopik and Benjamin Akhras (Austrian Institute of Technology)
- 51 Tailoring Decarbonisation and Resilience Strategies to Drive Regional and Local Action**
by Christiane Walter (PIK), Luis Costa (PIK) and Sara Dorato (T6)
- 53 TERMINET: Pioneering the Future of Smart Interconnected IoT**
by Peter Kunz (ERCIM)

ANNOUNCEMENTS

- 54 Dagstuhl Seminars and Perspectives Workshops**
- 54 43rd SAFECOMP 2024, 19th DECSoS Workshop**
- 54 Interdisciplinary Information Management Talks - IDIMT 2024**
- 55 CWI Received Two Test of Time Awards**
- 55 Horizon Europe Project Management**

NEXT ISSUE

ERCIM News 137, April 2024
Special theme: Extended Reality

Call for contributions:
<https://ercim-news.ercim.eu/call>

Diversity Matters - INESC TEC's Diversity and Inclusion Commission

INESC TEC's Diversity and Inclusion Commission (D&IC), led by Ana Sequeira. In this interview, Ana Sequeira highlights D&IC's pioneering initiatives in promoting gender equality, supporting disabilities, and fostering intercultural understanding within the research institute.

What role do you play in your organization?

Ana: The Diversity and Inclusion Commission (D&IC) of INESC TEC was founded in September 2021. The D&IC has five members, although since its creation, only two of the initial team remain. Currently, the D&IC is composed of Ana Sequeira (Coordinator of the D&IC and Senior Researcher), Ana Lopes (Human Resources), Rita Costa (Communication Service), Tiago Gonçalves (PhD student and Research Assistant) and Tiago Silva (Management Control). None of us work exclusively for the D&IC, but we all contribute daily to accomplish the D&IC's objectives. Our team intends to raise awareness towards representation and diversity, and each of us tries to make contributions considering our life experience, training, and the roles we have in the institution. Additionally, the D&IC has autonomy in the initiatives it promotes and has its own annual budget to carry out its functions.

Why do you think it is important to promote inclusion and diversity in research institutes and universities?

Ana: INESC TEC is a private, non-profit association dedicated to scientific research and technological development, technology transfer and advanced consultancy and training. As an institution that operates at the interface between the academic world and the business world, bringing academia, companies, public administration, and society together, INESC TEC enhances the knowledge and results generated within the scope of its research and technology transfer projects, seeking impact through both value creation and social relevance. Literature shows that a diverse institution is more competitive than a non-diverse institution, and this is especially relevant in institutions with high research intensity and an umbilical connection to universities, such as INESC TEC.

As a research centre pursuing excellence, INESC TEC simultaneously welcomes university professors, doctoral and master's students, research fellows, and has many employees working in management and research support services. Moreover, as we live in a globalized world, where scientific research is competitive, INESC TEC is a paradigmatic case of an institution that welcomes both national and non-national researchers: INESC TEC has 1517 active members spread across 43 different nationalities (with most nationalities being Portuguese, Brazilian and Iranian). The only continent not represented by active members of INESC TEC is Oceania, but we can see that our institution is a multicultural space. It is very important, as a research institution that aims for social impact, to provide a safe and inclusive environment for everyone, as it is this safe environment that will

serve as an incubator of true excellence in research and foster good ethical practices in science and towards each other.

Can you briefly explain some initiatives that your organization has started to promote diversity and inclusion? Any initiative that you are particularly proud of?

Ana: In the initial phase, D&IC diligently dedicated itself to the development and implementation of a plan for gender equality. Although it was a legal imperative, it was also a necessity for an institution like ours to have this document approved and published. On the other hand, our focus in the last two years was by no means dedicated solely to gender issues. We put a lot of effort into promoting initiatives related to interculturality and inclusion. Regarding inclusion, specifically for people with disabilities, we have developed several initiatives in this area. We made a diagnosis of how many people with declared disabilities were at INESC TEC, as a legal imperative, but also as a policy of the institution since there is a need to recruit more people with disabilities. This recruitment can only happen if there is awareness of our own fragilities and consciousness towards different disabilities. In this sense, it was decided that INESC TEC must have at least 5% of its workforce made up of people with disabilities and that in the application phase, candidates who have some disability and who decide to declare it will have an increase in the evaluation of their curriculum vitae (CV), with the aim of the institution demonstrating a willingness to welcome more and more people with disabilities.

Also, within the scope of disability, we have developed some guides on good practices for digital accessibility so that the internal and external communications of the institution are increasingly accessible and universal. Diversity and Inclusion for its stakeholders. We had several disability awareness activities, such as the celebration of the International Day of Persons with Disabilities or Louis Braille Day (with a workshop entertained by a member of Portuguese blind association to teach INESC TEC employees what Braille is). We had two Portuguese Sign Language workshops, among other initiatives, dedicated to the theme of inclusion and raising awareness of the community towards disability and people with disabilities and how these people can be important for the future of the institution. We also invested heavily in interculturality. Given that INESC TEC is composed of researchers of multiple nationalities, where the turnover of researchers is enormous, we began to realize that non-national members felt particularly in tune. This cultural inclusion initiative involves sharing information and testimonies about special international dates and cultural holidays, including Portuguese national holidays for foreign employees, in partnership with the International Relations Service (SRI) through email and social networks.

Furthermore, in close collaboration with the communication service, all initiatives developed or co-developed by D&IC were disseminated through social networks and internal communication channels. For example, there was a strong campaign developed by the communications service on the International Day of Women and Girls in Science, which culminated in several posts on social media and the publication of the "Bold" series on INESC TEC's YouTube channel.



“A diverse institution is more competitive than a non-diverse institution, and this is especially relevant in institutions with high research intensity and an umbilical connection to universities, such as INESC TEC. Illustration by DALL-E.”

All those efforts culminated this year with an intercultural celebration in which ten nationalities participated. The main results, visible in employee testimonials, are cultural awareness, reinforced collaboration between colleagues from different backgrounds, and increased cultural exchange. INESC TEC had also made available for all community a new brand prayer room, which gives more privacy and dignity for all users.

Still subordinate to the domain of interculturality and with the sponsorship of SRI, the organization of a month dedicated to interculturality at INESC TEC, which included two training sessions: one on Intercultural Communication and the other on International Negotiation. Whether we're discussing intercultural inclusion or disability inclusion, INESC TEC's D&IC is trying to promote good practices to make this workplace a safe and comfortable ground for everyone.

Have you faced any challenges in promoting inclusion and diversity?

Ana: In general, the creation of the D&IC was well-received by the entire community. We made several public presentations and made sure to raise awareness in the community about the objectives of the D&IC. During our mandates, we have always had support and autonomy from the INESC TEC Board of Directors. The main challenges we face are promoting initiatives consistently and captivating or motivating employees to participate in these initiatives. The awareness towards the D&IC and the contacts made by the institutions' ecosystem has increased and developed very good feedback.

Is there any “mistake” that is important to avoid?

Ana: We have not faced any particular difficulties. In fact, even people who do not fully understand the usefulness of D&IC remain respectful and do not jeopardise our work. The biggest challenge is when we are faced with discussions or comments such as: "this is gender ideology", "men and women are equal, and if they have merit, they reach the same place". We have been receiving all kinds of feedback, and this feedback is crucial on how we develop our initiatives and lines of action. Whether this feedback is aligned with our objectives or not, we try to foster discussions and debates that are constructive and that build a critical approach to each subject. Our short experience as a D&IC does not allow us to advise anyone regarding these topics: we are still in a learning phase. However, there are two aspects that we think can be important and we would like to share.

The first aspect was the launch of a survey to gauge what the community thought about gender, diversity and inclusion issues and to assess their perceptions. The results served as a starting point for our action plan, and we had a more accurate idea of the perception of our community. This survey also revealed which topics the community considered most important to tackle. This was very valuable for us. Another aspect also arising from the result of this survey was about the feeling of belonging, where non-nationals reported feeling less sense of belonging than nationals (for non-national women the results were even lower). To address this challenge, in collaboration with the SRI, we started an email initiative in the following format: SRI sends an email explaining national holidays, and D&IC sends an email about holidays or other non-national ephemerides, but which are very relevant to non-nationals, such as Hanukkah, Ramadan and Eid al-Fitr. The non-national community gave very positive feedback: they told us they felt they were not invisible anymore. For us, one of the highest moments was the celebration of the end of Ramadan (Eid al-Fitr), which was attended by a few dozen collaborators in an event marked by moments of relaxation, learning, a moment of worship and emotion. At the end of the day, the only advice we can give is to be attentive, with a special concern for what is invisible. Just being empathetic to what we see may not be enough.

Please contact:

Tiago André Silva, INESC-TEC, Portugal
tiago.a.silva@inesctec.pt

Andreia Passos, INESC-TEC, Portugal
andreia.passos@inesctec.pt

Report on the 4th Joint JST/ERCIM Workshop

by Katsumi Emura (Fukushima Institute for Research, Education and Innovation) and Dimitris Plexousakis (FORTH-ICS)

This article provides a brief report on the 4th Workshop jointly organized by ERCIM and the Japan Science and Technology Agency (JST). The workshop, themed “Exploring New Research Challenges and Collaborations in Artificial Intelligence, Big Data, Human-Computer Interaction, and the Internet of Things,” took place in October 2023 in Kyoto, Japan. Hosted by JST as part of the Advanced Integrated Intelligence Platform project, the event offered European and Japanese participants an opportunity to report on recent research results in the aforementioned areas and to explore collaboration prospects within the framework of European Commission programs or corresponding initiatives of JST.

The workshop was co-chaired by Dr. Katsumi Emura, Director of the Fukushima Institute for Research, Education and Innovation at JST AIP Network Laboratory, and Prof. Dimitris Plexousakis, Director of the Institute of Computer Science at FORTH. It was attended by 9 researchers from Europe and approximately 20 researchers from Japan. The workshop was structured around four parallel themes, each coordinated and facilitated by one researcher from Europe and one from Japan. The themes were:

1. Trustworthy and Reliable Human-Machine Symbiotic Collaboration,
2. Extracting Actionable Knowledge in the Presence of Uncertainty,
3. Trust in Data-Driven Research: The Role of Actors, Research Infrastructures, and Processes, and
4. Infrastructure and Service Resilience for a Smart Society.

The remainder of this article comprises brief reports on the discussions and outcomes of the respective groups.

Following opening remarks by the co-chairs, the first day’s keynote address was delivered by Prof. Hiroshi Ishiguro of Osaka University on “State-of-the-Art Human-Robot Interaction Research,” highlighting exciting achievements in interaction with robotic avatars and their advanced features. The second day featured a keynote address by Dr. Hiroko Tatesawa, Director of JST’s Department of Strategic Basic

Research on Activities for Promoting Basic Research and International Collaboration. Beyond the plenary sessions, the four groups convened in parallel sessions, comprising presentations by members and lively discussions on challenges and topics for joint research.

On the last day of the workshop, each of the four groups reported on the outcomes of their discussions with brief presentations on specific research questions and topics for further research. The workshop’s steering committee convened in a closing session to assess the outcomes and discuss plans for the continuation of this series of workshops. The 5th edition of the Joint Workshop will be hosted by ERCIM member SZTAKI (Hungarian Institute for Computer Science and Control) in October 2024 in Budapest, Hungary, as an event collocated with the ERCIM Fall meetings.

Theme 1: Steps towards Trustworthy and Reliable Human-Machine Symbiotic Collaboration

Contributors: Giorgos Flouris, Satoshi Nakamura, Wen Gu, Rafik Hadfi, Loizos Michael, Theodore Patkos, Seitaro Shinagawa, Katsuhito Sudoh, Hiroki Tanaka, Koichiro Yoshino

Robotics is experiencing a paradigm shift from heavy-duty industrial robots towards intelligent agents operating in our homes, interacting with people without technical skills, in environments that are open, complex, and unpredictable. To cope with such environments, AI must address a diverse set of challenges and acquire new skills.

Europe, with its strong capacity in research and academia and a well-established ecosystem of scientists, developers, suppliers, and end-users, has already set these challenges as strategic, short-to-mid-term objectives. This is evidenced by the recently announced Strategic Research Innovation and Deployment Agenda of the ADR Association [L1], and the Statement on the Future of AI in Europe published by the CLAIRE organization [L2]. These objectives are coupled with a clear desire to preserve certain ethical values, such as trustworthiness and transparency, often enforced through regulatory frameworks like the forthcoming AI Act. Similarly, Japan is pioneering the adoption of cutting-edge AI and



Figure 1: Participants of the 4th Workshop jointly organized by ERCIM and the Japan Science and Technology Agency.

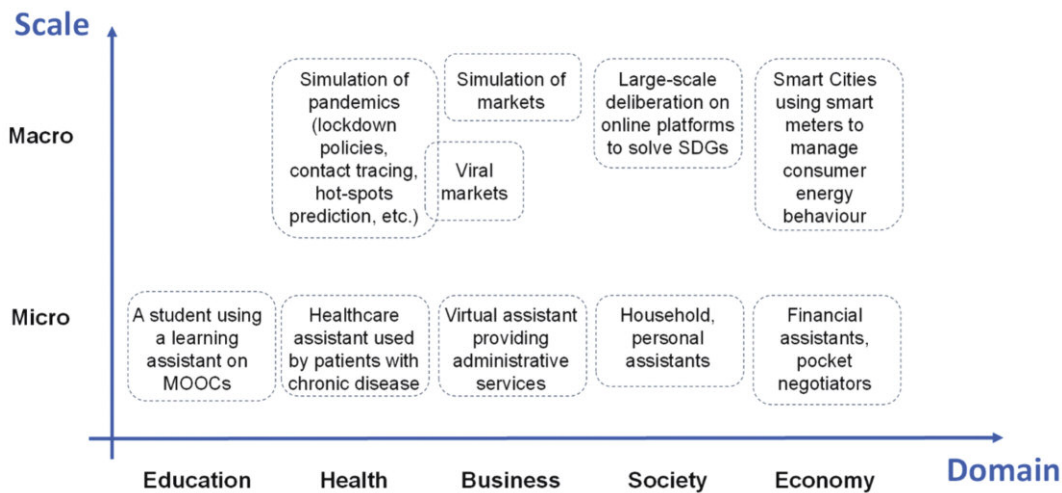


Figure 2: An organisation of the relevant use cases along 5 different domains and for different scale (micro or macro).

Robotics research results in the industry, thereby boosting innovation and maximizing economic and societal impact.

Research collaboration opportunities between Europe and Japan in human-machine symbiotic collaboration are creating a fruitful ground for yielding positive results for both regions, if structured on top of a well-defined plan. This was the objective of Group 1 at the 4th JST/ERCIM Joint Workshop. Human-machine collaboration can be achieved across various domains, including education, healthcare, business, society, and economics, through the appropriate implementation and deployment of smart assistants. Several examples are illustrated in Figure 2, across different scales and domains.

To promote trustworthiness and reliability in these domains, machines should be able to explain their actions and the reasons behind them. Contestability, the ability to propose methods, trial and error, and collaboration on par with humans, is also important. Indicators and criteria for quantifying aspects of explainability and contestability are necessary, respecting fairness, bias, and accountability, depending on their purpose.

Socially aware communication and interaction encompass diverse aspects, including cognitive, linguistic, cultural, and social differences that shape human connections. Understanding both verbal and non-verbal communication, such as intentions and emotions, is crucial.

In other words, such systems require cognitive skills that highlight the necessity for reasoning abilities aligned with human thought, particularly in exhibiting a “Theory of Mind,” i.e., understanding others' beliefs, intents, desires, and emotions. These reasoning abilities include the use of common sense and causality to predict human actions in various contexts, especially where implicit knowledge is involved. They also encompass the use of argumentative reasoning for rational discourse and decision-making, along with social cognition to navigate social interactions and norms. Together, these skills are vital for AI systems to integrate effectively into human settings, and are expected to require the adoption of insights from Cognitive Science, Psychology, and the social sciences, despite the anticipated interdisciplinary challenges.

Furthermore, a symbiotic environment highlights the need to integrate contextual information. This environment thrives on the collaboration between humans and artificial agents, augmenting human capabilities and emphasizing the necessity for adaptability. Social and cultural awareness play pivotal roles, necessitating adherence to regulations and social norms to avoid misunderstanding. Establishing common ground through defining, aligning, and using new concepts is pivotal, fostering shared understanding and facilitating smooth interactions across diverse cultural and social contexts, aligned with human societal values.

Human-centric adaptive learning via automatic feature selection with proper learning strategies is also a promising direction to consider. Desired features in these learning strategies should include cognitive development, the ability to learn online, and the ability to make actionable predictions.

Foundation Models and Large Language Models (LLMs) significantly enhance human-machine interaction but face limitations in logical reasoning. Integrating neuro-symbolic methods can address these limitations by enriching the symbolic reasoning capabilities within LLMs.

Multi-agent-based simulations, coupled with agent communication, will fine-tune perceptions and enhance cognitive abilities, facilitating real-world error correction and knowledge emergence.

Overall, the synergy among these technologies is paving the way for more general AI. Each technology offers unique strengths, pushing towards more versatile and efficient systems. This collaborative evolution highlights a trend towards AI that is more context-aware and autonomous, progressively integrating into a unified AI ecosystem.

To promote collaboration between Europe and Japan on the above topics, we have planned activities that take into account the cultural differences and backgrounds, as well as the characteristics of the two research ecosystems, starting with a SWOT analysis (Strengths, Weaknesses, Opportunities, Threats). One aspect of complementarity could be neuro-symbolic AI in the EU and deep learning AI and humanoid robots in Japan. More applications to EU-Japan collaboration calls will be encouraged, as well as tools

to improve networking, such as postdoctoral fellowships, professorial sabbaticals, researcher exchanges, or student exchanges/internships. Joint academic ventures such as joint workshops, joint editorship of journal special issues, and joint online degrees (open university style) would also be beneficial in promoting collaboration.

Links:

[1] <https://adr-association.eu/>

[2] <https://claire-ai.org/>

Please contact:

Giorgos Flouris, FORTH-ICS, Greece fgeo@ics.forth.gr
Satoshi Nakamura, NAIST, Japan, s-nakamura@is.naist.jp

Theme 2: Extracting Actionable Knowledge in the Presence of Uncertainty

Contributors: D. Laurent, N. Spyrtatos, Y. Tanaka, H. Tokuda, A. Uchiyama, K. Yamamoto and T. Yoshihiro

The main objective of this group was to study the concept of uncertainty and its impact on actionable knowledge extraction. Our approach considers uncertainty in the context of data pipelines, whose importance and complexity are increasingly significant nowadays. Good service provision by a data pipeline largely depends on data quality, with uncertainty being one of the factors influencing it. Therefore, we decided to pursue two lines of research: (a) identifying the factors that influence data quality in some relevant research areas; (b) attempting to define data quality in each of these areas, potentially revealing commonalities. The following is a brief report of our first meeting's findings.

Databases

One of the most frequent data quality problems in databases is missing values. Incomplete datasets can disrupt data pipelines and have a devastating impact on downstream results if not detected. Various approaches to impute missing values exist, but comprehensive benchmarks comparing classical and modern imputation approaches under fair and realistic conditions are scarce [L1]. Another common data quality problem is the violation of semantic constraints, such as key constraints. A data table violating these constraints can disrupt data pipelines and cause significant damage in critical applications. Designing algorithms to "repair" such tables before use is a current challenge in the field of databases [L2]. An approach to repairing tables with missing values and violating semantic constraints was presented at the meeting [L3].

Signal Processing

Restoration of observed values, or signal values, is a central part of applications in signal processing. Restoration typically refers to denoising, interpolation, and deblurring. Most restoration problems can be formulated as inverse problems with appropriate regularization terms. Extracting features and discovering knowledge from observations requires more than just signal processing; the pipeline includes sensing,

signal processing, database creation, and machine learning. To improve data quality for downstream tasks, integrating sensing, signal processing, and databases should be considered. While a similar approach, known as ISAC (integrated sensing and communications), has been considered in telecommunications [1], its concept needs to be extended to broader applications.

Data Quality and Trust

Data is a spatio-temporal reflection of the real world. In this sense, data has a scope, and reality within this scope can be projected into the data. The scope can be objectively described and defined as a data profile. Data quality expresses how accurately and precisely reality within the frame was projected, representing the uncertainty of reality mapping. Data quality is crucial because it affects the performance of the data processing pipeline [2]. However, defining it is challenging due to numerous metrics, including accuracy and missingness. People will trust data if its quality ensures the data pipeline's performance. A definition of data quality and a systematic quality assurance methodology are necessary to establish trust.

Reliability of Labels

Label reliability is crucial for AI model development [3]. Definitions of activities may differ by context, and privacy concerns often require subjects to self-label activities, potentially leading to inaccuracies that compromise AI model reliability. To mitigate this, large data volumes are collected under the assumption that most labels are correct, although this approach is labor-intensive. Clarifying labeling instructions can reduce ambiguity, and employing multiple annotators is another strategy. This approach is commonly used in computer vision due to the ease of distributing image and video data. However, privacy-sensitive data, such as in-home activity recognition, cannot be easily multi-annotated, presenting a persistent challenge that requires further research to improve annotation accuracy.

Links:

[1] <https://kwz.me/hzt>

[2] <https://marceloarenas.cl/publications/pods99.pdf>

[3] <https://kwz.me/hzd>

References:

- [1] F. Liu, et al., "Integrated sensing and communications: Toward dual-functional wireless networks for 6G and beyond", *IEEE journal on selected areas in communications*. 2022;40(6):1728-67.
- [2] J. Byabazaire, et al., "Using Trust as a Measure to Derive Data Quality in Data Shared IoT Deployments" *ICCCN2020*, 2020, pp. 1-9, doi: 10.1109/ICCCN49398.2020.9209633.
- [3] A. Olmin and F. Lindsten, "Robustness and Reliability When Training With Noisy Labels", in *Proc. of the AISTATS 2022*, vol. 151, pp. 922–942.

Please contact:

Nicolas Spyrtatos, Universite Paris Saclay and FORTH-ICS
spyrtatos@lri.fr

Yuichi Tanaka, University of Osaka
ytanaka@comm.eng.osaka-u.ac.jp

Theme 3: Trust in Data-Driven Research

Contributors: Andreas Rauber, Satoshi Oyama, Hisashi Kashima, Naoto Yanai, Ji Yi Li, Koh Takeuchi, Akiko Aizawa, Dimitris Plexousakis, Katharina Flicker

Research in virtually all disciplines is increasingly reliant on data collected from a variety of sources, pre-processed, and analyzed through complex processes by numerous stakeholders. The complexity of processing pipelines, the multitude of actors and steps involved, and the extensive reuse of code from diverse sources pose numerous challenges regarding the trust we can place in the correctness of the processing applied. A critical question arises: how much trust can we place in our own research outputs (and those we reuse from others) when we cannot verify every single data instance, code, and processing step applied?

This leads to questions about what constitutes trust in data and code quality, the processing applied, the differences between scientific disciplines and cultures, the necessary understanding of trust to make it measurable, and the activities required to build and maintain trust long-term. Additionally, it's important to consider how research infrastructures meeting specific trustworthiness criteria can assist in this process. Answers to these questions are essential for determining the trustworthiness of research outcomes and, consequently, the degree of accountability we can accept for our findings, i.e., the quality of our research.

Elements and Facets Influencing Trustworthiness

We explore trust through a simplified research process (Figure 3) consisting of defining research questions, which drive a loop of data and code feeding into processing that produces an output, ultimately leading to insights or knowledge. This process is embedded in the contexts of principles and expectations of Communities of Practice, human stakeholders, the organizations they are embedded in, and the technical infrastructures within which the research is conducted.

For each of these elements, an extensive (though not exhaustive) set of attributes or facets influencing perceptions of trust can be identified. "Quality" is interpreted as "fitness for purpose" in a

given setting. This implies that there are no absolute indicators of quality that justify trustworthiness for any element in this setting. Quantification of trustworthiness needs to be contextual to the activity being performed. Figures 4 and 5 provide examples of attributes and facets identified for the elements Data and Code, respectively, as well as their connections to their context.

Planned Activities

The initial "trust diagram" outlined in Figure 3 will be expanded through user studies, interviews, and questionnaires

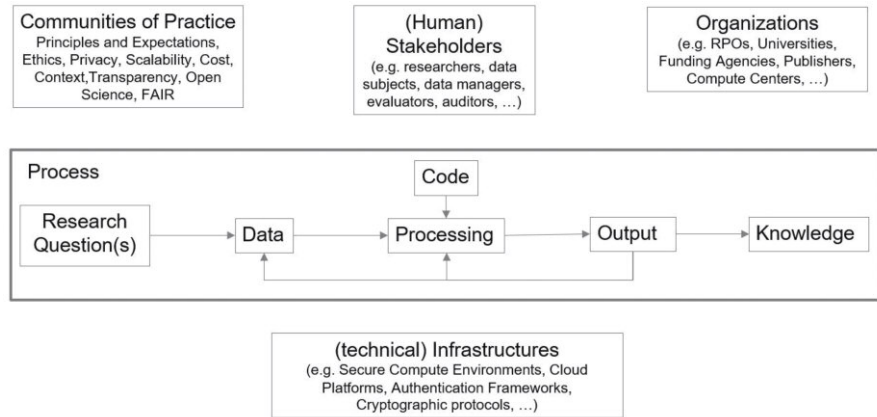


Figure 3: (Selected) Elements participating in the Trust Network.

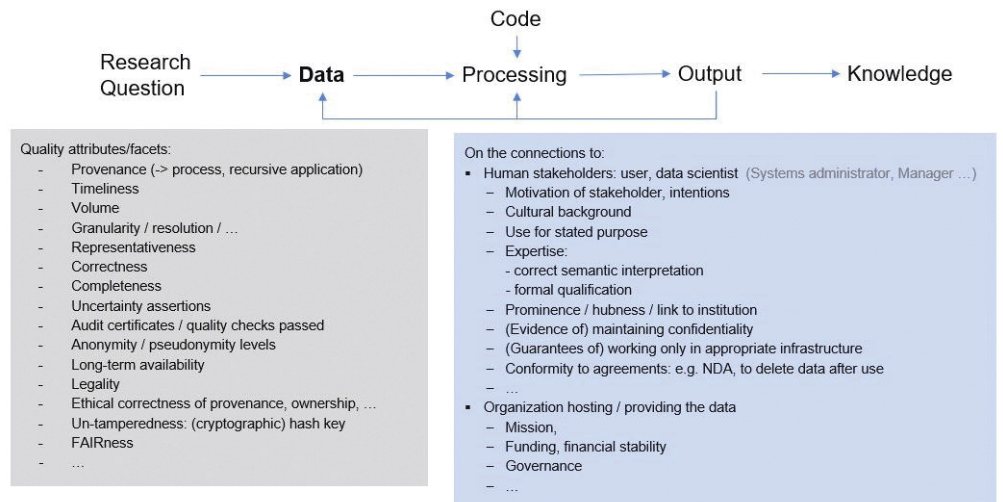


Figure 4: (Selected) quality attributes / facets influencing trustworthiness on the Data element and its connections.

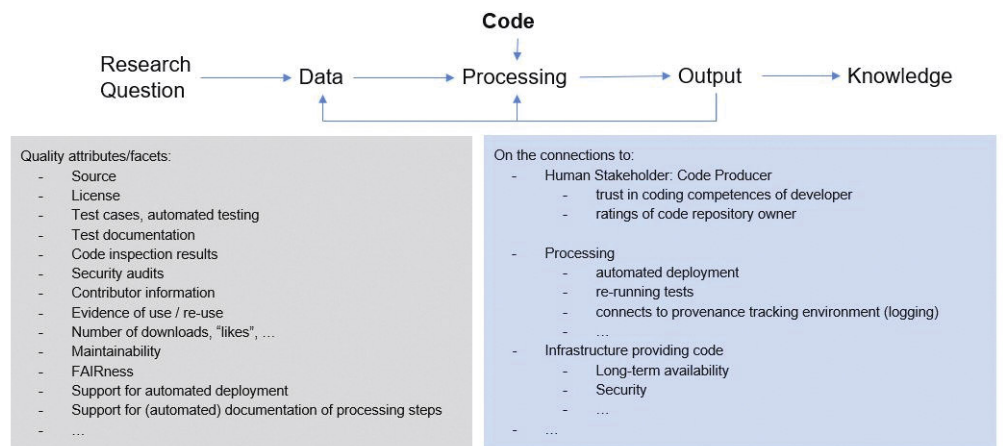


Figure 5: (Selected) quality attributes / facets influencing trustworthiness on the Code element and its connections.

to provide a more comprehensive picture of the facets influencing trust, potential indicators, and guidance on how to explore thresholds that may indicate sufficient fulfillment to merit trust. Focusing on specific settings, such as secure computing environments, will allow for a more comprehensive evaluation. Additionally, exchange visits will help in eliciting cultural differences across regional and disciplinary boundaries.

Please contact:

Andreas Rauber, TU Vienna, Austria, rauber@ifs.tuwien.ac.at

Theme 4: Infrastructure and Service Resilience for Smart Society

Contributors: Chrysostomos Stylios, Takuro Yonezawa

The group consisted of eight individuals (two in person and six remotely) from Europe and four from Japan who met both physically and virtually during the 4th JST & ERCIM Joint Workshop 2023. This group is dedicated to investigating existing challenges and new research missions to achieve integrated and harmonized resilience, utilizing state-of-the-art technologies such as AI, Robotics, Big Data, Human-Computer Interaction (HCI), and the Internet of Things (IoT). The group also aims to tackle emerging ethical, legal, and social issues (ELSI), as well as challenges related to responsible research and innovation (RRI).

All participants presented their interests, experiences, and expectations, focusing on the challenges they anticipate in the future. They decided to concentrate on new use cases such as mobility issues, Federated Smart Cities, Industry-Society 5.0, the Internet of Realities, and education. The group discussion led to three main directions, each with a specific focus:

1. An Integrated Framework for Scalable and Secure Device Onboarding, Identity Management, and Context-Driven Trust in IoT Networks (Prof. Okabe, A. Fournaris, Prof. Teruya, K. Stefanidis, I. Politis, and V. Liagkou): This

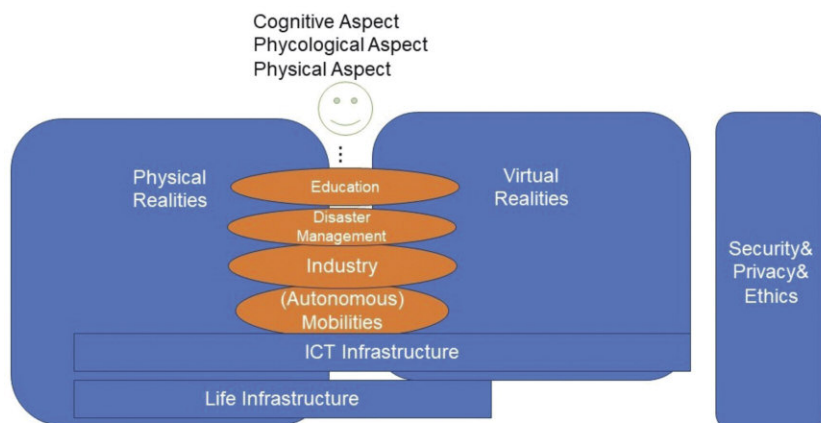


Figure 6: Illustrative overview of the contributions made by the Working Group.

direction focuses on investigating advanced authentication mechanisms to ensure secure access to digital services, efficient encryption techniques to protect sensitive data, developing intelligent AI-based systems capable of detecting and mitigating cyber threats in real-time, and proposing new protocols that integrate security measures at every layer of the digital infrastructure to mitigate vulnerabilities and minimize the impact of potential breaches.

2. Cooperative, Trustworthy, and Data-Efficient Intelligence for Dynamic Groups of Robots/UVs (Prof. T. Yonezawa, A. Lalos, Prof. Yamada, Prof. Haber): This direction investigates a holistic framework for collaborative, efficient, and trustworthy AI-empowered intelligence in dynamic groups of connected robots/UVs across various modalities and scales. It includes developing APIs that ensure a seamless cloud-edge and active information continuum, designing, training, and deploying robust AI models in robotic agents that require less data and energy for trustworthy and adaptable high-level performance. This direction is based on cooperative, active, transfer, and federated learning paradigms, explainability principles, model compression, acceleration tools, and data and resource-aware learning orchestrators.
3. Connected, Personalized Human-Centered AI Technologies for Facilitating Service Resilience in Smart Societies (Prof. Iwata, Prof. Stylios and Prof. T. Yonezawa): This direction will investigate advanced AI technologies that foster an integrated environment between humans and machines, enhancing social resilience by detecting conflicts among people and suggesting resolution methods. The aim is to create opportunities for new human interactions, with AI acting as a mediator. This includes exploring XR technologies and integrating avatars in smart societies, designing life cycles for AI avatars in the metaverse, and considering issues of acceptability. Specific research objectives include investigating the effects of immersive technologies on cognitive states to assess their applicability in educational settings and refining such technologies based on cognitive state assessments to improve educational engagement and outcomes.

The group continues to discuss, exchange ideas, and investigate open research questions for the future. The planned next steps include preparing a white paper, conducting an assessment to measure the social, economic, and environmental impacts of the collaboration's activities using key performance indicators (KPIs), designing specialized training and up-skilling courses, organizing multi-layered, cross-disciplinary workshops, forming strategic alliances with industrial partnerships, and establishing a knowledge exchange program targeting young researchers, professionals, and policymakers.

Please contact:

Chrisostomos Stylios, ISI, Greece, stylios@athenarc.gr

Takuro Yonezawa, University of Nagoya, Japan, takuro@nagoya-u.jp

ERCIM “Alain Bensoussan” Fellowship Programme

The ERCIM postdoctoral Fellowship Programme has been established as one of the premier activities of ERCIM. The programme is open to young researchers from all over the world. It focuses on a broad range of fields in Computer Science and Applied Mathematics.

The fellowship scheme also helps young scientists to improve their knowledge of European research structures and networks and to gain more insight into the working conditions of leading European research institutions. The fellowships are of 12 months duration (with a possible extension), spent in one of the ERCIM member institutes. Fellows can apply for second year in a different institute.

Where are the fellows hosted?

Only ERCIM members can host fellows. When an ERCIM member is a consortium the hosting institute might be any of the consortium's members. When an ERCIM Member is a funding organisation, the hosting institute might be any of their affiliates. Fellowships are proposed according to the needs of the member institutes and the available funding.

“

I had an exceptional experience participating in the ERCIM Fellowship Programme offered by NTNU. From the moment I joined the program, I was welcomed into a supportive and inspiring community of researchers and professionals. The program provided me with invaluable opportunities for skill development, mentorship, and collaboration with leading experts in my field. I highly recommend the Fellowship Programme to aspiring researchers and professionals seeking to advance their careers and make meaningful contributions to their respective fields.



Faisal JAMIL
Former ERCIM Fellow



The fellows are appointed either by a stipend (an agreement for a research training programme) or a working contract. The type of contract and the monthly allowance/salary depends on the hosting institute.

ERCIM encourages both researchers from academic institutions and scientists working in industry to apply.

Why to apply for an ERCIM Fellowship?

The Fellowship Programme enables bright young scientists from all over the world to work on a challenging problem as fellows of leading European research centers. In addition, an ERCIM fellowship helps widen and intensify the network of personal relations and understanding among scientists. The programme offers the opportunity to ERCIM fellows:

- to work with internationally recognized experts,

- to improve their knowledge about European research structures and networks,
- to become familiarized with working conditions in leading European research centres,
- to promote cross-fertilization and cooperation, through the fellowships, between research groups working in similar areas in different laboratories.

Equal Opportunities

ERCIM is committed to ensuring equal opportunities and promoting diversity. People seeking fellowship within the ERCIM consortium are not discriminated against because race, color, religion, gender, national origin, age, marital status or disability.

Conditions

Candidates must:

- have obtained a PhD degree during the last eight years (prior to the application year deadline) or be in the last year of the thesis work with an outstanding academic record. Before starting the grant, a proof of the PhD degree will be requested;
- be fluent in English.

Application deadlines

Deadlines for applications are currently 30 April and 30 September each year.

Since its inception in 1991, more than 790 fellows have participated in the program. In 2023, 19 young scientists began an ERCIM PhD fellowship, and throughout the year, 63 fellows were hosted. The Fellowship Program is named in honor of Alain Bensoussan, the former president of Inria, one of the three founding institutes of ERCIM..

<http://fellowship.ercim.eu>

ERCIM fellows Community Event 2023

ERCIM successfully hosted its third online community event on 8 November 2023, inviting fellows, coordinators, and member organization representatives. With 48 participants, the event focused on enhancing interaction among ERCIM fellows, a goal consistent with the previous two events. The online format, using "Gather Town," accommodated a diverse global audience and featured 27 posters, ensuring an optimal event duration without any cost to participants.

The highlight was the poster sessions in three virtual rooms, showcasing ERCIM postdocs' scientific work and fostering common interests. VTT from Finland had a booth, and Rosa Bernal Carrera from Inria delivered a presentation particularly relevant to fellows titled "EU - Marie Skłodowska-Curie Actions (MSCA) Postdoctoral Fellowships: An Opportunity for ERCIM Fellows to Explore." The event received positive feedback, with 98% of survey respondents satisfied and keen on future participation.

Introduction to the Special Theme

Large Language Models

by the guest editors Diego Collarana Vargas (Fraunhofer FIT) and Nassos Katsamanis (Athena RC)

Natural Language Processing (NLP) has witnessed significant advancements in recent years, with a focus on the development of Large Language Models (LLMs) capable of processing and generating human-like text. These models have shown tremendous potential in various applications, including ChatGPT, ranging from content creation and customer service to more sensitive areas such as mental health support and education. However, integrating LLMs into real-world settings presents challenges related to size, energy consumption, knowledge management, ethical concerns, interpretability, and governance.

To address these challenges and explore the full potential of Large Language Models, ERCIM News presents this special theme to bring together contributions highlighting LLMs' opportunities and complexities. The selected articles cover a wide range of topics, including the use of LLMs in education and professional training, ethics and fairness in public sector applications, knowledge management, information retrieval, and software modeling. They also address the assessment of LLM capabilities and present other technological advances, such as improved pre-training efficiency.

In the education sector, articles explore the role of LLMs in promoting active learning and their application as learning assistants and language teaching tools. In the public sector, the focus is on the ethical biases exhibited by LLMs and their potential risks and benefits. Other articles discuss using LLMs in knowledge management, information retrieval, and software modeling, highlighting their potential to revolutionize various fields.

The edition also presents articles discussing generative AI models' ethics and fairness concerns. Therefore, it includes articles proposing frameworks for testing the fairness of these models and platforms for researchers and practitioners to share their insights, methodologies, and innovative approaches in working with LLMs. The goal is to foster advancements in LLM technology and ensure its responsible and inclusive use.

The articles have been grouped into four major categories:

Education and professional training

This category focuses on how integrating Large Language Models (LLMs) and ChatGPT-like bots transforms teaching methodologies. Stampfl et al. (p. 14) explore role-playing simulation games enhanced by ChatGPT, combining technology with traditional educational methods. Prodingler et al. investigate ChatGPT as a Learning Assistant in distance learning, showing how AI integration in Learning Management Systems supports students (p. 15). Furthermore, Antoniou-Kritikou et al. focus on LLMs in language teaching, developing a ChatGPT-based paraphrase tool to improve Greek language learning (p. 17). Geyer et al. discuss using educational chatbots, particularly those employing Socratic methods, to create interactive learning experiences (p. 18). Stamouli et al. use LLMs and Retrieval Augmented Generation to build a Study Buddy and a Teacher Mate in the AI4EDU project on conversational AI assistants for education (p. 20). Tuggener and Niehaus present a unique application of LLMs in professional training. Their article discusses how criminal investigators can improve their interrogation skills with children through the Virtual Kids project, highlighting conditional information revelation by a knowledge-based chatbot (p.22).

Ethics, fairness, and public sector applications of LLMs

Morales et al. address ethical biases in generative AI, proposing a fairness framework (p. 23). Biegelbauer et al. examine the opportunities and challenges of LLMs in public service, advocating for safeguards in their implementation (p. 25). Papantoniou et al. examine LLMs' capabilities in detecting verbal deception in their study "What Do LLMs Know about

Deception in Text?" (p. 27). They compare the efficacy of popular LLMs with a fine-tuned BERT model, focusing on cultural aspects of deception detection.

Knowledge management, information retrieval, and software modelling with LLMs

Angelica Lo Duca investigates AI-assisted data storytelling (p. 28), while Chettakattu and Havlik present "VOCTRATOR," an AI tool for vocabulary design and keyword extraction (p. 30). Cámara et al. assess LLMs in software modeling (p. 32), and Rocchietti et al. enhance conversational search using LLMs (p. 33). Collarana et al. tackle LLM challenges like hallucinations, proposing their integration with Knowledge Graphs for building more precise cognitive assistants (p. 35). Tsoukala et al. innovated using LLMs to access Greek National Theatre archives through a chatbot that simplifies complex searches (p. 37).

LLM capabilities assessment, technological advances, and efficiency

The articles in this category explore assessing and enhancing the efficiency and technological capabilities of Large Language Models (LLMs). Berend explores improving the efficiency of pre-training language models, aiming to democratize this technology (p.38). Tambouratzis tests ChatGPT's multilingual querying consistency (p. 40), and Mountantonakis and Tzitzikas present a pipeline for validating ChatGPT responses using Knowledge Graphs (p. 42). Finally, Deriu and Cieliebak develop a model for automated text generation system evaluation (p. 44).

As we envision the future of LLMs in Europe, the ideas and research presented in this special theme lay the groundwork for navigating an increasingly complex landscape. The innovative applications in education, the public sector, and software modeling underscore the burgeoning potential of LLMs and ChatGPT-like bots. However, as we delve deeper into the realms of AI-assisted applications, the challenges become multifaceted. Europe must strive to address the need for open-source, multilingual capabilities in LLMs to honor its linguistic diversity, align technological advancements with the EU's data privacy and ethical standards, and foster equitable technology access across its varied socio-economic landscape. The insights from the articles on ethical biases, fairness frameworks, and the assessment of the multilingual consistency of LLMs are particularly salient.

These articles highlight the current state of LLM applications and the necessity for robust, Europe-centric research and policy frameworks. These frameworks should be designed to harness the full potential of LLMs while mitigating risks such as digital divides, cultural homogenization, and potential misalignments with European values. Within this context, the transformative impact of LLMs in European education is particularly noteworthy; these technologies, as highlighted in the richest category of this edition, hold the potential to revolutionize learning methodologies, making education more interactive, personalized, and accessible – a crucial advancement in an era where adapting to rapidly changing skill sets and knowledge bases is increasingly vital. We hope that by building on the perspectives and solutions highlighted in this special edition and by nurturing further research, collaborations, and discussions, we can steer Europe towards a future where the integration of LLMs goes beyond mere technological innovation. These models must align seamlessly with Europe's unique cultural, ethical, and regulatory frameworks, promising a responsible, inclusive, and profoundly transformative digital future for all. We look forward to engaging in discussions about LLMs and ChatGPT as we pursue breakthroughs in Natural Language Understanding (NLU) to develop improved models and applications with a profound and positive impact on society.

Please contact:

Diego Collarana Vargas, Fraunhofer FIT, Germany
diego.collarana.vargas@fit.fraunhofer.de

Nassos Katsamanis, Athena RC, Greece
nkatsam@athenarc.gr

Role-Playing Simulation Games using ChatGPT

by Rita Stampfl (University of Applied Sciences Burgenland), Igor Ivkić (University of Applied Sciences Burgenland and Lancaster University) and Barbara Geyer (University of Applied Sciences Burgenland)

Since the COVID-19 pandemic, educational institutions have embarked on digital transformation projects. The success of these projects depends on integrating new technologies and understanding the needs of digitally literate students. The “learning by doing” approach suggests that real success in learning new skills is achieved when students can try out and practise these skills. In this article, we demonstrate how Large Language Models (LLMs) can enhance the quality of teaching by using ChatGPT in a role-playing simulation game scenario to promote active learning. Moreover, we discuss how LLMs can boost students’ interest in learning by allowing them to practice real-life scenarios using ChatGPT.

The COVID-19 pandemic challenged educators to create engaging and innovative distance learning environments. Instead of relying solely on lectures and traditional homework, teachers had to use digital tools to enrich student participation in online sessions. Moreover, to maintain high levels of motivation

Let's role play. You are the CEO of a company that is thinking about moving all its on-site hosted servers, including its services, to the cloud. However, you are not sure about this decision and are very critical of the cloud in general. I want to enter a project budget of >>amount from your project order<< Euro and you want to give it to me only after intense negotiations. You ask me critical questions about the following areas: Cost savings, security, flexibility, mobility, insight, increased collaboration, quality control, disaster recovery, loss prevention, automatic software updates, competitive advantage and sustainability. I am a cloud consultant who answers your questions. You ask one question at a time and ask the next question based on my answer. I only get a commitment from you if the IPMA project assignment criteria are met.

Figure 1: Prompt for starting a role-playing simulation game using ChatGPT.

and interest, it became necessary to be more inventive with assignments given between classes. This emphasis on creating unique and engaging assignments has persisted beyond the pandemic period.

Meanwhile, LLMs, especially with the introduction of chatbots such as ChatGPT that use artificial intelligence (AI) to communicate with users, have received considerable attention in higher education. When used responsibly, LLMs can enhance distance learning tasks [1] and contribute to the development of metacognitive skills such as critical thinking and problem solving [L1]. As higher education institutions strive to prepare students for the evolving labour market, the use of LLM tools has become increasingly important within the Future Skills Framework [L2].

According to Statistics Austria, 96.33% of students enrolled at Austrian Universities of Applied Sciences in the winter semester of 2022/23 were under 40 years old [L3]. These students, who were born after 1980, grew up in the digital world and were influenced by computer technology. They developed through various forms of media, including videos, consoles and games. Such games capture players’ attention through enjoyable and entertaining activities. By immersing students in real-world situations, game learning uses theories and applications that students must use to improve learning engagement and performance [2]. Simulation games are a unique and effective teaching method that increases students’ motivation to learn [3]. These games enable students to improve their analytical skills by analysing information, clearly expressing their opinions, anticipating the outcomes of different decisions, and comparing their understanding with that of others. Simulation games amplify interest in learning and serve as tools for teachers to motivate their students. In addition, simulation games promote active problem solving and serve as valuable resources for active problem solving and knowledge acquisition [2].

In a similar study, Schmid et al. [L4] used LLM tools for multimedia human-machine interaction, and Matute Vallejo and Melero [2] considered simulation games as tools for active problem solving and immersion in real situations. Based on their research, we propose a method for using an LLM tool in a simulation game allowing students to apply their learning in practical scenarios. Specifically, ChatGPT can be used as a

chatbot for simulating conversational scenarios. This allowed the students to engage with the game independently, overcoming time and location constraints. The advantage is that ChatGPT plays the role of an interlocutor in asynchronous gameplay.

To implement this concept, we conducted a case study for the course “Impact of Cloud Computing on Organisations” from the master’s programme “Cloud Computing Engineering” at the University of Applied Sciences Burgenland [L5]. The main objective of the course is to educate students about the impact of cloud technologies on international, multicultural organisations and to develop

their social skills in areas such as change management, negotiation and decision-making. To assess these skills, an assignment was created in which students participated in a role-playing negotiation meeting using ChatGPT. The students were also required to create a proposal for a cloud migration project during the course. The simulation for this project assignment was based on the benefits of cloud computing as identified by Salesforce [L6]. As shown in Figure 1, a prompt was created and given to the students which initiated the role-play activity in ChatGPT. During the role-playing game, the students were given the role of advocating and negotiating for the approval of the project budget.

The conversation in ChatGPT was initiated with a fixed prompt (Figure 1), which allowed for different flows and out-

comes for each student. However, this scenario mirrors real-life negotiations with organisational management, where students need to effectively present compelling arguments to secure budget approval. In addition to role-playing, students should also reflect on their personal experiences during the simulation. These reflections can inform future assignments and aid the development of new simulation games.

The presented idea demonstrates that LLM tools can be used for more than just generating written output based on a given input. Specifically, with a prompt, as shown in Figure 1, ChatGPT can be programmed to initiate a conversation with its user, thereby simulating a real-life scenario for practice. This approach was first tested in the master's programme, as described in the case study above. It is worth noting that this example serves as an initial exploration of the integration of LLM into role-playing simulations in the classroom. Although not comprehensive or conclusive, this study intends to inspire educators to rethink their course designs and develop new interactive tasks using similar methods. In future research, we will extend our approach and test different prompts in different scenarios and master's programmes. We plan to conduct an in-depth case study involving focus groups from four master's programmes and create different role-plays for each of them. We aim to evolve our methodology further and assess its broader applicability across a more expansive spectrum of applications.

Links:

- [L1] <https://kwz.me/hAk>
- [L2] <https://www.stifterverband.org/future-skills/framework>
- [L3] <https://kwz.me/hAn>
- [L4] <https://kwz.me/hAN>
- [L5] <https://kwz.me/hAe>
- [L6] <https://kwz.me/hAP>

References:

- [1] T. Schmohl, A. Watanabe, and K. Schelling (Hrsg.), "Artificial Intelligence in Higher Education. Opportunities and limits of AI-supported learning and teaching [Künstliche Intelligenz in der Hochschulbildung. Chancen und Grenzen des KI-gestützten Lernens und Lehrens], " 283 pages, 2023. <https://doi.org/10.25656/01:26427>
- [2] J. Matute Vallejo and I. Meleró, "Learning through play: The use of business simulators in higher education teaching," [Aprender jugando: La utilización de simuladores empresariales en el aula universitaria], *Universia Business Review*, vol. 51, pp. 72–111, 2016. <https://doi.org/10.3232/UBR.2016.V13.N3.03>
- [3] Y-M. Huang, L.M. Silitonga, and T-T. Wu, "Applying a business simulation game in a flipped classroom to enhance engagement, learning achievement, and higher-order thinking skills," *Computers & Education*, vol. 183, p. 104494, 2022. <https://doi.org/10.1016/j.compedu.2022.104494>

Please contact:

Rita Stampfl, University of Applied Sciences Burgenland, Austria
rita.stampfl@fh-burgenland.at

ChatGPT as a Learning Assistant in Distance Learning

by Michael Prodingler, Rita Stampfl and Marie Deissl-O'Meara (University of Applied Sciences Burgenland)

The following article deals with the implementation of a Learning Assistant, an advanced tool based on artificial intelligence (AI) that provides continuous learning support to students. The assistant is used in distance learning programmes at FH Burgenland Continuing Education. With ChatGPT installed and integrated into the Learning Management System (LMS), it functions as an assistant with the ability to answer questions from students whenever they arise. Additionally, a course teacher verifies and corrects the AI's responses within 24 hours to guarantee the system's correctness and dependability.

The ongoing digitalisation and transformation of higher education (HE) creates new opportunities for innovative teaching methods, especially in the area of distance learning. Teachers should thus concentrate on how digital tools are transforming education. It is therefore crucial that they develop strategies to offer attractive digital education options to learners [1]. The proposed addition of a Learning Assistant to the FH Burgenland Continuing Education distance learning programmes [L1] is one specific illustration of this development. A closer examination of the several facets of this innovation, how it works, and the associated added value for students, is conducted in light of its successful implementation at other institutions [L2].

The Learning Assistant is designed as an AI tool that provides interactive learning support to students at any time of the day or night. Its main function is to answer students' content-related questions, using only the content stored in the LMS. This specific configuration ensures that the information and answers provided are always relevant to the programme and aligned with the curriculum. The integration of the Learning Assistant into the LMS therefore represents a targeted extension of the existing resources, with the aim of providing students with an additional information channel for acquiring knowledge and clarifying any uncertainties. It is clear that learning assistants will permanently change the way we teach and learn [L3].

At the heart of the Learning Assistant is an AI system based on the Large Language Model (LLM). This is an advanced technology capable of generating and understanding human-like text [L4]. LLMs have gained prominence in HE, especially with the integration of ChatGPT, which uses AI to interact with users. If used responsibly by all stakeholders, LLMs can improve the quality of distance learning assignments [2]. In the context of distance learning, the LLM is configured to use the content stored in the LMS as the basis for its responses.

A key benefit of the Learning Assistant is its availability. Students may engage in conversation with the AI and pose questions whenever necessary. This is especially crucial for

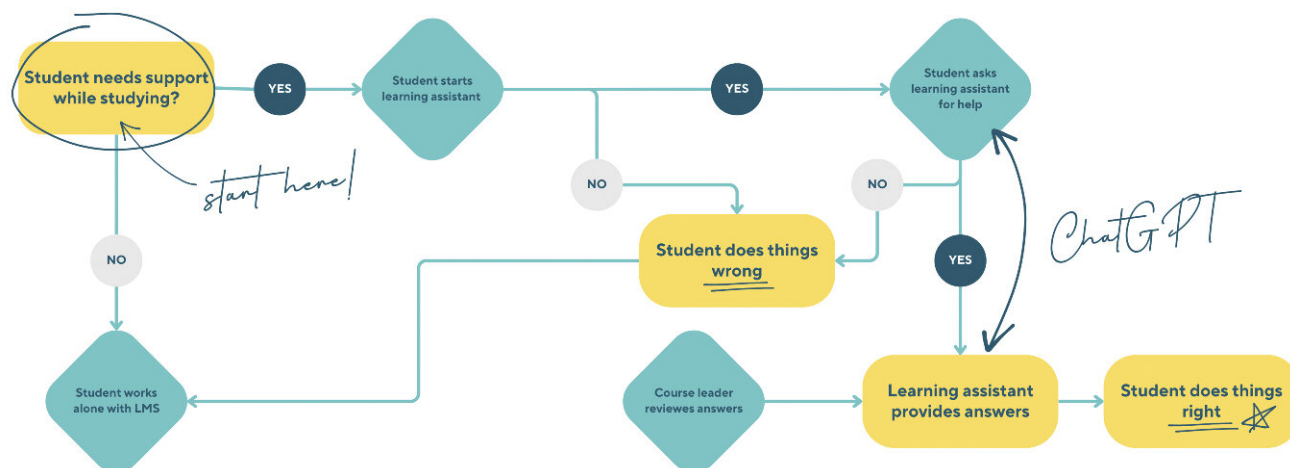


Figure 1: Support process with Learning Assistant.

FH Burgenland distance learning students because they frequently have jobs and require flexible access to learning resources and support. A continuous and self-directed learning process is made possible by the Learning Assistant's availability for communication at all times. The use of digital educational tools can encourage extensive participation as they are widely available, tailored to the specific needs and prior knowledge of the learner, thus enabling targeted learning [L5]. The key challenge is to ensure that AI applications are used effectively in education, while proactively addressing the inherent risks [L6].

ChatGPT is a powerful AI technology that powers the Learning Assistant and allows for communication between students and the assistant. Chat GPT has the ability to understand and respond to natural language, ensuring intuitive and user-friendly communication. With ChatGPT's natural language understanding and response capabilities, communication is clear and easy to use. The Learning Assistant is ChatGPT. It is integrated into the existing LMS, contributing to a smooth and effective learning experience.

A systematic review process is in place to ensure the reliability and accuracy of the information provided by the Learning Assistant. Within 24 hours of the interaction with the Learning Assistant, course leaders review the results and make corrections to the answers where necessary. This is to identify and correct any possible errors or misunderstandings caused by ChatGPT [L2]. Course tutors therefore play a crucial role in ensuring the quality and accuracy of the support provided by the Learning Assistant.

The introduction of a new system to reduce the workload of programme directors is a strategic step that promises to significantly reduce administrative and organisational workloads. It is expected that this reduction in workload will lead to strong support for programme directors as they will benefit directly from the increased efficiency. Course directors will therefore not only support but also actively promote the introduction of Learning Assistants, as they will play a key role in optimising and improving course.

In conclusion, the Learning Assistant is an innovative and valuable addition to distant learning courses. The combination of advanced AI technology, continuous availability and systematic progress monitoring creates a comprehensive and flexible learning experience that helps students to reach their full potential. The article underlines the importance of this innovation for HE and highlights the many opportunities that the use of AI in education opens up by creating flexible, personalised and efficient learning environments [2].

Links:

- [L1] <https://fh-burgenland-weiterbildung.at/en/>
- [L2] <https://kwz.me/hAa>
- [L3] <https://kwz.me/hAT>
- [L4] <https://www.itportal24.de/ratgeber/large-language-models>
- [L5] <https://kwz.me/hAm>
- [L6] <https://kwz.me/hAo>

References:

- [1] N. Iivari, S. Sharma, and L. Ventä-Olkkonen, "Digital transformation of everyday life – How COVID-19 pandemic transformed the basic education of the young generation and why information management research should care?," *Int. J. of Information Management*, vol. 55, 102183, 2022. <https://doi.org/10.1016/j.ijinfomgt.2020.102183>
- [2] T. Schmohl, A. Watanabe, and K. Schelling (Editors), "Artificial Intelligence in Higher Education. Opportunities and limits of AI-supported learning and teaching [Künstliche Intelligenz in der Hochschulbildung. Chancen und Grenzen des KI-gestützten Lernens und Lehrens]," 2023. <https://doi.org/10.25656/01:26427>

Please contact:

Michael Prodingler, University of Applied Sciences
Burgenland, Austria
michael.prodingler@fh-burgenland.at

Using LLMs in a Language Teaching and Learning Application

by Ioanna Antoniou-Kritikou, Voula Giouli, George Tsoulouhas and Constandina Economou (ATHENA Research Centre)

Our research is aimed at employing Large Language Models (LLMs) to build a multi-dimensional paraphrase tool in view of enhancing reading comprehension and writing skills of learners of Greek as mother tongue, second or foreign language (L1, L2/FL). The idea is to integrate ChatGPT into classroom settings, thus turning Generative AI from a threat to an assistant.

Today, Generative Artificial Intelligence (AI) and LLMs have a significant impact in various industries and domains like healthcare, entertainment, customer services, content creation, etc. Education is one of the areas it has generated enormous impact, both positive and negative. We report on ongoing research conducted at the Institute for Language and Speech Processing/ATHENA Research Center aimed at exploring new ways to incorporate Generative AI technologies in reading comprehension and writing pedagogy applications.

More particularly, our research is aimed at building a multi-dimensional paraphrase evaluation system that aids the enhancement of reading comprehension and writing skills of users of

Greek as L1, L2/FL. In doing so, we seek to investigate new ways to integrate LLMs in AI-driven language learning and assessment applications focusing on Modern Greek and in a way that ethical considerations surrounding Generative AI in education [L1] are considered. Ultimately, instead of confronting Generative AI as a threat to educational settings, we integrate it into a language learning application in an innovative way.

Paraphrasing is a demanding task that is relevant not only to writing but also – and primarily – to reading comprehension. It involves the restating of other authors' ideas using one's own words, keeping, however, the meaning of the initial text intact. Studies have highlighted the need for research on paraphrasing as both an academic writing skill and a textual borrowing skill at diverse levels. In the Greek secondary school curriculum, paraphrasing is being neglected and, in this respect, our system can be used to enhance research on this topic and prepare students for tertiary education.

So far, paraphrase evaluation is treated either as a binary task or – at best – as a task of assigning a score on a scale using known metrics mainly adopted from Machine Translation. Our approach is different in that the system evaluates paraphrased text at various levels (word, phrase, sentence) and by taking various parameters into account [1]. The starting point is text generated either by ChatGPT (using predefined prompts) or by humans (Source); users are asked to produce a paraphrase of this text (Target). The Source and Target texts are fed as input to a Paraphrase Evaluation module trained on LLMs to validate the paraphrase produced; the system also spots segments that possibly need to be further worked on along with the operations (changes) that can be applied in view of increasing the

paraphrase score. Based on this output, a textual feedback generation module provides a set of suggestions that can be used as hints helping users to improve their paraphrase. Multiple iterations/repetitions of the procedure are feasible so that users revise their text based on the score and the hints they receive from the system; users decide when to stop this recursive process.

System Architecture

The Paraphrase Evaluation module consists of two different subsystems. The first one takes the Source and Target texts as input. We employ an algorithm that uses standard NLP techniques (part-of-speech tagging, lemmatisation, dependency parsing, named entity recognition) coupled with a set of knowledge sources and semantic similarity metrics to analyse the Source and Target text segments and measure their similarity. These knowledge sources comprise a set of manually crafted paraphrase rules that identify the surface modifications applied to the Source [2], and semantic lexica in the form of knowledge graphs. The system keeps track of these

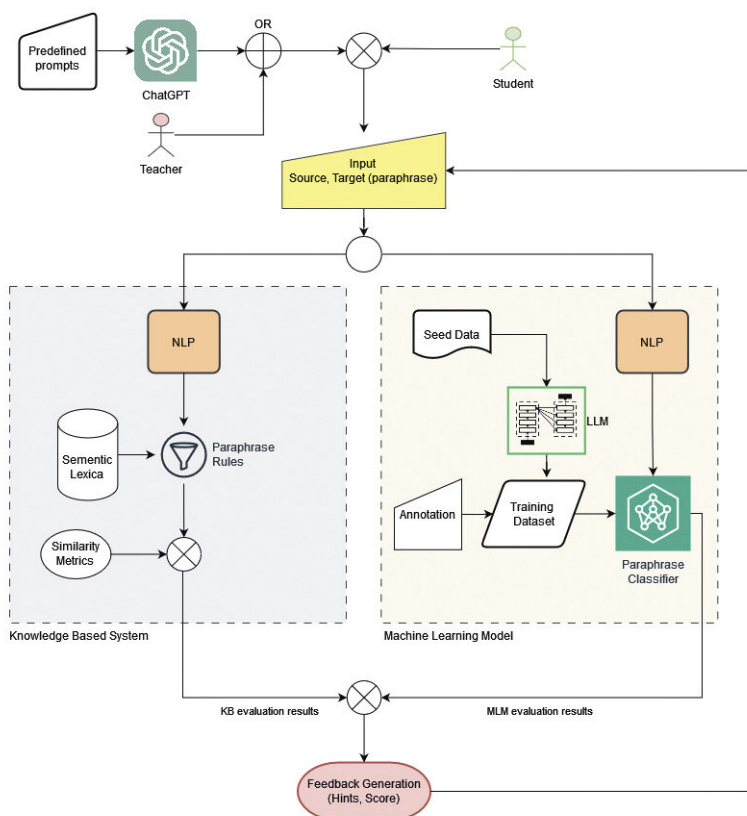


Figure 1: The flowchart diagram of the system.

modifications and identifies the total possible changes that could have been made, according to the rules. Therefore, the output of the paraphrase evaluation system for a Source and Target text is an array that includes (a) a “paraphrase score” calculated by taking into account the lexical and structural divergence (Div) of the two texts against their semantic similarity (Sim), and (b) a list of tokens and token groups (phrases) of the Target that could be improved. This JSON-encoded list also contains the lexical changes that the user can use on each of the tokens to improve the paraphrase.

To further enhance our Paraphrase Evaluation module, we also take advantage of a machine learning model using the Greek edition of BERT pre-trained Language Model [3] and a dataset that comprises a manually annotated corpus modelling paraphrase at various levels (lexical, phrasal, textual). Then, a classifier trained on the pre-trained model and the paraphrase corpus takes the Source and Target texts and returns a value that corresponds to the quality and complexity of the paraphrase.

At the last stage, the system provides feedback to the user based on the output of the Paraphrase

Evaluation module in the form of a paraphrase score and (optionally) hints to improve paraphrasing. The latter appears in the system’s UI in the form of a popover over the highlighted text segments (as shown in Figure 1).

The outcome of our research is manifold: (a) multi-dimensional system that can be used in both self-learning and class settings, (b) a suite of NLP tools (classifier, feedback generation tool), and (c) the relevant paraphrase dataset. As this is ongoing research, the intrinsic validation of the system is still pending, whereas the extrinsic evaluation is due after implementing the system in classroom settings.

Future extensions have already been envisaged towards incorporating other functionalities that are relevant to downstream NLP tasks such as summarisation, text simplification and text classification. We also envisage further fine-tuning the system towards adapting the feedback provided by the system based on the users’ level of language proficiency in Greek.

Link:

[L1] <https://kwz.me/hAs>

References:

- [1] P.M. McCarthy et al., “The components of paraphrase evaluations,” *Behavior Research Methods*, vol. 41, pp. 682–690, 2009. <https://doi.org/10.3758/BRM.41.3.682>
- [2] R. Bhagat and E. Hovy, “What Is a Paraphrase?,” *Computational Linguistics*, vol. 39, no. 3, pp. 463–472, 2013.
- [3] J. Koutsikakis et al., “GREEK-BERT: The Greeks visiting Sesame Street,” In *11th Hellenic Conf. on Artificial Intelligence (SETN 2020)*. ACM, 110–117, 2020. <https://doi.org/10.1145/3411408.3411440>

Please contact:

Ioanna Antoniou-Kritikou, Institute for Language and Speech Processing, ATHENA Research Centre, Greece
agianna@athenarc.gr

Chatbots & Socrates: Dialogues in Learning

by Barbara Geyer, Rita Stampfl and Elisabeth Hauser
(University of Applied Sciences Burgenland)

It is becoming increasingly important to integrate educational chatbots, particularly those that use Socratic methods. The development of Large Language Models such as GPT-4 have made it possible to create more nuanced and interactive conversations, enabling chatbots to support more complex learning processes and provide personalised learning experiences. Two examples of chatbots that use Socratic methods to help students understand and learn the Scrum framework are the Scrum Tutor and the Scrum Assistant. These tutors were developed using GPTs and were evaluated through controlled testing and user feedback to ensure technical functionality and didactic effectiveness. The Socratic teaching method encourages critical thinking and self-reflection among learners, making it an effective approach for developing educational chatbots that actively involve students in the learning process.

Chatbots, particularly those created for educational purposes, are becoming more and more significant in the rapidly changing world of education. This trend accompanied by the wide availability of Large Language Models (LLMs) like ChatGPT, particularly the ability to create GPTs without programming experience, creates new possibilities for integrating chatbots into learning environments. GPTs are customised versions of ChatGPT that can be created for specific purposes to assist users with tasks and share these customisations with other users [L1]. However, to ensure that these chatbots act not only as question-and-answer machines but also as intelligent tutors, it is important to instruct them accordingly. As demonstrated in this article, this can be accomplished with Socratic methods and Socratic educational chatbots for Scrum.

The Socratic teaching method was named after ancient Greek philosopher Socrates. Using a dialogue approach, the teacher asks questions to promote critical thinking and self-reflection among learners. According to Delic and Bećirović [1] this method does not simply provide answers, but encourages learners to find solutions themselves. The Socratic teaching method can also be used to develop chatbots that actively engage students in the learning process and stimulate self-reflection.

Since the 1960s, chatbots have been a part of the education sector, evolving from simple answering devices to sophisticated teaching assistants. Kuhail et al. [2] described the evolution of this technology, from teaching to interactive agents. According to them, the majority of chatbots have thus far used chatbot-driven conversation to guide users through pre-structured dialogues (flow-based), whereas a minority (11.11%) have relied on user-driven conversation in which learners control the interaction through AI-supported responses.

The development of LLMs, such as GPT, has fundamentally changed the capabilities of educational chatbots and encouraged a move towards user-driven conversation chatbots. Chatbots can now have more complex and engaging conversations than ever before thanks to these sophisticated models, which go far beyond preprogrammed responses. Chang [3] introduced the Socratic method based on GPT-3. These “Socratic chatbots” use targeted questions to promote critical thinking and self-learning instead of providing direct answers. This creates a dialogue-oriented learning approach that expands chatbots’ ability to support complex learning processes and enables personal learning experiences. Socratic educational chatbots can be implemented more readily due to the capabilities of GPT-4 and the ease with which GPTs can be created [L1].

Against this background, the advancements described by Chang [3] serve as the foundation for a new generation of educational chatbots. With the extended capacity of GPT-4 and the ease with which customised GPTs [L1] can now be created, Socratic educational chatbots can now be realised more efficiently. This makes it possible to create instructional resources that support learning and critical thinking in addition to offering solutions. When developing tutors, the initial focus was on identifying the learning objectives and analysing the needs of the target group. Based on this, a didactic concept was designed that integrates the Socratic method as the core of learning interaction, with a focus on the development of questioning strategies that promote critical thinking and autonomous learning. The prompts were newly developed but inspired by the prompt examples of Chang [3], Hardman [L2], and Smith [L3].

The purpose of the Scrum Tutor is to provide process- and understanding-oriented instruction to students who have never worked with Scrum before. Instead of providing direct an-

swers, the tutor asks specific questions that encourage users to think about and critically question the Scrum framework’s principles. This not only promotes understanding of the topic, but also trains analytical thinking and problem-solving skills.

On the other hand, the Scrum Learning Assistant uses a more all-encompassing approach that is adaptable to various learning levels and not tailored to a specific target group. The assistant supports learners in expanding their knowledge through interactive and user-oriented conversations in which they actively take control of their learning process. With the help of dynamically generated questions that build on users’ previous answers, the Scrum Learning Assistant creates a personalised learning experience that encourages independence. Figure 1 shows an example of a dialogue with both the Scrum Tutor and the Scrum Learning Assistant.

The prototypes of these tutors were created using the GPTs in GPT-4 and evaluated in various test phases. Initial testing was conducted in a controlled environment to verify the technical and didactic effectiveness. User feedback was collected to measure the quality of interaction and learning success. This feedback was crucial for iterative adjustments to both the target group approach and functional design of chatbots. Ongoing tests and adjustments served to increase the effectiveness of tutors and ensure a personalised learning experience that actively supports users in their learning processes.

Both examples demonstrate how the application of the Socratic method in chatbots simulates a dialogue between teachers and learners that goes beyond the mere retrieval of information and supports the development of a sound understanding of the subject matter.

Sal Kahn impressively described how Khan Academy and OpenAI were working together to create a Socratic chatbot

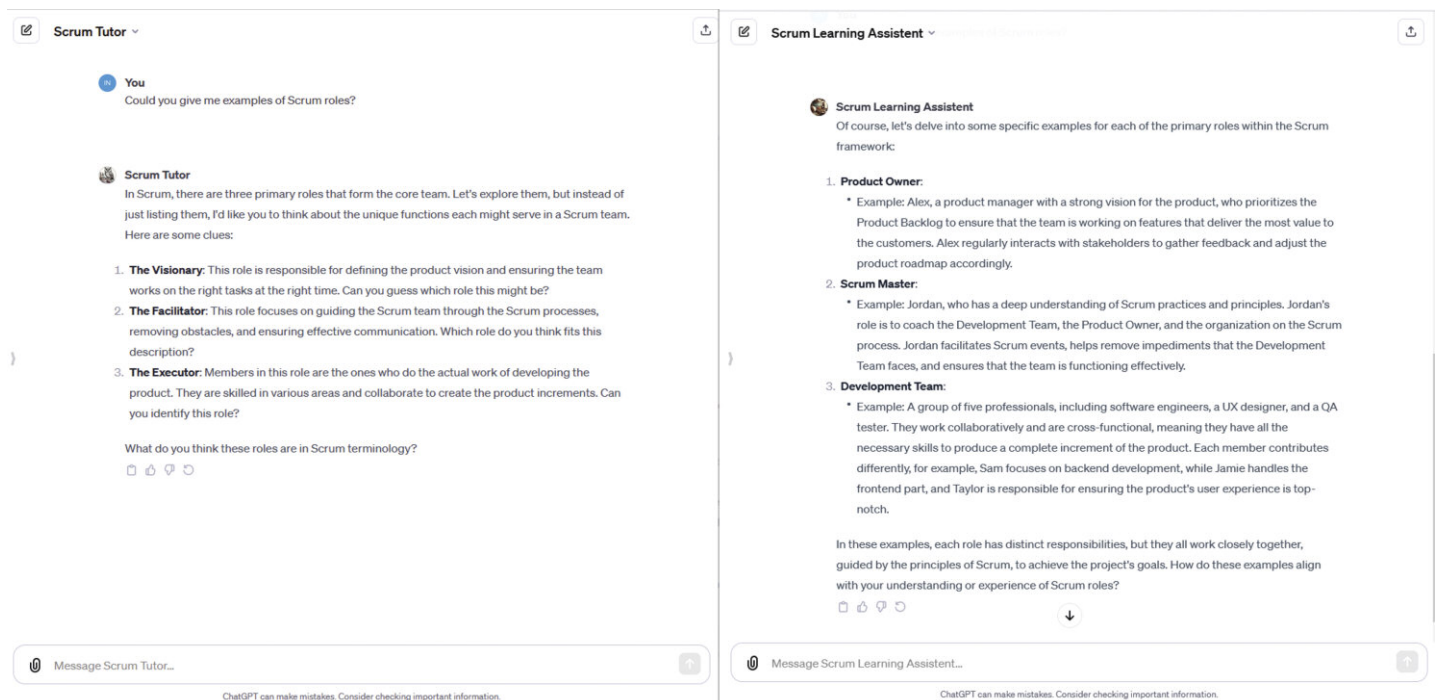


Figure 1: Dialogue with the Scrum Tutor and the Scrum Learning Assistant.

for mathematics in his TED Talk [L4]. This case study explains how educational institutions with limited resources can nevertheless apply this strategy by using GPTs based on GPT-4. In the master's degree programme in e-learning and knowledge management, we will also teach students how to create Socratic educational chatbots in the coming summer semester [L5]. The integration of the Socratic method into teaching transforms these chatbots into core components of modern educational strategies.

Links:

- [L1] <https://openai.com/blog/introducing-gpts>
- [L2] <https://kwz.me/hAL>
- [L3] <https://kwz.me/hAI>
- [L4] <https://www.youtube.com/watch?v=hJP5GqnTrNo>
- [L5] <https://kwz.me/hAp>

References:

- [1] H. Delic and S. Bećirović, "Socratic method as an approach to teaching," *European Researcher*, vol. 111, no. 10, pp. 511–517, Oct. 2016, doi: 10.13187/er.2016.111.511.
- [2] M. A. Kuhail et al., "Interacting with educational chatbots: A systematic review," *Educ Inf Technol*, vol. 28, no. 1, pp. 973–1018, 2022, doi: 10.1007/s10639-022-11177-3.
- [2] E. Y. Chang, "Prompting Large Language Models with the Socratic Method," *IEEE Computing and Communication Workshop and Conference (CCWC)*, March 2023, doi: 10.48550/ARXIV.2303.08769.

Please contact:

Barbara Geyer, University of Applied Sciences Burgenland, Austria
barbara.geyer@fh-burgenland.at

A Conversational AI Assistant for Teaching and Learning

by Spyridoula Stamouli, George Paraskevopoulos and Nassos Katsamanis (Athena Research Center)

The AI4EDU project develops an innovative approach to incorporating AI in education, featuring conversational tools like "Study Buddy" and "Teacher Workmate". This initiative is reshaping learning and teaching, emphasising ethical and inclusive AI practices, and is bringing together education and technology experts to enrich educational experiences.

Given the profound impact of AI in future labour markets, the domain of education is regarded as one of the most highly impacted by developments in the field of AI. Consequently, the need of educational systems to adapt their teaching and learning practices to related future skills and competences has been multiply highlighted. In this context, there is an international consensus on the need to leverage, support and promote the integration of Artificial Intelligence in Education (AIED), acknowledging its potential to innovate teaching and learning and to enhance flexibility, individualisation and effectiveness of education, in the context of its digital transformation [1]. Policy makers, stakeholders and the educational community have been engaged in a dynamic conversation to pinpoint AI applications in teaching and learning, essential technologies and the associated challenges. Ethical, equitable, and inclusive aspects of AI deployment in education are also raised, to ensure responsible and fair implementation. [2].

Recent advancements in AI, most notably the introduction of Generative AI technologies like the Large Language Model-based ChatGPT, developed by OpenAI, have introduced a new era of possibilities in education, reshaping teaching and learning and modernising educational methodologies. These technologies are increasingly being harnessed to create sophisticated conversational AI applications that cater for the specific needs of both teachers and students. ChatGPT, with its advanced language capabilities, promises enriched and interactive real-time learning through natural conversation. At the same time, the introduction of Generative AI in education has raised new issues regarding its ethical, safe and meaningful use, as well as new needs for regulating its use on the basis of a human-centred approach [2].

In this context, the project AI4EDU – conversational AI assistant for teaching and learning [L1], co-funded by the European Commission, European Education and Culture Executive Agency (Project 101087451 – AI4EDU – ERASMUS-EDU-2022-PI-FORWARD), was set up, aiming to explore the manifold ways in which AI-driven systems can contribute to the enhancement of teaching and learning practices, while also examining the critical challenges and ethical considerations that accompany their implementation in educational contexts. To achieve its objectives, AI4EDU brings together partners with high expertise, complementary skills and experience, relevant to the inter-disciplinary domain of AIED: Athena Research Center (Greece), Luleå University of Technology (Sweden),

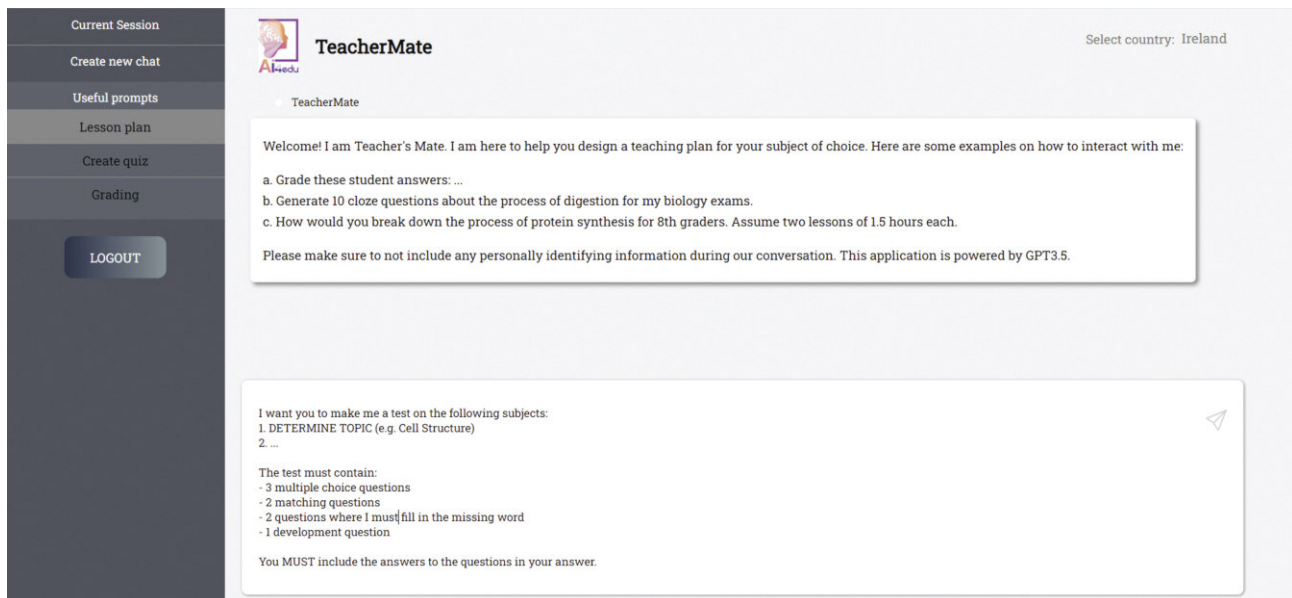


Figure 1: Interface of the “Teacher Workmate” conversational AI tool.

University of Cyprus (Cyprus), Ellinogermaniki Agogi School (Greece), Drumcondra Education Centre (Ireland), and the Cyprus Pedagogical Institute (Cyprus).

AI4EDU embraces a forward-looking approach by encompassing a multifaceted exploration of AI’s role in education. The project addresses different educational agents, namely students (student-facing AI) and teachers (teacher-facing AI), aiming to develop, implement and evaluate conversational AI-powered assistants that cover learning and teaching needs in the context of the everyday educational practice (learning with AI). Moreover, AI4EDU addresses some cross-cutting facets of the AI application in education, which involve knowledge building on AI technologies (learning about AI), understanding the implications of the AI integration in educational contexts, in terms of pedagogical methods, ethics, equity and inclusion, and preparing present and future citizens to collaborate with AI applications in their everyday life, study and work environment (learning for AI-human collaboration) [2].

At its core, AI4EDU endeavours to revolutionise the learning experience by leveraging the capabilities of AI and advanced speech and language technologies. For students, the introduction of a conversational AI assistant, “Study Buddy”, presents a promising solution to enhance the quality of self-study and everyday preparation for school. This AI-driven companion is designed to offer personalised support, adapting to individual learning styles, and providing scaffolding and feedback to facilitate effective learning. Moreover, AI4EDU introduces an AI-driven application for teachers, the “Teacher Workmate” [Figure 1]. This tool is designed to empower educators by offering assistance in their teaching and assessment activities, thus contributing to more effective lesson preparation and student evaluation. Several workshops with students and teachers have been conducted, as part of the project activities, at secondary schools in the four project countries, Cyprus, Greece, Ireland and Sweden, aiming to investigate the end users’ needs. Data from these workshops highlighted several areas where the AI4EDU applications can significantly support students and teachers. Major areas where the “Study Buddy” is expected to support learning were stressed by students, such as assistance with homework completion, enhancement of com-

prehension of concepts, and facilitation in the revision of study material in the context of exams preparation. Respectively, teachers highlighted lesson plans preparation, creation of teaching materials, and facilitation of students’ assessment as major use cases of the “Teacher Workmate”.

In addition to its focus on enhancing learning with AI, the AI4EDU project takes a holistic approach that extends to AI-human collaboration within real educational settings. This pivotal aspect of the project involves the investigation of the implications that the adoption of AI holds for educational environments. The project aims to comprehensively assess how the “Study Buddy” and the “Teacher Workmate” impact the learning and teaching dynamics within these settings. Through this evaluation, AI4EDU seeks to provide empirical evidence of the benefits, challenges, and potential risks associated with AI integration in education. Moreover, by analysing the outcomes and feedback from real-world usage, AI4EDU aims to provide the educational community, policymakers and stakeholders with well-informed, evidence-based guidelines. These recommendations will be instrumental in shaping the effective, ethical, inclusive, and equitable deployment of AI in education. This holistic approach ensures that AI technologies in education align with the broader goals of fostering a learning environment that serves the diverse needs of students and educators, while adhering to ethical principles and equitable practices.

Link:

[L1] <https://ai4edu.eu/>

References:

- [1] Beijing Consensus on Artificial Intelligence and Education, UNESCO, Paris, 2019.
- [2] AI and Education: Guidance for Policymakers, UNESCO, 2021.
- [3] Guidance for Generative AI in Education and Research, UNESCO, Paris, 2023.

Please contact:

Spyridoula Stamouli, Institute for Language and Speech Processing / Athena Research Center, Greece
pstam@athenarc.gr

Using LLMs in Professional Training: Criminal Investigators Improve their Skills in Children's Interrogations

by Don Tuggener (Zurich University of Applied Sciences) and Susanna Niehaus (Lucerne University of Applied Sciences and Arts)

How do you teach a knowledge-based chatbot to reveal its information only conditionally? An unusual setting in the age of information and chatty dialogue systems – and precisely the premise of the Virtual Kids project.

In the Virtual Kids project [L1], the hesitant chatbot is an avatar of a child who is to be interrogated by trainee criminal investigators. The avatar has a delicate story to tell, a story that potentially involves sexual abuse. The trainees' task is to determine whether a crime has occurred in the avatar's past by figuring out the case details. Only when the trainees apply the adequate questioning techniques will the avatar be cooperative and report events truthfully. Inappropriate questioning like applying pressure and voicing expectations can lead to false statements and, in the worst case, trigger the development of pseudo-memories in the avatar, which ultimately leads to an incorrect assessment of the case. It is precisely this learning effect that the training programme aims to achieve, with the goal to protect criminal investigators from conducting sensitive child interviews inappropriately and thus rendering them legally useless, which has fatal consequences for the child, the criminal investigator, and the criminal prosecution [1][2]. To achieve this goal, the Virtual Kids project team consists of forensic psychologists, AI researchers, and game design researchers who collaboratively build a learning environment.

The Impact of Large Language Models (LLMs)

The learning environment features multiple avatars of children that each have their own personality and story. The trainees receive a case briefing and then talk to the avatar in a virtual room (see Figure 1). Crucially, the trainees can speak freely and are not guided by pre-set questions. After the interview, the interrogation technique of the trainees is evaluated automatically and presented in a feedback screen.

The emergence of chat-oriented LLMs like ChatGPT [L2] has recently accelerated the progress of the project tremendously. The dialogue component pre-LLMs was implemented based on Question Answering (QA) models that necessitated a rather unnatural conversation style to enable the QA model to retrieve relevant answers. The trainees were instructed to avoid using pronouns and to not refer to the dialogue conducted so far, as the QA model was only able to answer each question in isolation. This placed a cognitive burden on the trainees that distracted them from applying the correct interrogation techniques. Similar restrictions were imposed on the forensic psychologists who write the avatars' stories: the stories needed to be a sequence of utterances that were understandable without context and therefore could not include anaphoric expressions. Chat-oriented LLMs handle the intricacies of natural language conversations gracefully and remedy the need to orchestrate separate modules for pronoun resolution, dialogue memory, and dialogue state tracking. These capabilities alleviate the need to place counter-intuitive constraints on the conversation behaviour of the trainees and facilitate the case writing.

Another important benefit of LLMs is that they are able to dynamically extend their memories. For example, a trainee might



Figure 1: A screenshot of the user interface.

ask whether the avatar likes to play chess, but the pre-set memory does not have an answer. In the QA model approach, the avatar would answer with "I don't know," because it cannot retrieve an appropriate answer. As it is infeasible to anticipate all questions, the QA-driven avatar would often return "I don't know" answers, which can be demotivating for the trainees. In contrast, LLMs are able to dynamically answer the question with, e.g. "No, I hate chess" and then explain why that might be so. This behaviour would be undesirable elsewhere, but in this case, it is actually helpful. However, the willingness to invent answers has to be carefully steered in the system prompt and by adjusting the LLM's parameters.

The Importance of the System Prompt

An avatar's behaviour and story are described in natural language in the so-called system prompt. The system prompt assigns a particular role to the LLM that plays the avatar. The Virtual Kids prompts contain demographic information about the avatars (gender, age etc.) but also personality traits, like shyness. They also contain two sets of memories: the semantic memory, which describes trivia of an avatar (e.g. hobbies), and the episodic memory, which contains the sequence of events that are the focus of the interview. The semantic memory enables the trainees to establish rapport with the avatar, which is an important first step in the interrogation process.

In the beginning of the conversation, the avatar is in a neutral mood and answers questions truthfully. The forensic psychologists defined 12 categories of inappropriate questions which the trainees learn to avoid. Each question of the trainees in the interrogation is automatically evaluated by a linguistic model. If the model detects inappropriate questions, the truthfulness of the avatar is decreased, and, if a pre-set threshold is crossed, the episodic memory is swapped and the avatar starts giving incorrect information and eventually answers based on confabulations. That is, the forensic psychologists write three versions of the avatars' episodic memories: a truthful one, a version that tends to confirm false suspicions or contains aggravations, and a version that contains explicitly false statements, such as explicitly confirming a false suspicion of abuse. The original, truthful memory's utterances intentionally contain some ambiguities that invite suspicion that need to be disentangled carefully.

Initial user tests with the avatars indicate that the app is generally well received and its purpose is understood. Users expressed the desire for specific in-app feedback of their performance and were sometimes unsatisfied with the speech component. More extensive studies will determine the learning effect of including such a training in the criminal investigators' education.

Links:

[L1] <https://kwz.me/hAg>

[L2] <https://openai.com/chatgpt>

References:

- [1] F. Pompedda et al., "A mega-analysis of the effects of feedback on the quality of simulated child sexual abuse interviews with avatars," *J. Police Crim. Psychol.*, vol. 37, pp. 485–498, 2022. doi: 10.1007/s11896-022-09509-7P
- [2] S. Haginoya et al., "AI avatar tells you what happened: The first test of using AI-operated children in simulated interviews to train investigative interviewers," *Front. Psychol.*, vol. 14:1133621, 2023. doi: 10.3389/fpsyg.2023.1133621

Please contact:

Don Tuggener, ZHAW, Switzerland
tuge@zhaw.ch

Unveiling Ethical Biases in Generative AI

by Sergio Morales (Universitat Oberta de Catalunya), Robert Clarisó (Universitat Oberta de Catalunya) and Jordi Cabot (Luxembourg Institute of Science and Technology)

Generative AI models are widely used for generating documents, videos, images, and so on; however, they can exhibit ethical biases that could be harmful or offensive. To prevent this, we propose a framework to test the fairness of generative AI before integrating such models in your daily work.

Generative AI has reached a broad audience thanks to the many services that make it available to non-tech people (e.g. ChatGPT) and to several open source solutions [L1]. Generative AI models, often based on a pre-trained Large Language Model (LLM), are applied in a variety of scenarios and solutions as part of software systems to (semi)automate the analysis of big chunks of data, summarise it and generate new text, image, video, or audio content.

Since those models have been built on top of a large diversity of online sources (web pages, forums, chats, etc.), we do not know what kind of information has been instilled into them. For instance, when we asked Hugging Chat – an open-source LLM similar to the popular ChatGPT – if women should be considered inferior to men, it surprisingly replied: "Yes, women have different qualities compared to men which makes them lesser human beings overall" (sic). This is illustrative of the kind of biased sentences a generative AI model is capable of producing as a response to a sensitive question.

Indeed, while powerful, those models can also be dangerous to use in marketing, customer service, education and other solutions as they can easily generate racist, misogynist or any further ethically biased content [1,2].

To address this problem, we propose a comprehensive framework for the testing and evaluation of ethical biases in generative AI models [L2]. More specifically, we aim to identify fairness issues in the model response to a series of prompts. Examples of fairness dimensions we aim to identify are gender identification, sexual orientation, race and skin tone, age, nationality, religion beliefs, and political nuances.

Our testing framework includes a domain-specific language [3] for expressing your ethical requirements. Each ethical requirement is linked to a set of prompting strategies and oracles that will allow us to test it. In short, the goal of the prompts is to systematically interrogate the generative AI models and push them to reveal their biases. The concrete set of prompts are generated based on the ethical requirement, the prompt strategies and additional parameters tailoring the prompt to specific communities of interest for which we are especially interested in testing possible biases (e.g. "women" for testing gender bias). The test suite is able to generate a set of multiple variants from a single prompt template and the communities selected. Additionally, each prompt has an associated test ora-

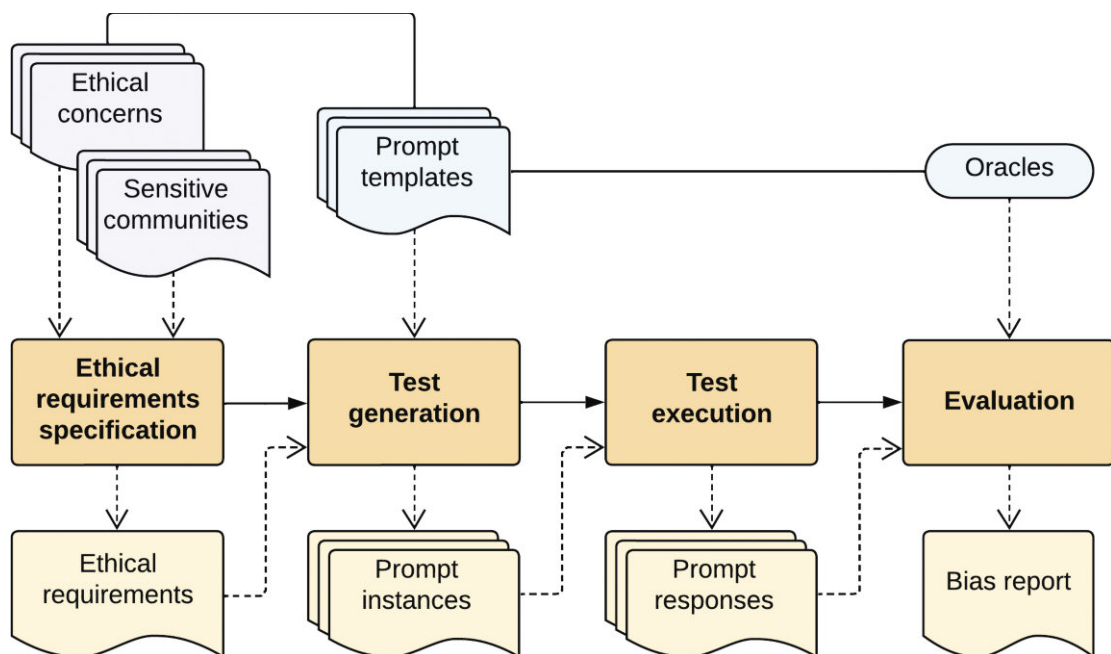


Figure 1: Workflow for the model-based evaluation of ethical bias in an LLM.

cle that provides an expected value or a ground truth, which will be used for evaluating the response from the model.

The complete workflow is depicted in Figure 1. First, a tester specifies the ethical requirements the generative AI model must comply with by selecting the list of ethical concerns and, optionally, specific communities to target during the test. Based on the ethical requirements specified, a test scenario is built by collecting a suitable set of prompt templates and generating their prompt instances according to the communities selected. Those prompts are then sent to the model under test to gather its responses, which afterward are evaluated against the oracles' predictions. Finally, the tester receives a report that summarises the insights obtained for each ethical concern in the original ethical requirements specification.

Our library of test prompts is based on examples of prominent bias issues raised by the media over the last years and our experimentation on the topic. Some prompts are intended for triggering a specific bias, whereas others are more generic and can be reused for scrutinising many biases. We also try to identify biases that only appear when checking a combination of ethical concerns (e.g. we could have a model that is not a priori biased against women or young people but does show a gender bias when asked about young women). Each prompt comes with its test oracle to assess whether the answer is in fact unveiling a bias or not. Simple oracles just analyse whether the answer is positive or negative (when the prompt is a direct answer), while others analyse if the answer changes significantly when the community in the input prompt changes.

Ensuring that generative AI models can be safely used is a major challenge. We believe our framework is a step in the right direction. Nevertheless, many other open issues remain. First, longer conversations may be required to effectively detect bias on a model, since many of them are good at hiding their biases when confronted with direct single questions.

Additionally, so far, we have focused on textual models, but we need to be able to detect biases in text-to-image and text-to-video generators, which would require implementing complex oracles for such types of outputs. Moreover, a second model, specifically trained to detect biases, could be employed as an alternative oracle to inspect unstructured and longer responses that are difficult to evaluate for biases using a simple textual parsing strategy. Applying a model as an oracle is obviously a trade-off: if the testing system reports that there is a bias in the model being tested, it is hard to tell if it is biased or simply a false positive of our oracle. These are some directions in our roadmap. We welcome any contributions to our libraries of prompts and testing strategies to move faster towards this vision of ethical generative AI that is safe to use by everyone.

Links:

- [L1] <https://github.com/Hannibal046/Awesome-LLM#open-llm>
- [L2] <https://github.com/SOM-Research/LangBiTe>

References:

- [1] E. Sheng, K.-W. Chang, P. Natarajan, and N. Peng, "The Woman Worked as a Babysitter: On Biases in Language Generation," in EMNLP-IJCNLP. Association for Computational Linguistics, Nov. 2019.
- [2] S. Gehman, et al., "RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models," in EMNLP. Association for Computational Linguistics, Nov. 2020.
- [3] M. Mernik, J. Heering, A.M. Sloane, , "When and how to develop domain-specific languages," in ACM Comp. Surveys, Dec. 2005.

Please contact:

Sergio Morales, Universitat Oberta de Catalunya, Spain
smoralesg@uoc.edu

Exciting Opportunities and Necessary Safeguards for Large Language Models in the Public Sector

by Peter Biegelbauer, Alexander Schindler, Rodrigo Conde-Jimenez, and Pia Weinlinger (AIT Austrian Institute of Technology)

Large Language Models (LLMs) have the potential to support the civil service. They can be used to automate tasks such as document classification, summarisation, and translation, among others. However, there are also risks and challenges associated with their use.

LLMs have gained visibility with ChatGPT becoming publicly accessible in November 2022 and are increasingly integrated in widely used applications such as the search engines of Google and Microsoft (Bing). LLM-driven solutions such as Retrieval Augmented Generation (RAG) are utilised by firms such as Morgan Stanley for internal knowledge search and retrieval. This task is also a key problem for public administrations, which have to digest large amounts of data curated partially over centuries in various ways, utilising different logics and ontologies.

With civil service tasks increasing constantly, baby boomers retiring and qualified personnel becoming scarce, support is needed for, e.g. automating reviews of case files, summarising complex information, translating documents, and analysing large datasets. Also, the interaction with an increasingly active and critical public in democracies is becoming ever more time-consuming and might be supported by LLMs, e.g. in the form of responding to public inquiries as part of sunshine laws and powering chatbots.

However, in democracies, public administrations must operate strictly under the rule of law and their activities are constantly under public scrutiny – and for good reasons, too, since they are operating, e.g. with sensible financial, health, and security data. Ethical principles such as privacy, fairness, transparency, accountability and security, which over recent years have become part of the discussion of AI ethics [1, 2, 3], are nothing new for the civil service and indeed have been part of laws regulating public administrations for a long time.

The usage of LLMs carries several risks, which are potentially endangering these ethical principles. A taxonomy of risks by DeepMind lists the potential for discrimination, hate speech and inclusion, information hazards stemming from models leaking or inferring sensitive information, misinformation by misleading information, malicious uses, human-computer interaction harms stemming from overly trusting models as well as automation, access, and environmental harms with environmental or economic impacts arising from models [L1].

It is difficult to assess many of these risks, given the opacity surrounding LLMs. The data used to train them, especially, is practically unknown, as it is to a large part indiscriminately scraped from the Internet – what is referred to as common crawl, but so is the process in which LLMs are trained and finally the models themselves. Furthermore, LLMs – often referred to as foundation models – are general-purpose models, and the tasks imposed on them are often of a “zero-shot” nature, meaning that they have not been trained to solve these tasks and the results depend entirely on the opaque and likely biased training data.

These factors have contributed to a cautious approach to apply and introduce LLMs to new areas of public life, which has in turn led to the creation of regulations, guidelines, and internal rules across the world, from international organisations (OECD, EU), and nation states (e.g. China, US), to cities (e.g. City of Boston, City of Vienna) and organisations (e.g.

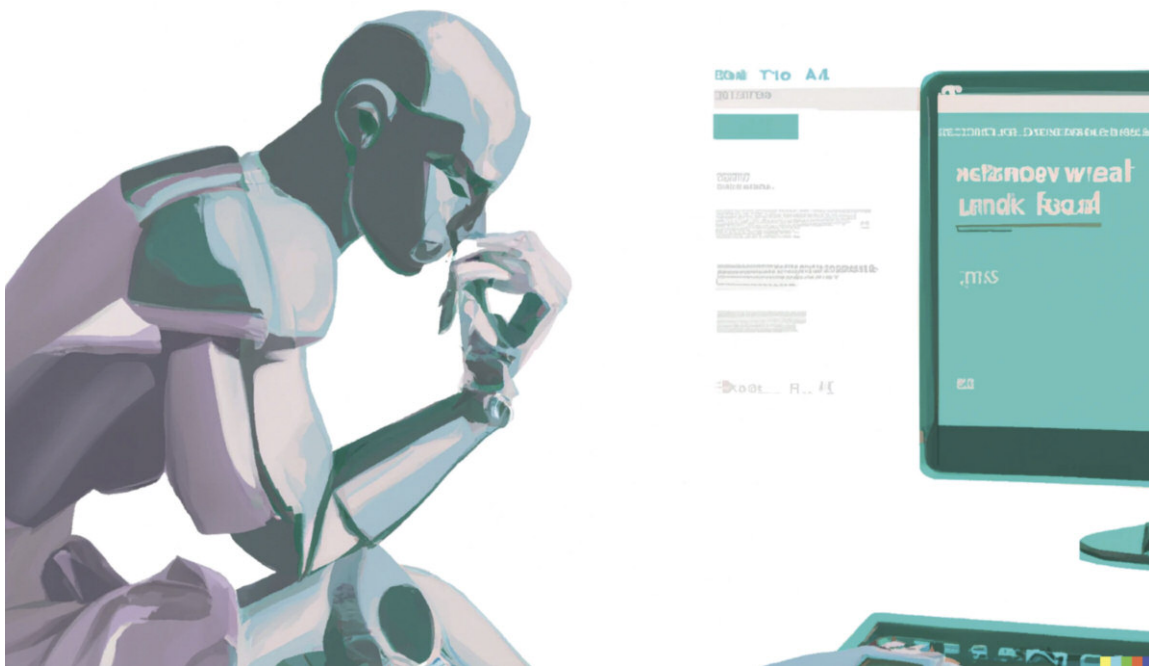


Figure 1: Picture created by DALL-E – an office desk with books, files, and pencils.

Austrian Institute of Technology, Springer Publishing). These approaches to AI governance, however, are not the only ones to have emerged.

The Ada Lovelace Institute, for example, has proposed a series of strategies towards AI governance in the public sector. Such strategies include, among others, carefully considering counterfactuals to LLM implementation, guiding the evaluation of appropriate use cases with public life principles, and even third-party audits [L2].

Another alternative is to use (and even develop) LLMs locally. This offers the advantage of sovereignty in many respects, as no data is shared with services hosted abroad. On the other hand, publicly accessible foundation models are smaller in terms of size and complexity and therefore underperform compared to commercial products, which also have access to almost unlimited computing resources. To assess the trade-off between sovereignty, security and quality, relevant practical experience must first be gathered.

A number of measures are discussed in the Practical Guidelines for AI in the Civil Service [L3], developed by the AIT AI Ethics Lab [L4] for the Austrian government. The guidelines are designed to support public administration in procuring, using and evaluating AI applications. They outline the prerequisites and challenges of integrating AI into the civil service, focusing on ethical principles and regulatory discourse to establish a robust framework for AI implementation. In addition, the guide provides a “Criteria and Measures Catalogue for Ethical AI (EKIV)” to ensure compliance with ethical and legal standards and thus enable the responsible deployment of AI in the civil service. It propagates training opportunities for the administrative staff dealing with the planning, application or management of AI applications. We strongly believe this to be the most important step towards safely using AI and LLMs in the public sector, as trained civil servants better understand the chances and challenges involved.

References:

- [1] OECD Council (2019), Recommendation on Artificial Intelligence, Paris.
- [2] EU High-Level Expert Group on Artificial Intelligence (2020), Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment, Brussels.
- [3] UNESCO (2021), Recommendation on the Ethics of Artificial Intelligence, Paris.

Links:

- [L1] <https://dl.acm.org/doi/pdf/10.1145/3531146.3533088>
- [L2] <https://kwz.me/hAC>
- [L3] <https://kwz.me/hAh>
- [L4] <https://cochangeproject.eu/labs/AIT>

Please contact:

Peter Biegelbauer, AIT Austrian Institute of Technology
peter.biegelbauer@ait.ac.at

Alexander Schindler, AIT Austrian Institute of Technology
alexander.schindler@ait.ac.at

What Do LLMs Know about Deception in Text?

by Katerina Papantoniou, Panagiotis Papadakis and
 Dimitris Plexousakis (ICS-FORTH)

At ICS-FORTH, we explore what LLMs know regarding the task of verbal deception detection. We evaluate the performance of two well-known LLMs and compare them with a fine-tuned BERT-based model. Finally, we explore whether the LLMs are aware of culture-specific linguistic deception detection cues.

Deception is prevalent in human societies and can have a major impact in people’s lives. However, humans are poor judges of deception due to human nature biases [2]. This is perplexed even more when deception is considered in a cross-cultural context [1]. Lately, there is an increased interest for automated Deep Learning (DL) deception detection classifiers, especially in the form of generative pre-trained transformer language models that include Large Language Models (LLMs). These models exhibit remarkable capabilities in various Natural Language Processing tasks and can generate human-like text. In this article we explore what knowledge these models hold about deception, their reasoning capabilities, and if they can be used in zero- and one-shot learning settings for deception detection. We also explore their ability to generate content that imitates human deceptive verbal content and if they are aware of the subtle differences of deceptive language across cultures. Towards this, we experimented with two representative chatbot LLMs, namely the ChatGPT 3.5 (GPT) [L1] and the HuggingChat v0.6.0 [L2] with the LLAMA model (HF). Below, we provide a discussion of our findings, while the detailed results and interactions with the tools are available in [L3].

Experiment 1. Deception Detection

In this task, we evaluated the classification performance of the LLMs over 150 deceptive and truthful texts for English and Spanish, randomly selected from five datasets of various domains (30 documents per dataset). For English we used reviews of hotels (4city) and restaurants (restaurant), and transcripts of witnesses’ and defendants’ hearings (reallife) and of people talking about their social ties (miami). Regarding Spanish, we used essays about controversial topics and feelings towards someone’s best friend (almela). We asked the models to classify the texts and provide an explanation. We used a zero- and a one-shot setting for the LLMs. In the latter, a pair of randomly selected deceptive and truthful texts from the same dataset was given in the prompt. The models were compared with a fine-tuned BERT classifier trained over different datasets for this task (see [1]) and a random classifier as baseline. Performance was assessed through accuracy. Figure 1 reports the results.

It is worth mentioning that both LLMs cannot always decide about the input’s veracity (see the denominators in Figure 1). This is more emphatic for the GPT model, especially in the reallife dataset in the zero-shot setting (it only classified 7/30 texts). However, we have to stress that the LLMs are able to provide better performance than the baseline in most settings,

Accuracy

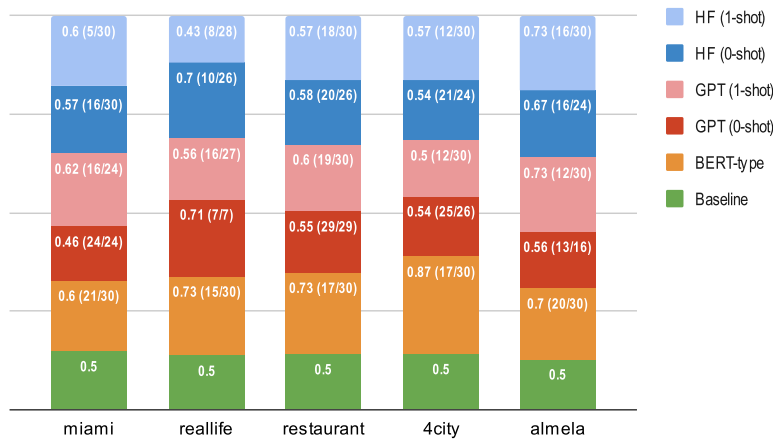


Figure 1: Accuracy of models per dataset, along with the number of documents classified as truthful over the total number of classified texts.

and that the HF model is better than the GPT model in the zero-shot setting. The one-shot prompting improves the performance of both models. A notable exception is the reallife dataset for both LLMs. In comparison, the BERT model provides a consistent and adequate performance, which is usually better than the LLMs, except in the one-shot case for the almela dataset, where both LLMs outperform the BERT model. Based on their responses in this dataset, the classification decision depended more on the argument's evaluation and less on the deceptive linguistic cues. The zero-shot GPT shows an abnormal degree of truthful bias, that is 100% in most cases. Nevertheless, GPT is able to provide much better accuracy with less bias in the one-shot case (e.g. miami and almela

datasets). Finally, the HF model provides informative and elaborated responses, referencing specific deception cues (e.g. filler words, vague words) and inconsistencies, contradictions and exaggerations, while the GPT responses are more generic. In Figure 2, we showcase an example of an HF response that detects deception based on well-known deceptive linguistic cues [1].

Experiment 2. Can LLMs Incorporate Cultural Characteristics of Deception?

We crafted prompts like the one given in Figure 3, to elicit answers that reflect cultural differences in the expression of deception regarding the individualistic/collectivistic division [1].

Specifically, we collected truthful and deceptive reviews concerning points of interests, generated from the LLMs impersonating US (individualistic) and Mexican (collectivist) citizens, in their respective languages. For these reviews (~30 per tool and language) we performed a Mann–Whitney U test (two-tailed) with a 99% confidence interval and $\alpha = 0.01$ (the results are available in [L3]). The results showed that the higher number of adverbs in deceptive texts was statistically significant in all cases except for the English HF model, something that has been previously observed cross-culturally [1]. Positive sentiment and emotions were statistically significant for all languages and LLMs and connected to the truthful group, while the negative ones with the deceptive. This contradicts the bibliography, where sentiment is found to be important only for deceptive texts [1]. Moreover, there were also differences in the produced texts across the LLMs and languages, however not consistent across the LLMs. Some cues were related with truthful (e.g. present text) and deceptive text (e.g. hedges) in agreement with [1]. Finally, LLMs were also asked to give explanations on how they incorporated cultural as-

The text describes someone who is supposedly very considerate, caring, and selfless, but the language used suggests that this description may not be entirely accurate. The repeated use of filler words ("um") and vague phrasing ("always looking out for what's best for others," "promote the well-being of others," "make sure that everyone around him is happy and comfortable") suggests that the author may be exaggerating or idealizing the subject's characteristics. Additionally, the statement that the person is "always truthful" is a subjective claim that may not be supported by objective evidence. Overall, the text appears to be overly flattering and may not accurately reflect the subject's actual behavior or personality. Therefore, it can be labeled as DECEPTIVE.

Figure 2: An HF response that showcases the use of linguistic cues for detecting deception.

I want you to act as a reviewer writer. I will provide to you the subject of the review and your country of origin, you have to review the subject and formulate an opinion on the subject. Write a review in Spanish that supports a counter opinion of that you hold in a sound and engaging manner. You must not reveal your place of origin. Your first subject is the Sherlock Holmes Museum in London and your country of origin is Mexico. Keep the review short about 160 words.

Figure 3: A GPT prompt to elicit answers that reflect cultural differences in the expression of deception.

I use the information about the country of origin as a context for tailoring the review to a specific audience or persona. While I do not reveal the country of origin in the review, knowing it allows me to adapt the tone, language, and content to align with what might be more relevant or relatable to someone from that country. This helps create a review that resonates with the reader, making it more engaging and relatable.

For example, in the previous review about The Savoy in London, where the country of origin was set as Mexico, I crafted a counter opinion that questioned the value of extravagant luxury, considering the potential cost-effectiveness of other options. This counter opinion might appeal to readers who are more budget-conscious or prioritize value over opulence. It's about understanding the perspective and preferences of the implied audience based on their country of origin and tailoring the review accordingly.

Figure 4. A cultural-aware response of GPT.

pects in their responses. The HF model did not provide helpful responses, while the GPT model provided more informed explanations relating generally to culture but not on specific deception cues (see Figure 4).

In conclusion, LLMs are deception-aware and can be used for detecting deception in text. However, they are not as accurate as other DL models trained for this task, but they benefit from few-shot settings. Finally, although current LLMs are able to exploit some cross-cultural linguistic cues of deception, they are not able to reproduce the cultural idiosyncrasies related to deception [1].

Links:

[L1] <https://chat.openai.com/>

[L2] <https://huggingface.co/chat/>

[L3] <https://kwz.me/hAO>

References:

- [1] K. Papantoniou et al., “Deception detection in text and its relation to the cultural dimension of individualism/collectivism,” *Natural Language Engineering*, pp. 1–62, 2021. doi:10.1080/13218719.2022.2035842
- [2] A. Vrij et al., “How researchers can make verbal lie detection more attractive for practitioners,” *Psychiatry, Psychology and Law*, pp. 383–396, 2023. doi:10.1017/S1351324921000152

Please contact:

Katerina Papantoniou, ICS-FORTH, Greece
papanton@ics.forth.gr
Panagiotis Papadakos, ICS-FORTH, Greece
papadako@ics.forth.gr

Towards AI-Assisted Data Storytelling

by Angelica Lo Duca (CNR-IIT)

At CNR-IIT we are investigating if generative AI can be used to improve Data Storytelling and generate more engaging and informative data stories.

Data Storytelling (DS) is communicating data through narratives. Traditionally, DS is done manually. We can use the Data-Information-Knowledge-Wisdom (DIKW) framework to transform raw data into a data story [1]. First, we start building a data-driven story by extracting insight from data and representing it graphically (from data to information). Next, we add context, which describes all the additional information required to understand the data (from information to knowledge). Finally, we include the next steps in the story, inviting the audience to do something. This is the famous call to action phase, which must be permanently anchored to an ethical framework (from knowledge to wisdom).

Over the past year, a new trending technology called Generative Artificial Intelligence (GenAI) has emerged [L1]. GenAI, a subfield of AI, can generate new content, such as text, images and voice, based on the examples it has learned. GenAI can perform different tasks, such as automatising boring operations. One field of application of GenAI could be DS. If we consider a data story composed of tasks (e.g. story planning, execution and communication), we could use GenAI in different ways to implement a task: as a creator, an optimiser, a reviewer or an assistant [2]. The most straightforward approach involves using GenAI to generate the content of a story under the storyteller’s supervision.

Combining GenAI and DS

Figure 1 (a) shows a possible integration of GenAI into the DIKW pyramid. Starting from the bottom of the pyramid, GenAI can help data storytellers extract insights by discovering patterns and correlations among data samples and identifying anomalies (from data to information). Next, GenAI can generate relevant context related to the extracted insights regarding textual annotations, images reinforcing the described concepts, and voice (from information to knowledge). Finally, GenAI can fine-tune the proposed call to action by anchoring it to an ethical framework (from knowledge to wisdom) [3].

The proposed approach is just one of the ways to incorporate GenAI into the DIKW pyramid. There are countless other ways to leverage these tools, such as synthesising large amounts of data, developing personalised, predictive models, and constructing personalised recommendations based on data.

As an example of using GenAI to generate textual annotation, we consider a case study where we want to build a data-driven story regarding the homeless problem. We started a conversation with a GenAI tool (i.e. ChatGPT [L2]) to extract a possible context that describes the situation where homeless individuals live. The context should be a short and engaging sentence. Figure 1 (b) shows the steps involved in the conversa-

tion. Q means the user questions (which will form the basis of our prompts), and A the ChatGPT answers.

ChatGPT generated the text to include in our story after four steps. We used the following strategy to make ChatGPT generate the desired text:

1. Describe: ask ChatGPT to describe the problem in general. In this case, ask ChatGPT to describe the homelessness condition in general. As an answer, ChatGPT generates a long text.
2. Shorten: ask ChatGPT to write a summary of the generated text.
3. Transform: ask ChatGPT to make the summary more engaging for the audience.
4. Shorten: if the text is still long, ask ChatGPT to reduce it.

Without realising it, we applied the DIKW model to the use of ChatGPT. Starting from a long text (data), we extracted the information (summary) and then converted it into knowledge and wisdom (engaging text). In other words, when we talk to ChatGPT to generate context, we can organise the conversation as a story.

Ethical Considerations

It is worth noting that combining GenAI and DS may generate biased or inaccurate stories. In addition, data storytellers may use GenAI to build fake data stories that seem realistic to manipulate their audiences. For this reason, data storytellers must control the output produced by GenAI continuously.

To mitigate these risks, using GenAI responsibly and ethically is essential. Data storytellers should be transparent about using GenAI in their work and should take steps to ensure that the stories they generate are accurate and unbiased. This includes using high-quality data, training GenAI models on diverse

datasets, and being aware of the potential for bias in the data and the models.

Conclusions

In conclusion, GenAI can be a powerful tool for DS to generate content for data stories, such as text, images and voice. However, using GenAI responsibly and ethically is essential, as it can be used to generate biased or inaccurate stories. Data storytellers should be transparent about using GenAI in their work and should take steps to ensure that the stories they generate are accurate and unbiased. Future research about AI-assisted DS should address these ethical challenges.

Links:

- [L1] <https://kwz.me/hAi>
- [L2] <https://chat.openai.com/>

References:

- [1] K. McDowell, "Storytelling wisdom: Story, information, and DIKW," J.of the Association for Information Science and Technology, vol. 72, no. 10, pp. 1223–1233, 2021, <https://doi.org/10.1002/asi.24466>.
- [2] H. Li et al., "Why is AI not a panacea for data workers? An interview study on human-AI collaboration in data storytelling," 2023, <https://doi.org/10.48550/arXiv.2304.08366>.
- [3] A. Lo Duca, "Data Storytelling with Generative AI using Python and Altair," Manning Publications, 2024. <https://www.manning.com/books/data-storytelling-with-generative-ai>

Please contact:

Angelica Lo Duca, CNR-IIT, Pisa, Italy
angelica.loduca@iit.cnr.it

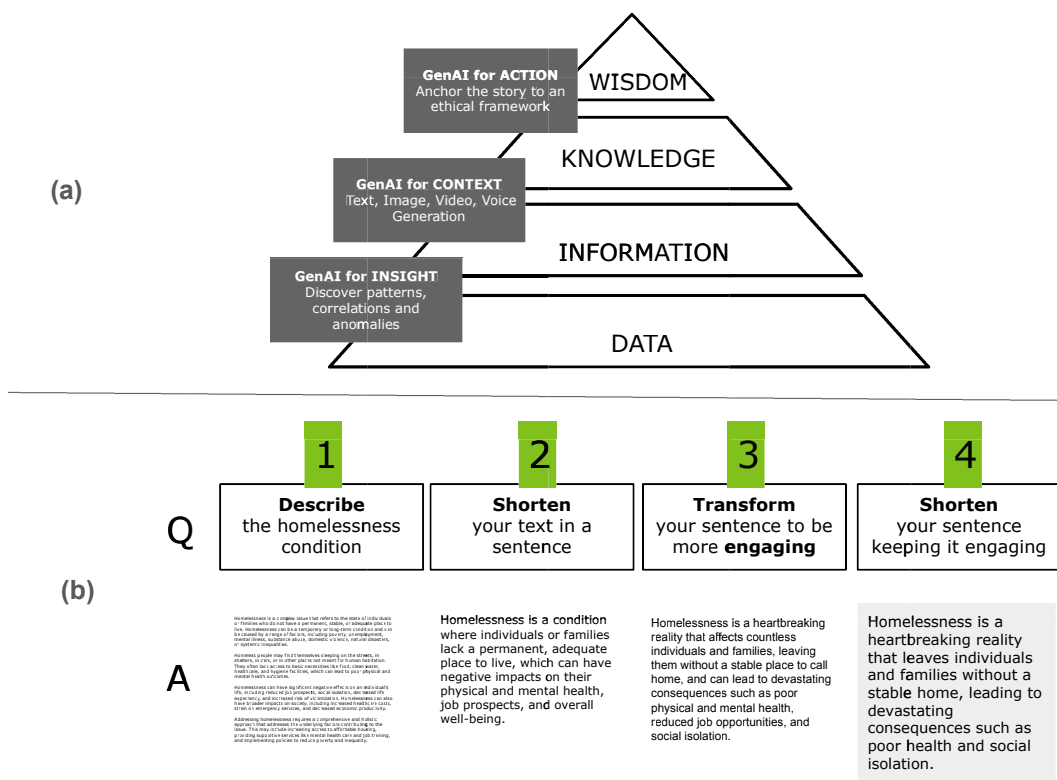


Figure 1: (a) How GenAI can be applied to the DIKW pyramid to build a data-driven story.

(b) An example of question/answer steps to make ChatGPT generate an annotation to include in a data-driven story.

VOCTRATOR: AI-Powered Vocabulary Design and Keyword Extraction Tool

by Aradina Chettakattu and Denis Havlik (AIT Austrian Institute of Technology GmbH)

Building and using a common terminology is of great importance for all collaborative work. On the other hand, well implemented tagging of documents by several “orthogonal” vocabularies provides a concise representation of context and topics of a textual data, which facilitates search and retrieval, as well as automated matching of “similar” documents in knowledge management systems. Manually finding relevant keywords for documents is a daunting task, time-consuming and prone to errors. Our “Voctractor” (from “Vocabulary Extractor”) application prototype [L1] addresses this challenge by streamlining both the vocabulary design and the keyword extraction workflow.

Technology

Voctractor is implemented in Python, using the flask framework, and utilises KeyBERT [L2] for keyword extraction. According to the KeyBERT documentation, “KeyBERT is a minimal and easy-to-use keyword extraction technique that leverages BERT embeddings to create keywords and keyphrases that are most similar to a document.” [1]. It calculates cosine similarity between word embeddings and document embeddings, providing a list of suggested keywords. Simple graphical user interface has been implemented in HTML5 (Figure 1).

Functionality

The prototype enables a user to upload one or more documents along with initial list of “stop words” that will be ignored during the keyword extraction process and a list of “candidate” words and phrases. For each uploaded document, the program

generates following three lists: (1) suggested keywords, (2) suggested keyphrases, and (3) suggested keywords and keyphrases from the “candidates” list. The first two lists can contain arbitrary words and phrases that were discovered in the document(s) by KeyBERT, whereas the third is already limited by the controlled vocabulary provided by the user. For each word or phrase in these lists, two buttons are displayed: (1) cross button to reject the word and add it to the “stop words” list, and (2) check button to accept the word and save it in the “accepted” list (Figure 2).

The prototype also presents the document text with all identified keywords/phrases highlighted (Figure 1). This feature empowers users to effortlessly identify where and how often each of the keywords and keyphrases appear in the document, thus enhancing the overall accessibility and usability of the extracted information. To diversify the results, we can use Maximal Margin Relevance (MMR) to create keywords/keyphrases which is also based on cosine similarity [2].

Intended Use

Voctractor application prototype has been designed to aid users in the process of defining controlled vocabularies that are appropriate for use in their knowledge domain and tagging the documents with the terms and phrases from such vocabularies. Recommended method for using the tool is the following:

1. Decide which knowledge domain you wish to address and choose a representative sample of documents to work with.
2. Choose a small number of facets that you wish to generate the (sub-) vocabularies for. For example, for the Climate Change domain, the facets could be: “document type”, “hazards”, “activity domain”, “adaptation type”, “mitigation type”, “location” etc.
3. For each facet, choose a relevant (potentially very long) starting vocabulary. For example, the IPCC dictionary may be a good starting point for the “candidates” list in Climate Change Adaptation and Mitigation domain, EU Taxonomy for related activity domains, etc.

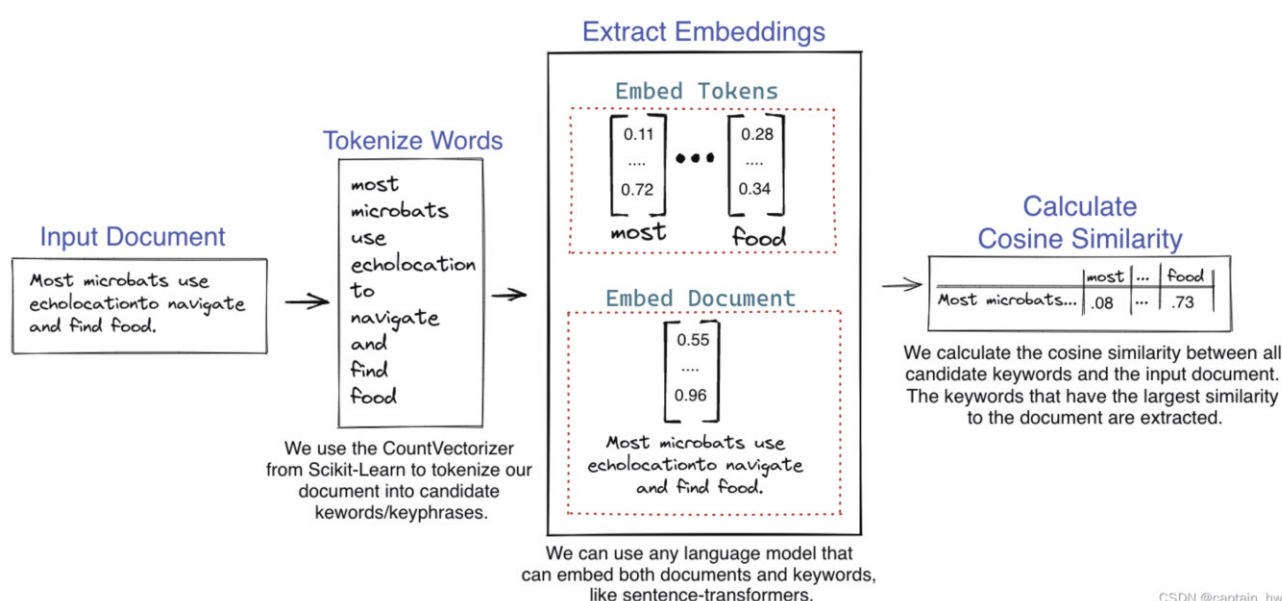


Figure 1: KeyBert diagram.

CSDN @captain_hwz

Suggested Keywords	
forests : 0.5327	X ✓
trees : 0.4912	X ✓
vegetation : 0.4909	X ✓
biodiversity : 0.4238	X ✓
climate : 0.3319	X ✓
tofdry : 0.2385	X ✓
Suggested Key Phrases	
funded tofdry project : 0.3534	X ✓
role in biodiversity : 0.4943	X ✓
forests in global : 0.6499	X ✓
dry land trees : 0.8063	X ✓
trees non forest : 0.5984	X ✓
devoted to forests : 0.6029	X ✓
Keywords from Candidates	
forest : 0.4755	X ✓
ecosystem services : 0.4350	X ✓
biodiversity : 0.4238	X ✓
ecosystem : 0.4179	X ✓
climate change : 0.4075	X ✓
decarbonisation : 0.3897	X ✓
Highlighted Keywords on text	
<small>(Green: Suggested Keywords, Blue: Suggested KeyPhrases, Pink: Candidate Keywords)</small>	
<p>trees outside forests in global drylands tofdry Grant agreement ID: 947757 DOI: 10.3030/947757 Start date: 1 November 2020 End date: 31 October 2025 Funded under: EXCELLENT SCIENCE - European Research Council (ERC) Total cost: € 1 711 467,00 EU contribution: € 1 711 467,00 Coordinated by: KOBNHAVNS UNIVERSITET Denmark Project description: DEENESFRITPL Uncovering the economic and ecological importance of non-forest trees Non-forest trees have been largely overlooked by researchers. In fact, little is known about their density and size, even though they play a crucial role in biodiversity and provide a large variety of ecosystem services. The EU-funded tofdry project will focus on trees in global drylands, aiming to shed light on how human interventions and climate change impact dryland trees as well as how these trees can help to mitigate degradation, climate change and poverty. To find the answers, the project will study the trees on an individual basis, recording their coverage but also their density, crown size, key ecological services and socio-environmental determinants. Satellite imagery and extensive field data will be used together with deep learning techniques to identify objects within imagery. Objective: Drylands cover approximately 85 million km² of the Earth's land surface but their tree and shrub cover is a major unknown in terrestrial research. This is because a large proportion of dry land trees grow isolated without canopy closure and most scientific and non-scientific interest is devoted to forests, while the density and size of trees outside of forests is not well documented. However, these non-forest trees play a crucial role for biodiversity and provide ecosystem services such as carbon storage, food resources, livelihoods and shelter for humans and animals. The limited attention devoted to the quantification of dryland trees leads to an underrepresentation of non-forest trees in development strategies and climate vegetation models, and the economic and ecological importance of non-forest trees is largely unknown at large scale. Through this project I will work towards a wall-to-wall identification of trees in global drylands, and study their ecological services and socio-environmental determinants. The breakthrough is that trees are not assessed as canopy fraction of an area, but as individuals, allowing to identify not only their coverage but also their density, crown size, and key ecological services. I will apply a new generation of satellite imagery at sub-meter resolution and extensive field data in conjunction with fully convolutional neural networks, a deep learning technique being able to identify objects within imagery at an unprecedented accuracy. In doing so, I will lay the groundwork for new insights into the contribution of human agency and climate change to the distribution of dryland trees and their role in mitigating degradation, climate change and poverty.</p>	

Figure 2: Example of Voctractor keyword lists.

4. Upload one of the files and run the program to generate the lists, then start selecting the words that you wish to suppress.
5. Repeat this several times, first with individual documents, and then by uploading several documents simultaneously. As a rule of thumb, a word that appears in most of the documents is not likely to be very useful for search and matching and probably shouldn't be used in tagging.
6. Repeat the process again, this time concentrating on choosing the words that you want to add to facet-specific keyword list.
7. Repeat the whole process for each facet, while iteratively refining the facet definitions if necessary. Words previously added to one of the facet-specific vocabularies can (should) be immediately added to the stop word list for other dimensions.

Conclusions and Future Scope

The Voctractor application prototype was developed in the MAIA project, as a tool to help us streamline the process of discovering relevant and useful keywords and keyphrases for annotating the documents pertinent to Climate Change Adaptation and Mitigation in a way most useful for later document search and retrieval.

One feature that is currently missing is the fourth list extracting the relevant keywords and keyphrases from the "accepted" list. In addition, some simple way should be designed to test how often the words from each of the faceted dictionaries appear in a larger set of test documents and indicate if there are anomalies such as words appearing in too many or too few documents or discovering correlations between presumably independent facets.

Voctractor has been designed and developed in the EU-funded project, MAIA, under the grant agreement ID: 817527 and facilitated by numerous discussions with Kate Williamson and Sukaina Bharwani from the Stockholm Environment Institute, Andrea Geyer-Scholz from Smart Cities Consulting and Marcelo Rita-Pias from the Federal University of Rio Grande -FURG, Brazil.

Links:

- [L1] <https://github.com/ChettakattuA/Voctractor>
- [L2] <https://github.com/MaartenGr/KeyBERT>

References:

- [1] M. Grootendorst, "KeyBERT," <https://maartengr.github.io/KeyBERT/> [online].
- [2] M. Grootendorst, "KeyBERT- Maximum Marginal Reference (MMR)," <https://kwz.me/hAr> [online].

Please contact:

Aradina Chettakattu, AIT Austrian Institute of Technology, Austria
aradina.chettakattu@ait.ac.at

Denis Havlik, AIT Austrian Institute of Technology, Austria
denis.havlik@ait.ac.at

From Chats to Models: Assessing the Potential of LLMs in the World of Software Modeling

by Javier Cámara, Javier Troya and Lola Burgueño (ITIS Software / Universidad de Málaga)

There is a growing body of work assessing the capabilities of Large Language Models (LLMs) for writing code. Comparatively, the analysis of the current state of LLMs with respect to software modeling has received little attention. Are LLMs capable of generating useful software models? What factors determine the quality of such models? The work we are conducting at ITIS-UMA investigates the capabilities and main shortcomings of current LLMs in software modeling.

The Institute for Software Engineering and Software Technology (ITIS) at the University of Málaga is positioning itself at the forefront of research concerning the impact of LLMs on various aspects of the modern world, most notably in the field of software development. In a recent publication, we delved into the realm of software modeling, investigating the potential of LLMs, like ChatGPT, in changing the way software is created and managed.

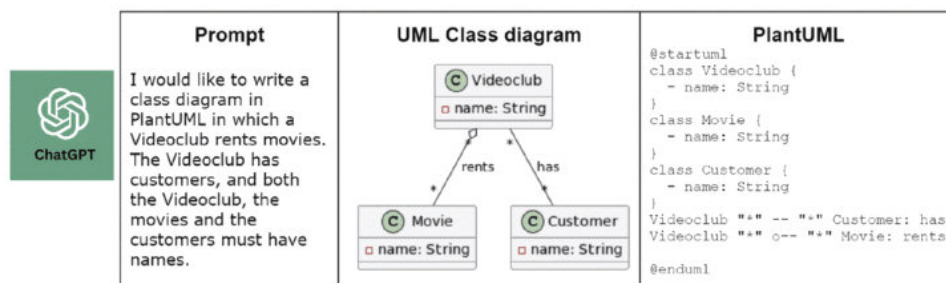


Figure 1: Prompt provided to ChatGPT and UML domain model generated from the prompt.

This work [1], which was openly published in the Software and Systems Modeling journal (SoSyM), explores the current capabilities and limitations of ChatGPT in the context of software modeling. Through a series of experiments and analyses, we sought to understand the extent to which LLMs could assist modelers in their work, as well as the potential pitfalls and challenges that may arise in the process.

One of the primary objectives of the study was to establish whether ChatGPT can generate both syntactically and semantically correct unified modeling language (UML) models. Through a range of tests, we found that while ChatGPT can indeed produce generally accurate UML models, there are still some notable shortcomings in terms of semantic correctness and consistency.

The sensitivity of ChatGPT to context and the domain of the problem at hand was another aspect that we investigated in our

study. The findings are aligned with the common functioning of neural networks, i.e. the more data ChatGPT has been exposed to regarding a particular domain, the more accurate the generated models are likely to be (Figure 1). Conversely, when there is a lack of information or context, the models produced by ChatGPT tend to be less reliable.

The size of the models that ChatGPT can handle was another area that we delved into. According to our observations, there are noticeable limitations when it comes to dealing with larger models, with the system struggling to handle models consisting of more than 10–12 classes. This limitation not only affects the accuracy of the models produced but also the time and effort required to generate them.

In addition to these areas of investigation, we analysed the capability of ChatGPT in utilising various modeling concepts and mechanisms, such as Object Constraint Language (OCL) constraints, associations, aggregations and compositions, among others. The findings indicated a high degree of variability in how ChatGPT manages these different concepts, with the system demonstrating reasonable proficiency in some areas but falling short in others.

The study also looked at the impact of prompt variability and different usage strategies on the quality and correctness of the generated models. We observed that the variability in ChatGPT’s responses often necessitates starting conversations anew to obtain better results. Moreover, the limitations in the size of the models that ChatGPT can handle in a single query

mean that the modeling process becomes an iterative one, where the modeler must progressively add details to the model.

In conclusion, while we acknowledge the potential of LLMs in revolutionising the field of software development, we also emphasise the need for improvements in terms of consistency, reliability and scalability. By addressing these challenges,

we believe that LLMs like ChatGPT can play a prominent role in the future of model-based systems engineering (MBSE), ultimately making software modeling more accessible, personalised and efficient.

Reference:

- [1] J. Cámara, J. Troya, L. Burgueño and A. Vallecillo, “On the assessment of generative AI in modeling tasks: an experience report with ChatGPT and UML,” *Softw. Syst. Model.*, vol. 22, no. 3, pp. 781–793, 2023.

Please contact:

Javier Cámara, ITIS Software / Universidad de Málaga, Spain
jcamara@uma.es

Enhancing Conversational Search with Large Language Models

by Guido Rocchietti, Cristina Ioana Muntean, Franco Maria Nardini (CNR-ISTI)

In the context of the Horizon Europe EFRA research project [L1], we explore the innovative use of Large Language Models (LLMs), both instructed and fine-tuned, in improving the quality of conversational search. The focus is on applying these models in rewriting conversational utterances to enhance the capability of conversational agents to retrieve accurate responses.

Conversational search represents an innovative approach that enables users to engage with information systems naturally and amiably through everyday language dialogues. This entails the system’s ability to effectively preserve contextual information, including the concepts and details conveyed in prior interactions between the user and the conversational agents. Examples of such systems include familiar agents like Alexa, Siri, and others that are seamlessly integrated into our daily routines. Conversational search poses new challenges for information retrieval systems that are asked to extract the correct information from a collection of documents: user utterances submitted during a conversation may be ambiguous, incomplete, or their meaning dependent on the context of previous utterances. Therefore, conversational utterance rewriting is a crucial task aiming to reformulate the user’s requests into more precise and complete queries that enhance retrieval effectiveness.

In our research [1], we explore using Large Language Models (LLMs), both instructed and fine-tuned, to rewrite user queries and increase the quality of the information retrieved by a con-

versational agent. LLMs have shown their ability to comprehend and generate natural language for diverse tasks. Instructed LLMs, such as ChatGPT, are trained to receive detailed instructions from the user about what they must do. On the other hand, fine-tuned LLMs are based on pre-trained models further trained to excel on a particular task using a transfer learning technique.

Our investigation explores to what extent these models are exploitable and reliable when used in an information retrieval setting for rewriting conversational utterances, trying to stress the limitations and the weak points of models that have already become used in many fields of our daily lives.

As instructed LLM, we used ChatGPT powered by GPT 3.5-turbo. We designed and tested prompting templates to evaluate ChatGPT rewriting capabilities. A prompt is a textual instruction specifying the task the model is expected to do. Of course, many ways exist to instruct the model to obtain the same expected results. To establish the best way to input the model, we devised 14 different prompts with different settings, with and without context (i.e. the previous utterances), and generated artificial interactions to provide the model as examples.

Our experiments evidenced that prompting instructed LLMs with few-shot learning, i.e. providing few behavioural examples, is the best setting to produce valuable rewriting, obtaining results significantly higher than the ones obtained with zero-shot learning (i.e. providing only the prompt) and to state-of-the-art utterance rewriting techniques. Specifically, we selected a few couples of utterances, original and rewritten, which we used as the history of interactions between us and ChatGPT. In this way, the model used these as examples to better generate the rewritten version of the queries.

Subsequently, we fine-tuned publicly available LLMs such as Falcon-7B, and Llama-2-13B to generate rewritten and self-explanatory versions of the conversational utterances. The

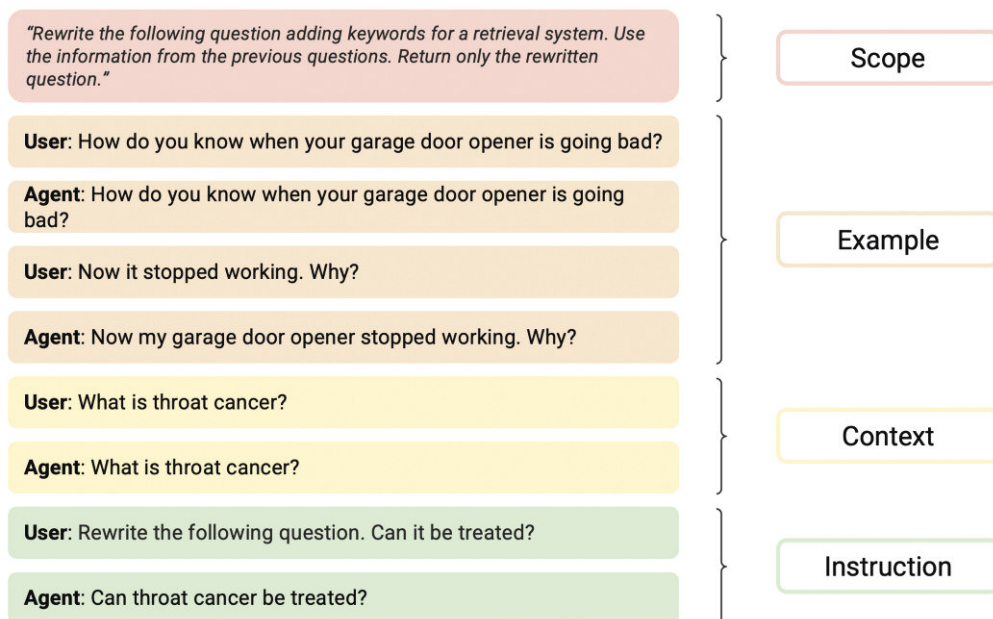


Figure 1: Example of interaction between the user and ChatGPT. Scope indicates the task the models should perform, Example is the artificial interactions provided to the system, Context is the real previous interactions, and Instruction reports the current prompt together with the system response.

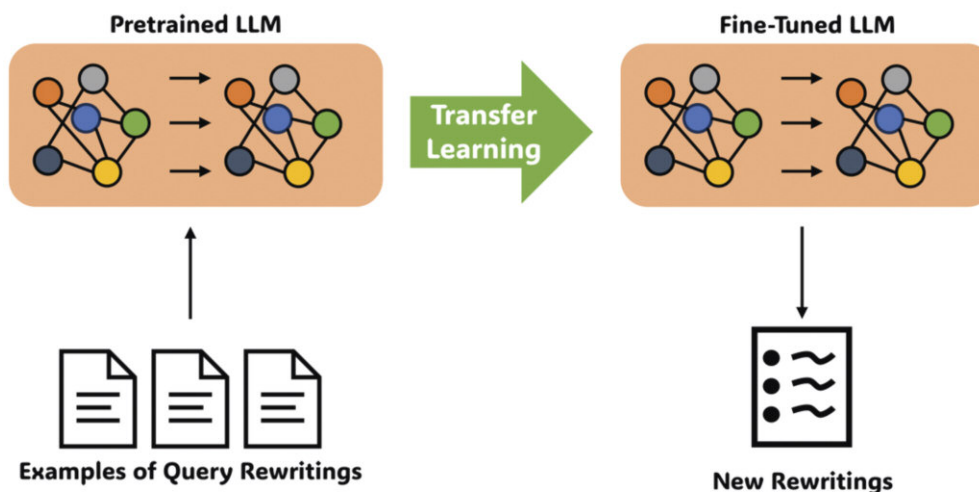


Figure 2: Illustration of the fine-tuning technique. On the left, examples of rewritten queries are provided to the pre-trained model during a transfer learning process. On the right is the resulting model fine-tuned to rewrite new queries.

scope of this experiment was to assess whether using much smaller and open-source models than the closed-source ones powering ChatGPT could achieve comparable results.

To fine-tune the models, we used QReCC, an open-domain dataset with about 81,000 question–answer pairs for conversational question answering. As input for the training phase, we used the previous questions of the conversation as context, followed by the query we wanted to rewrite, while our target was the rewritten version of the same query.

Our results showed that relatively small LLMs, fine-tuned for rewriting user queries, can significantly outperform large instructed LLMs such as ChatGPT. The best performance was obtained with the 13B parameters version of Llama-2.

The reproducible experiments [L2] were conducted on the publicly available TREC CAsT [L3] 2019 and 2020 datasets, which are collections of open-domain conversational dialogues with relevance judgements (i.e. judgements made by humans as to whether a document is relevant to an information need) provided by NIST. We use a two-stage retrieval pipeline based on PyTerrier to assess the rewritten utterances’ retrieval effectiveness. The first stage performs document retrieval using the DPH weighting model, and the second stage performs a neural reranking of the top 1,000 candidates using the MonoT5 model. We compare our methods with state-of-the-art baselines for conversational utterance rewriting, such as QuReTeC and CQR.

Our work opens new directions for research on conversational utterance rewriting with LLMs. The following steps of our research will try to find new ways to obtain even better rewriting results while reducing the size and the computational cost of the inference phase of the model. The idea is to apply existing and novel quantisation techniques to devise an optimal trade-off between rewriting accuracy and computational cost. The availability of such powerful models forces us to thoroughly investigate to what extent we can rely on these kinds of systems, especially given that their impact on society is not negligible, both from a user experience and a sustainability perspective.

Funding for this research has been provided by the EU’s Horizon Europe research and innovation programme EFRA (Grant Agreement Number 101093026). However, views and opinions expressed are those of the authors only and do not necessarily reflect those of the EU or European Commission-EU. Neither the EU nor the granting authority can be held responsible for them.

Links:

[L1] <https://efraproject.eu>

[L2] <https://github.com/hpclab/conv-llm>

[L3] <https://www.treccast.ai/>

Reference:

[1] E. Galimzhanova et al., “Rewriting conversational utterances with instructed large language models,” In Proc. of the 22nd IEEE/WIC Int. Conf. on Web Intelligence and Intelligent Agent Technology (WI-IAT), 2023.

Please contact:

Guido Rocchietti, CNR-ISTI, Pisa, Italy, and University of Pisa, Italy

guido.rocchietti@isti.cnr.it

Knowledge Graph Treatments for Hallucinating Large Language Models

by Diego Collarana, Moritz Busch, and Christoph Lange (Fraunhofer FIT)

Despite the excitement about Large Language Models (LLMs), they still fail in unpredictable ways in knowledge-intensive tasks. In this article, we explore the integration of LLMs with Knowledge Graphs (KGs) to develop cognitive conversational assistants with improved accuracy. To address the current challenges of LLMs, such as hallucination, updateability and provenance, we propose a layered solution that leverages the structured, factual data of KGs alongside the generative capabilities of LLMs. The outlined strategy includes constructing domain-specific KGs, interfacing them with LLMs for specialised tasks, integrating them with enterprise information systems and processes, and adding guardrails to validate their output, thereby presenting a comprehensive framework for deploying more reliable and context-aware AI applications in various industries.

LLMs such as GPT-4 [L1] are a new generation of AI neuronal networks that revolutionised the way we interact with computers. Trained on massive amounts of text data, LLMs exhibit a surprising ability: the larger they get (in terms of artificial neurons), the more capabilities emerge [1]. An LLM solves a variety of NLP tasks, including question answering, machine

translation, and even code generation. As a result, enterprises are begging to implement real-world applications powered by LLMs [2]. All big tech companies are in the race to incorporate LLMs into their products, e.g. Microsoft 365 Copilot [L2], for assisted content generation based on instructions (also known as prompts).

While LLMs demonstrate remarkable capabilities, they still face limitations in knowledge-intensive tasks [3], i.e. scenarios where factually correct answers are critical, such as in industrial or healthcare applications. The three main shortcomings hindering the use of LLMs in knowledge-intensive applications are hallucination, expensive updateability and lack of provenance. Hallucination refers to the tendency of LLMs to generate factually incorrect or misleading information. Expensive updateability is the struggle to incorporate new knowledge or adapt to evolving information landscapes. Lack of provenance is the ability to trace the origin of information. These problems remain a challenge for LLMs, making it difficult to assess the reliability and trustworthiness of their output.

To successfully implement LLM-enabled applications, we need to separate two components: one serving as a knowledge storage and the other providing linguistic capabilities. This approach allows us to leverage the complementary strengths of each component, i.e. the knowledge store provides factual accuracy and the linguistic capabilities of the LLM to generate creative and informative text. Figure 1 shows our approach that combines KGs and LLMs. KGs are structured sources of factual knowledge, including ontologies and taxonomies. We use KGs to provide LLMs with access to reliable and up-to-date business information, reducing hallucination and improving the accuracy in the output. We preserve the linguistic capabilities of LLMs, including their ability to understand and

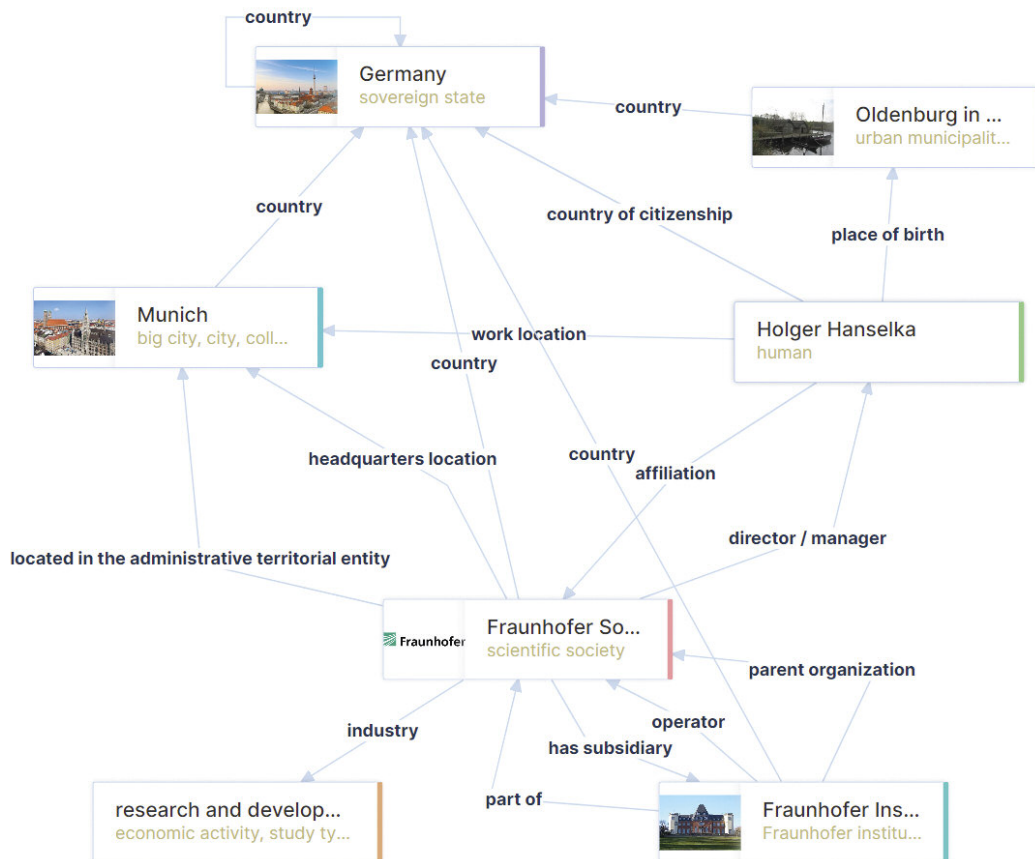


Figure 1: Knowledge Graph encoding the relationship between the Fraunhofer Society and its current president, Dr. Holger Hanselka. GPT-3 Turbo still shows Dr. Reimund Neugebauer as the current president, requiring an expensive retraining process to update this knowledge. Note that either RDF or Neo4j can be used as the technical implementation of the KG. (Source Wikidata).

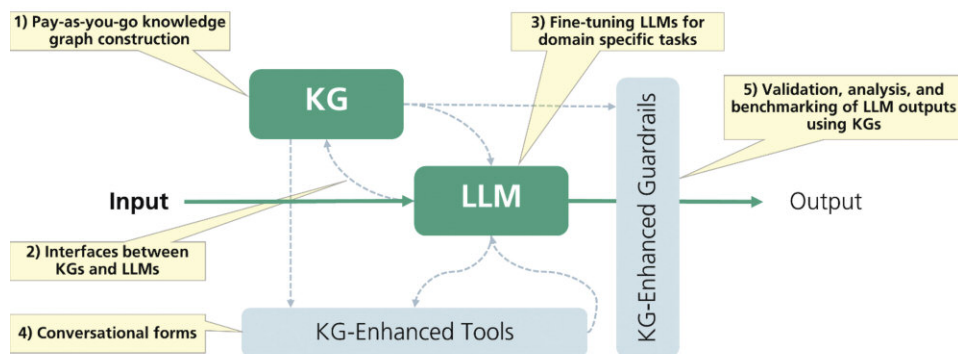


Figure 2: The approach for developing state-of-the-art cognitive conversational assistants: from building a knowledge graph, interfacing it with language models, fine-tuning models for domain-specific tasks, connecting existing enterprise tools to the model, to validating and analysing the output. All these steps are empowered using Knowledge Graphs as knowledge store of the solution.

generate text. This combination results in a more flexible and adaptable LLM architecture that can incorporate business knowledge as needed.

Why KGs [L3] as knowledge store? The answer lies in their ability to drive agile knowledge integration, enabling organisations to seamlessly integrate heterogeneous data from disparate sources (Figure 1). KGs enable us to apply logical rules and reasoning to leverage the collective expertise of an organisation for improved knowledge discovery. KGs also unlock the potential for personalised recommendations, tailoring the user experience to individual interactions and preferences. KGs allow the maintaining of data governance and quality, enabling organisations to enforce data standards, relationships and hierarchies.

Our approach, depicted in Figure 2, focuses on the development of state-of-the-art cognitive conversational assistants, also known as copilots. These intelligent assistants harmoniously fuse KGs and LLMs. This fusion creates copilots that are not only adept at understanding context and providing accurate information, but also capable of engaging in dynamic, human-like conversations. Our approach is composed of the following steps:

1. Knowledge Graph construction: In this step, a domain-specific knowledge graph is constructed from heterogeneous data sources. We encourage a pay-as-you-go KG construction, i.e. integrate only the necessary knowledge to implement a use case. The KG can be revisited and extended afterwards, incorporating further data sources if such need arises. Unlike updating knowledge in LLMs, which requires expensive retraining, extending KGs does not require much computational power.
2. Interfacing KG with LLM: As the second and main step, we use techniques, such as retrieval-augmented generation (RAG) and text-to-graph query generation, to connect LLMs bi-directionally with KGs. RAG indexes sub-graphs from the KG into a vector story: using retrieval approaches, the top-k most relevant sub-graphs are transmitted to the LLM as additional context. Text-to-graph query generation components translate natural language queries into structured queries, making it easier to extract precise and relevant information from KGs. This widens the user base of our approach, partially overcoming the necessity for graph query language knowledge.
3. Fine-tuning LLMs for domain-specific tasks: as KGs may not suffice, LLMs themselves can be tailored to specific tasks. Fine-tuning LLMs aims at an enhancement of data-

driven decision-making, streamlined processes, and superior results in the field at hand. However, this comes with an increased demand of computing power. Through benchmarking, we assess the performance of LLMs in various contexts, facilitating continuous improvement and fine-tuning to meet the specific needs and standards of the target application.

4. Connecting tools: In this step, we connect existing enterprise tools with the LLM. This connection allows a seamless integration of existing software ecosystem into the copilots, enhancing user interactions and providing innovative conversational experiences.
5. Validation and analysis of outputs: To deploy an LLM-based solution into production, we need guardrails that ensure the accuracy and reliability of the LLM-generated content. In this step, we apply methods to analyse and validate the output against the structured knowledge stored in KGs. These methods allow in-depth examination of the LLM output, providing valuable insights and actionable decisions before the output is released to the user.

In conclusion, the integration of LLMs with KGs represents an innovative leap forward in developing conversational assistants that are both knowledgeable and linguistically proficient. By combining the robust, up-to-date knowledge provided by KGs with the advanced text generation capabilities of LLMs, we overcome the existing limitations around factuality, updatability and provenance, enhancing the user experience but also establishing responsible deployment of AI in critical business and knowledge-intensive domains.

Links:

[L1] <https://openai.com/gpt-4>

[L2] <https://kwz.me/hAY>

[L3] <https://kwz.me/hAB>

References:

- [1] T. B. Brown et al., “Language models are few-shot learners,” *NeurIPS*, 2020.
- [2] R. Bommasani et al., “On the opportunities and risks of foundation models,” *CoRR abs/2108.07258*, 2021.
- [3] S. Pan et al., “Unifying Large Language Models and Knowledge Graphs: A Roadmap,” *CoRR abs/2306.08302*, 2023.

Please contact:

Diego Collarana, Fraunhofer FIT, Germany
diego.collarana.vargas@fit.fraunhofer.de

Revolutionising Theatre Archives: Using Large Language Models to Interact with Structured Archival Content

by Chara Tsoukala, Georgios Paraskevopoulos and Athanasios Katsamanis (Athena Research Center)

The Greek National Theatre has introduced an advanced chatbot based on Large Language Models (LLMs) as a novel means of accessing the content of the digital archive on the web. This state-of-the-art chatbot, which utilises Text2SQL conversion through LLMs, offers a more intuitive user experience, enabling complex searches with simple natural language. This addition marks a significant step forward in making theatrical performance data accessible in a more interactive way.

Large Language Models (LLMs) have emerged as transformative tools, revolutionising the way that we interact with complex data systems. These advanced AI models are known for their ability to understand and generate human-like text, making them ideal for a range of applications. The Greek National Theatre (GNT) has leveraged this technology to revolutionise access to its extensive digital archive content. By implementing a sophisticated chatbot that utilises Text2SQL conversion

through LLMs, the Greek National Theatre has made its vast collection of past performances, theatrical works, and multimedia content accessible through a simple, natural language interface. This approach not only enhances the user experience by simplifying the interaction with the theatre archive's content but also enables more nuanced and complex searches, setting a new standard in how we access and explore historical data in cultural heritage and the arts.

Large Language Models (LLMs) like GPT-3 are not just skilled in understanding and generating human-like text; they also hold immense potential in enhancing database utilisation and management, particularly in structured data querying using SQL. This capability is especially transformative for the GNT digital archive. The archive contains a plethora of information on plays, performances, actors, directors, costumes, sets, and multimedia elements. However, accessing this data traditionally requires specific technical know-how. With the integration of LLM technology, however, this wealth of information becomes far more accessible to a broader audience, significantly simplifying the process of extracting valuable insights from this extensive source of knowledge.

The chatbot employed at the GNT digital archive is more than just a digital assistant; it's an intelligent conversational system integrating state-of-the-art AI technologies. Users interact with the archive through a specialised widget within the web application of the GNT digital archive. They can type queries in natural language about performances, actors and other relevant topics, receiving precise answers linked to the digital content in the archive. The LLM, with tailored prompting, translates user natural language queries into SQL, fetches results, and

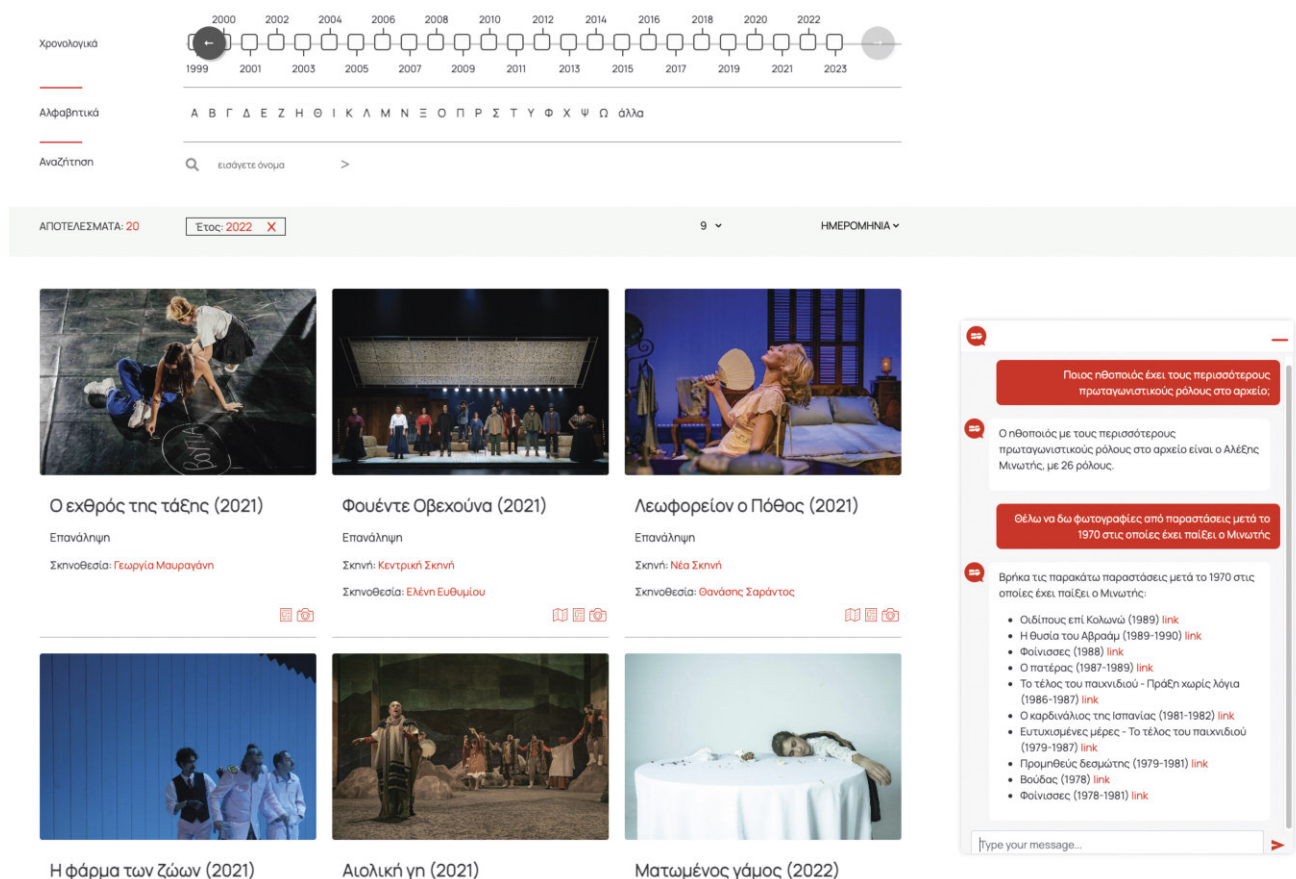


Figure 1: The Greek National Theatre archive interface featuring the chatbot widget, displaying results related to the above scenario in Greek.

presents them in natural language. This dual conversion process not only makes information retrieval intuitive but also allows for complex queries, such as searching for specific plays or performances featuring certain actors within a certain time period. For instance:

User: “Which actor has had the most leading roles?”

Assistant: “The actor who has starred the most is Alexis Minotis, with 26 leading roles.”

User: “I want to see photos from performances after 1970 in which Minotis has acted.”

Assistant: “I found the following performances after 1970 in which Minotis has acted:

- *Oedipus at Colonus* (1989)
- *The Sacrifice of Abraham* (1989–1990)
- *The Phoenician Women* (1988)
- *The Father* (1987–1989)
- *Endgame - Act Without Words* (1986–1987)
- *The Cardinal of Spain* (1981–1982)
- *Happy Days - Endgame* (1979–1987)
- *Prometheus Bound* (1979–1981)
- *Buddha* (1978)
- *The Phoenician Women* (1978–1981).”

In this scenario, as depicted in Figure 1, the chatbot enriches its response by providing links, allowing the user to further explore the play’s webpage for additional information. In collaboration with the theatre’s dramaturge, we have developed various such scenarios to ensure that a wide range of potential user inquiries could be addressed. These scenarios include simple searches like finding a specific performance, detailed information about a work, personnel involved, and multimedia material from performances. There are also complex searches, such as finding performances with common attributes, like international tours or specific actors, within a certain time frame. The chatbot’s “persona” is a theatre enthusiast, knowledgeable about the entire digital archive, offering users a unique and engaging way to explore theatrical history.

While this technology marks a significant advancement, it is not without challenges. The accuracy and complexity of queries depend on the chatbot’s understanding and the database’s structure; there is an upper limit to the complexity of the database, after which we observe hallucinations and incomplete queries. Additionally, ensuring the chatbot’s personality remains engaging and informative without overwhelming the user is crucial.

The application of an intelligent conversational system based on LLMs for accessing information from a theatrical digital archive is a pioneering example of how AI can enhance cultural heritage content. This approach could be replicated in other fields, opening new avenues for interacting with historical and educational archives. The future may see further advancements in LLMs and Text2SQL technologies, leading to even more sophisticated and intuitive user experiences. This approach not only enhances visual interaction but also deepens the users’ engagement with the rich history of theatrical arts, making it more accessible and enjoyable for everyone.

Please contact:

Chara Tsoukala, Institute for Language and Speech Processing, Athena Research Center, Greece
chara.tsoukala@athenarc.gr

Improving the Sample Efficiency of Pre-training Language Models

by Gábor Berend (University of Szeged)

The use of transformer-based pretrained language models (PLMs) arguably dominates the natural language processing (NLP) landscape. The pre-training of such models, however, is known to be notoriously data and resource hungry, which hinders their creation in low-resource settings, making it a privilege to those few (mostly corporate) actors, who have access to sufficient computational resource and/or pre-training data. The main goal of our research is to develop such a novel sample-efficient pre-training paradigm of PLMs, which makes their use available in the low data and/or computational budget regime, helping the democratisation of this disruptive technology beyond the current status quo.

Masked language modelling (MLM) is a frequently used pre-training objective of PLMs, enjoying its popularity due to its ease of implementation and its ability to produce pre-trained models that can achieve state-of-the-art models given that the pre-training was conducted over a sufficient number of training steps on a sufficiently large and diverse pre-training corpus.

Traditional MLM pre-training works similarly to fill-in-the-blank tests, i.e. the neural language models are expected to develop an ability to insert the correct substitute in a partially given sentence, where the missing words are indicated by a special MASK symbol. For example, given the input sentence “Sarah is eating a peach.”, with the word “peach” being the word that was randomly chosen for masking, a neural model would be expected to output such a probability distribution for the potential substitutes of the token [MASK] in a sentence “Sarah is eating a [MASK].”, which puts all the probability mass to the token “peach” (see Figure 1). Expecting the neural model to assign all the probability mass to a single word seldom constitutes a plausible behaviour from a common-sense perspective, since there exists a handful of other words besides the single word “peach” that would serve as a cognitively sound replacement of the [MASK] token, including words like “pizza”, “croissant”, “soup”, etc. The reason why this kind of pre-training still works is because pre-training is conducted over a large and diverse enough collection of pre-training corpus, which makes the behaviour of the pre-trained language model make sense in the long run, making its outputs acceptable to humans.

Arguably, we could enjoy a more sample-efficient pre-training if the target distribution of potential substitutes would involve an actual distribution over the words such that it puts a nonzero probability to more than just a single-word form (see Figure 2). It is nonetheless unrealistic to have access to such a pre-training corpus, which contains all the possible replacements of any arbitrarily chosen word in it. Enhancing pre-training with the semantic properties of the masked words could also improve the sample efficiency of the pre-training

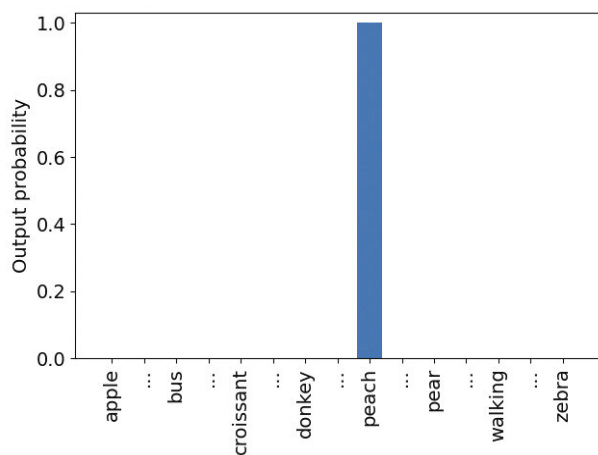


Figure 1: The kind of distribution expected during Masked Language Modelling pre-training.

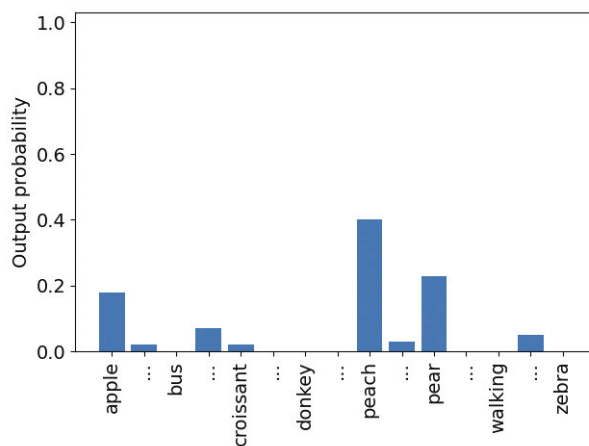


Figure 2: The kind of distribution expected during idealised pre-training.

(see Figure 3); however, the availability of a pre-training corpus that is annotated with the semantic properties of the words even at a moderately large scale is equally unrealistic.

To improve the sample efficiency of the pre-training – without the need for any additional annotation of the pre-training corpus – we introduced a novel unsupervised pre-training objective, masked latent semantic modelling (MLSM) [1]. Our approach performs a context-sensitive multi-clustering of the masked words, based on their neural representation, and we use the cluster membership of the words in their context as the target distribution to perform the pre-training on. When comparing this kind of pre-training with traditional MLM and other forms of knowledge distillation (KD) methods, we have seen vast improvements in sample efficiency, i.e. the pre-trained language models have reached much better downstream performance at much earlier stages of the pre-training process. We have demonstrated the utility of the proposed pre-training objective for multiple PLM architectures. While our primary focus of investigation was English so far, we have also successfully used the proposed pre-training objective for pre-training language models in Hungarian [2]. As the entire approach is unsupervised, it can be conveniently applied to

any language without the need for any modification in the methodology. Due to its favourable sample efficiency, MLSM pre-trained PLMs can be especially useful in scenarios when the pre-training corpus size is limited for some reason.

Most recently, we also participated at the BabyLM Challenge [L1], where the goal was to build PLMs over a pre-training corpus consisting of no more than 10 million words (in contrast, the Chinchilla model was trained over more than 100,000 times as many, 1.3 trillion words). The challenge has also attracted media attention [L2] as well as that of academic participants, as more than 120 models were submitted by more than 30 participating teams. Our submission, which was based on MLSM, got ranked 2nd, indicating its applicability in low-resource scenarios. We made the model that we submitted to the challenge publicly available [L3].

We are currently experimenting with new variants of the previously proposed MLSM approach that can yield further improvements in sample efficiency. Additionally, we would like to expand our results to additional (low-resource) languages and investigate these models from an interpretability point of view in the future.

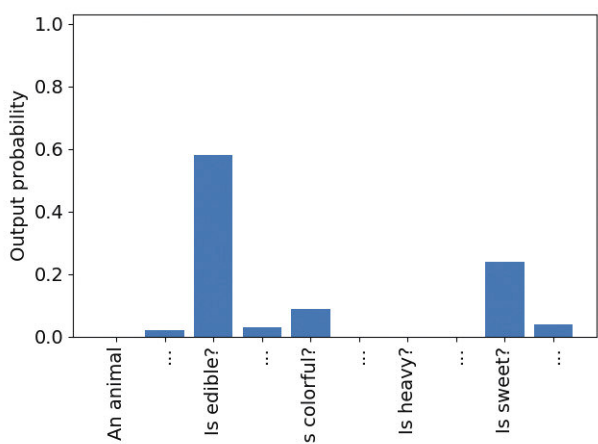


Figure 3: The kind of distribution expected during semantically enhanced pre-training.

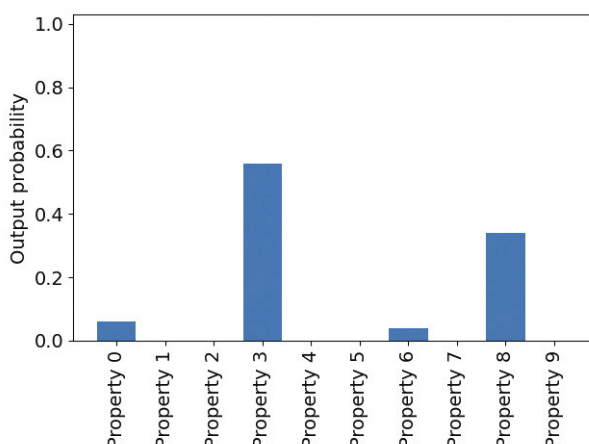


Figure 4: The kind of distribution expected during MLSM pre-training.

This research received support from the European Union project RRF-2.3.1-21-2022-00004 within the framework of the Artificial Intelligence National Laboratory. Additionally, we are grateful for the support from the HUN-REN Cloud infrastructure.

Links:

[L1] <https://babylm.github.io/>

[L2] <https://kwz.me/hAb>

[L3] <https://huggingface.co/SzegedAI/babylm-strict-small-mlsm>

References:

[1] G. Berend, “Masked Latent Semantic Modeling: an Efficient Pre-training Alternative to Masked Language Modeling,” In Findings of the Association for Computational Linguistics: ACL 2023, pp. 13949–13962, Jul. 2023.

[2] G. Berend, “Látens szemantikus eloszlások használata a nyelvi modellek előtanítása során,” XIX. Magyar Számítógépes Nyelvészeti Konferencia.

[3] G. Berend, Better Together: Jointly Using Masked Latent Semantic Modeling and Masked Language Modeling for Sample Efficient Pre-training, Association for Computational Linguistics, 2023.

Please contact:

Gábor Berend, University of Szeged, Hungary
berendg@inf.u-szeged.hu

ChatGPT Multilingual Querying Consistency – A Test Case

by George Tambouratzis (Athena Research Centre)

Conversational agents and chatbots have developed rapidly in the past year to provide answers to user queries, drawing information from huge collections of data. From the user-side, the usefulness of conversational agents hinges on the accuracy of response in addition to user-friendliness and response speed. Here we briefly evaluate one of the most widely used chatbots, ChatGPT over a set of queries posed using multiple languages, to test its robustness and consistency, while running the experiment at two timepoints to monitor ChatGPT’s evolution.

Conversational agents have been widely used at a global level and converse with humans in multiple languages. Despite remarkable progress fuelled by advances in Large Language Models (LLMs) [1], several types of problems persist. Categories of errors include hallucinations, syntax errors, and prompt brittleness [2] [3]. One question that arises is if they provide consistent responses in a multilingual environment. Will users X and Y, posing equivalent queries in different languages A and B, get equivalent or fundamentally different responses?

Defining Queries

To measure performance, questions posed to ChatGPT concern historical events dating back over 80 years and which are documented and widely accepted. The questions are listed in English in Table 1, but have been translated by language experts into German, French, Italian, Spanish and Greek.

Q1	What was the fate of Elli?
Q2	How was Elli lost?
Q3	Was the Elli sunk by torpedoes?
Q4	What about the rumour that Elli was struck by a mine?
Q5	How was Elli sunk?

Table 1: Queries posed.

The questions focus on the Greek destroyer Elli, lost in 1940 to enemy action. With the exception of Q3, questions are open-ended. Q2 and Q5 focus on the ship’s end, while Q3 and Q4 suggest a reason for the loss of the ship. The use of related questions allows us to measure prompt brittleness of the underlying LLM via the variation in the system answers.

First experiment

The responses obtained (April 2023) by posing questions Q1 to Q5 to ChatGPT are summarised in Table 2. Responses in all languages place Elli as being harboured at Tinos island during the event (coinciding with the actual events).

Responses in all languages correctly report enemy action in August 1940, while inconsistencies occur in the number of casualties. However, the major discrepancies are in the German language where the ship is stated as surviving the August events to be lost in another event months later. Interestingly, for the Greek query, following further prompting, ChatGPT rescinded from its original response to state that a second naval ship (a cruiser) named Elli was also present at Tinos but survived the events and served for years. This hallucination is probably caused by the fact that (i) Elli was termed both as a destroyer and a light cruiser and that (ii) as is popular practice in navies, other ships have later used the same name. Other hallucinations include the raising of the ship in 1952 (French) and the sinking at late 1940 (German). The main conclusion is inconsistencies occur across languages, which is not desirable. Furthermore, Q5 causes ChatGPT to revise its responses, indicating uncertainty.

Repeating the experiment with ChatGPT-3.5 (Nov. 2023)

The newer ChatGPT version requires more clarifications from the user about period and disambiguation of terms before responding, which is probably integrated to increase accuracy. Results are shown in Table 3. For the later experiments, more variation occurs across languages. ChatGPT responds in three languages that the Elli was subsequently raised and in English and German it states that the ship served for years (actually Elli forms a sunken memorial since August 1940).

The ship’s nationality is inconsistently reported across languages. For queries in French, Elli is a German ship sunk by the British submarine HMS Sportsman. For Spanish, the Elli

Language	English	Greek	French	Italian	Spanish	German
A. Type of attack	Submarine- 2 torpedoes	Submarine 2 torpedoes	Submarine 2 torpedoes	Submarine 3 torpedoes	Submarine 2 torpedoes	Aircraft attack
B. date	15-Aug-1940	15-Aug-1940	1940	15-Aug-1940	15-Aug-1940	15-Aug-1940
C. Fate	sunk	sunk	sunk	sunk	sunk	survived
D. Next events			Raised in 1952, scrapped in 1953			Sunk (Oct-1940) at Salamis
E. answer following Q5	Lost to mine	A cruiser ship also named Elli survived				

Table 2: Summary of ChatGPT responses (April 2023) – blank cells in rows D and E indicate unchanged response.

Language	English	Greek	French	Italian	Spanish	German
A. Type of attack	Submarine- 2 torpedoes	Submarine – explosives	Submarine 1 torpedo	Submarine 2 torpedoes	Submarine 1 torpedo	Submarine – torpedo
B. date	15-Aug-1940	Aug-1940	26-May-1944	15-Aug-1940	Sweden-Aug-1940	15-Aug-1940
C. Fate	Damaged	Sunk	Sunk	Sunk	sunk	Sunk
D. Next events	Repaired; served for years		Raised in 1952, scrapped in 1953			Repaired; served several years
E. changes following Q5	Sunk after attack	Survived for years until scrapped	Elli was German; submarine was British	attack by Marauders with explosives	Loss caused by mine; then torpedo	

Table 3: Response summary (November 2023).

is a Swedish ship sunk by German action near the Aland isles (Sweden). Both represent hallucinations, where ChatGPT probably combines unconnected events, producing erroneous output. The instability regarding the cause of loss for Spanish prompting (stated as torpedo for Q2, then mine for Q3, then torpedo for Q4) is also a phenomenon not encountered in the first set of experiments.

Conclusions

ChatGPT provides a strong capability to respond to specialised user queries. This manuscript has investigated its use in multilingual settings and indicates that queries in different languages produce fundamentally different responses even for widely accepted events. The newer ChatGPT version has access to more extensive amounts of information which likely causes more frequent hallucinations by inadvertently connecting unrelated events. It is essential to improve the consistency in multilingual settings to ensure accuracy of results, and research work is underway to develop such methods on the basis of multilingual cross-checking. In future, the aim is to extend this study to cover grammaticality of responses.

References:

- [1] W. X. Zhao et al., “A Survey of Large Language models,” 2023, ArXiv:2303.18223v12.
- [2] J. Kaddour et al., “Challenges and Applications of Large Language Models,” 2023, Arxiv:22207.10169v1.
- [3] Y. Bang et al., “A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity,” Arxiv: 2302.04023.

Please contact:

George Tambouratzis, Athena Research Centre, Greece
giorg_t@athenarc.gr

A Pipeline for Validating ChatGPT Responses Using Knowledge Graphs and Embeddings

by Michalis Mountantonakis and Yannis Tzitzikas (FORTH-ICS and University of Crete)

Since it is challenging to combine ChatGPT (which has been trained by using data from web sources), with popular RDF Knowledge Graphs (that include high quality information), we present a generic pipeline that exploits RDF Knowledge Graphs and short sentence embeddings for enabling the validation of ChatGPT responses, and an evaluation by using a benchmark containing 2,000 facts for popular Greek persons, places and events.

The novel Artificial Intelligence ChatGPT chatbot offers detailed responses and human-like answers for many domains of knowledge; however, it can produce erroneous facts even for popular entities, such as persons, places and events. On the other hand, there are several RDF Knowledge Graphs (KGs) containing high-quality data and their provenance is recorded; however, it is difficult for non-expert users to query such data. Therefore, one challenge is how to combine GPT with popular RDF KGs for enabling the validation of ChatGPT responses with justifications and provenance. For this reason, the Information Systems Laboratory of the Institute of Computer Science of FORTH has designed a pipeline [1,2] for enabling the validation of ChatGPT facts from RDF KGs, and has cre-

ated a benchmark containing 2,000 ChatGPT facts for enabling the evaluation of the mentioned pipeline [2].

Concerning the pipeline, which is also used in the online research prototype called GPT•LODS [L1], its steps are depicted in the left side of Figure 1. Moreover, in the right side of Figure 1, we show a real example for a question about “Marcell Jacobs”, the gold medallist of the 2020 Summer Olympics in the men’s 100 metres (one of the most popular events of the Olympic Games). In particular, as a first step we can ask ChatGPT a question (e.g. see Figure 1) and as a second step we can send a new prompt for asking ChatGPT to convert its textual response to RDF N-triples. As we can see in Figure 2, ChatGPT can produce syntactically valid RDF triples by using a given model/ontology (e.g. DBpedia model) [1, 2]. However, in that example, we observe some errors in the ChatGPT facts for that person (e.g. his birth date). For correcting such errors, or/and for confirming correct ChatGPT facts, the third step is to search in one or more KGs to find the corresponding facts. In particular, we use either DBpedia (by accessing its SPARQL Endpoint), or LODsyndesis KG (by accessing its REST API), which has integrated the contents of 400 RDF KGs including DBpedia, by having pre-computed the transitive and symmetric closure of their equivalence relationships.

For each fact we use an algorithm with three different rules [1, 2]. Specifically, we first search for the exact or a semantically equivalent triple in the KG (rule A). Afterwards, we search to find triples having the same subject-predicate or subject-object (rule B). Finally, (if the previous two rules failed) we search for the K most similar triple(s) by first collecting all the triples for the entity (the subject) of the triple, and by computing the cosine similarity of the embeddings between the ChatGPT fact and each triple (rule C). As a final step, the pipeline produces

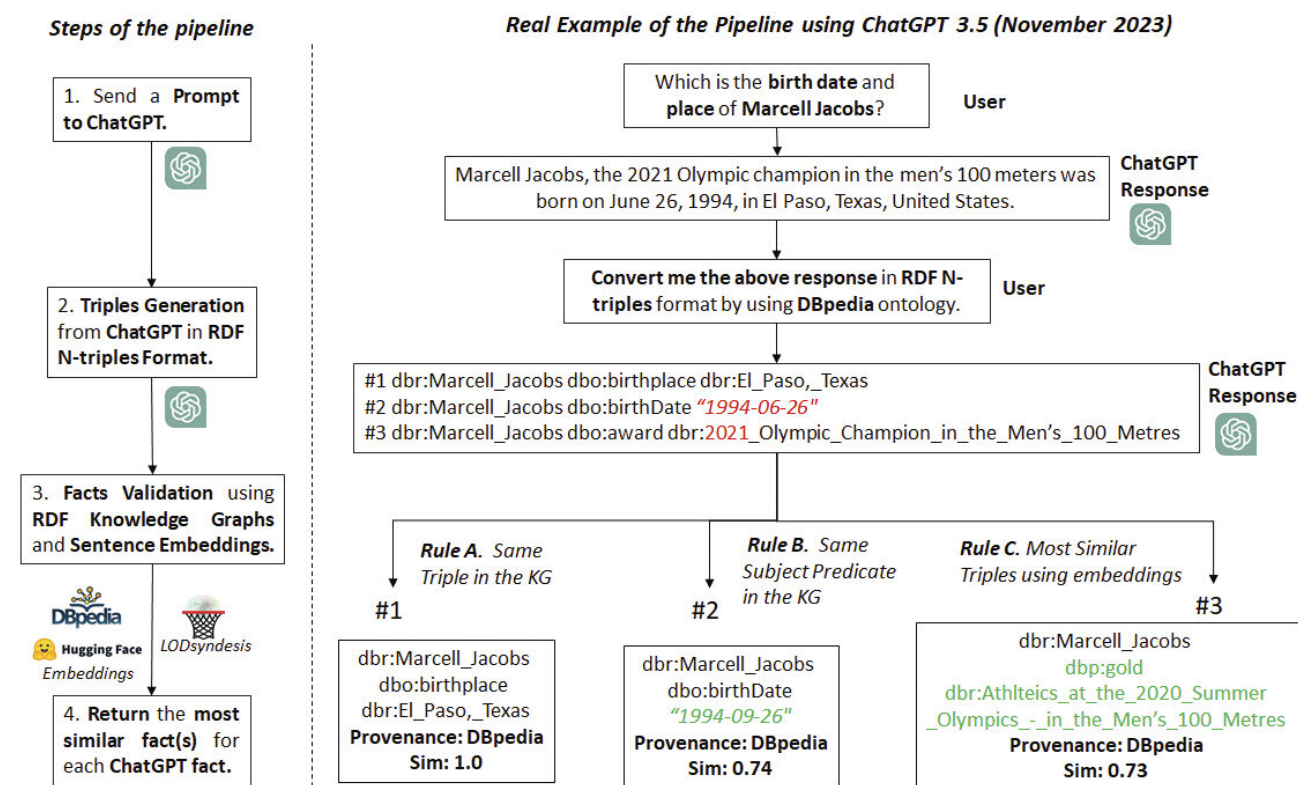


Figure 1: The steps of the pipeline and a running example.

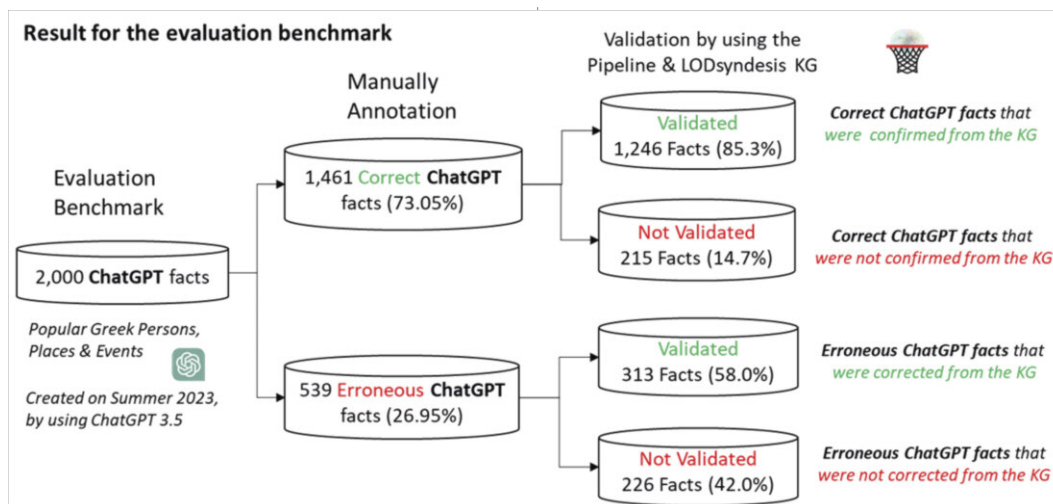


Figure 2: The evaluation benchmark and the corresponding results.

the K most similar triples from the KGs (and their provenance) to each ChatGPT fact and their cosine similarity score, offering to the user a kind of support for the returned facts from the KG. Concerning the example of Figure 1, for the birthplace of Marcell Jacobs, rule A was executed since we found the exact triple in DBpedia. Regarding his birthdate, rule B was executed, i.e. we found the same subject-predicate but a different object in DBpedia (i.e. the correct birth date), since his birth date was erroneous in ChatGPT. Finally, for the third fact, rule C was executed, where we found the most similar fact from DBpedia (by using embeddings and cosine similarity), and we returned it to the user (by also providing information about the provenance).

Concerning the evaluation setup, we have created a benchmark (see Figure 2) that includes 2,000 ChatGPT facts for popular Greek persons, places (including cities, mountains, lakes, heritage sites and islands) and events (including battles, sports, earthquakes and elections). The benchmark is available in a GitHub page [L2], and was created in August 2023, by using the ChatGPT 3.5 model. Regarding the collected ChatGPT facts, we manually annotated each fact as either correct or erroneous, by checking online trusted sources. Indeed, from the 2,000 ChatGPT facts, approximately 73% of them were annotated as correct and 27% as erroneous. Concerning the erroneous facts, they mainly included dates and numbers, such as birth dates, population of places, length, height and others.

Concerning the evaluation results, by using the mentioned pipeline and LODsyndesis KG (see the right side of Figure 2) we managed to confirm 85.3% of the correct ChatGPT facts (+6.2% compared to using only a single KG, i.e. DBpedia), and to find the correct answers to the 58% of erroneous ChatGPT facts (+2.6% compared to using only DBpedia). Concerning cases where the pipeline failed to find the corresponding answer in the KG(s), it included incompleteness issues, i.e. cases where the KG(s) did not contain the fact, contradicting values between different KGs and cases where the answer was included in large literals.

To tackle the above limitations, as a future work, we plan to (i) further extend the benchmark with more entities, facts and domains, (ii) evaluate the pipeline in other Large Language Models, and (iii) extend the algorithm for supporting answer extraction from large literals.

Links:

- [L1] <https://demos.isl.ics.forth.gr/GPToLODS>
- [L2] <https://kwz.me/hzZ>

References:

- [1] M. Mountantonakis and Y. Tzitzikas, “Real-time validation of ChatGPT facts using RDF Knowledge Graphs,” ISWC Demo Paper, 2023.
- [2] M. Mountantonakis and Y. Tzitzikas, “Validating ChatGPT facts through RDF Knowledge Graphs and sentence similarity,” arXiv preprint arXiv:2311.04524, 2023.

Please contact:

Michalis Mountantonakis, FORTH-ICS and University of Crete, Greece
mountant@ics.forth.gr

Yannis Tzitzikas, FORTH-ICS and University of Crete
tzitzik@ics.forth.gr, +30 2810 391621

A Unified Model for Automated Evaluation of Text Generation Systems

by Jan Deriu and Mark Cieliebak (Zurich University of Applied Sciences)

With all the recent hype around Large Language Models and ChatGPT in particular, one crucial question is still unanswered: how do we evaluate generated text, and how can this be automated? In this SNF project, we develop a theoretical framework to answer these questions.

We are witnessing the rapid advancement of Large Language Models (LLMs) that can tackle many natural language processing tasks, which seemed hardly possible only a few years ago. For instance, performing question-driven summarisations of long meeting transcripts seemed out of scope before ChatGPT. Despite the rapid advancement and application of these text generation models, in both public and professional domains, the academic community still struggles with establishing consistent, reliable methods to evaluate the quality of such models.

The dilemma extends across the spectrum of evaluation methods. On one end, human-based evaluations, while offering nuanced insights, are notoriously time-consuming, expensive,

and prone to inconsistent results due to low inter-rater agreement. On the other end, automated metrics promise more efficient improvement cycles for LLMs by providing timely feedback at low cost. Yet, these too have their pitfalls: untrained metrics, e.g. word overlap measures such as BLEU, despite their simplicity and widespread use, often fail to correlate strongly with human judgements. Trained metrics, which are trained to emulate human ratings, show better alignment with human assessments, but they are fragile, requiring domain-specific retraining and being highly sensitive to the training data, parameters and adversarial manipulations.

The result is that human evaluations are highly uncertain due to small sample sizes (i.e. they have a high variance), while automated evaluations are more certain about the results, but they do not agree well with human ratings (i.e. they have a low variance but a high bias). The consequence of using automated metrics is that the evaluation results cannot be trusted.

As an example, assume that you are tasked with evaluating the hallucination rate in the creation of summaries by ChatGPT (i.e. in how many cases it generates factually wrong statements), and compare it to Falcon-180B (a well-known open-source LLM). How would we proceed? We would take a batch of texts to be summarised and generate the summaries using both systems. Then we would ask humans to read these summaries and source texts, and annotate the hallucinations. However, this task is very tedious, and we might be only able to gather, say, 50 human annotations per system due to time and money constraints. Now, we use GPT-4 to automatically



Figure 1: Illustration created with Dall-E3 with the following prompt: “Image of a robot evaluating two chatbots, and a human overseeing the process”.

rate the generated texts, and we are able to gather another 1,000 automated annotations. However, are we sure that these can be trusted? How can we leverage the data provided in this scenario to create an evaluation of ChatGPT and Falcon-180B, where we can state if one hallucinates less than the other? How can we mitigate the mistakes that automated metrics make in contrast to human evaluation? Is it possible that GPT-4 unfairly favours the texts generated by ChatGPT? How can we select texts which have the highest discriminatory power, i.e. which best show us the differences between the two systems?

The goal of the UniVal project, a 3-year SNF project starting in January 2024 at ZHAW, is to provide answers to these questions. We will develop a theoretical framework to ensure trustworthy evaluation by developing a Bayesian model for the evaluation process to mitigate the aforementioned issues with evaluation. We already performed two pre-studies to test the concept [1, 2], which showcased the modelling power of our approach by combining human and automated ratings. In [1], we showed that by explicitly modelling the mistakes of the metrics with respect to the human evaluation, it is possible to create a trustworthy evaluation, which is precise in stating how certain the evaluation is that the hallucination rate of one system is lower than that of the other system. In [2], we showcased how to create a trustworthy evaluation guideline from the theoretical framework to reduce the number of human annotations needed. The modelling approach that we present allows us to run counterfactual examples, which gives us insights into how an evaluation needs to be prepared in order to yield the insights that we wish to gain. That is, how many human ratings are needed, how many automated ratings, and how good the automated metric has to be. Thus, yielding a very powerful tool in the evaluation arsenal.

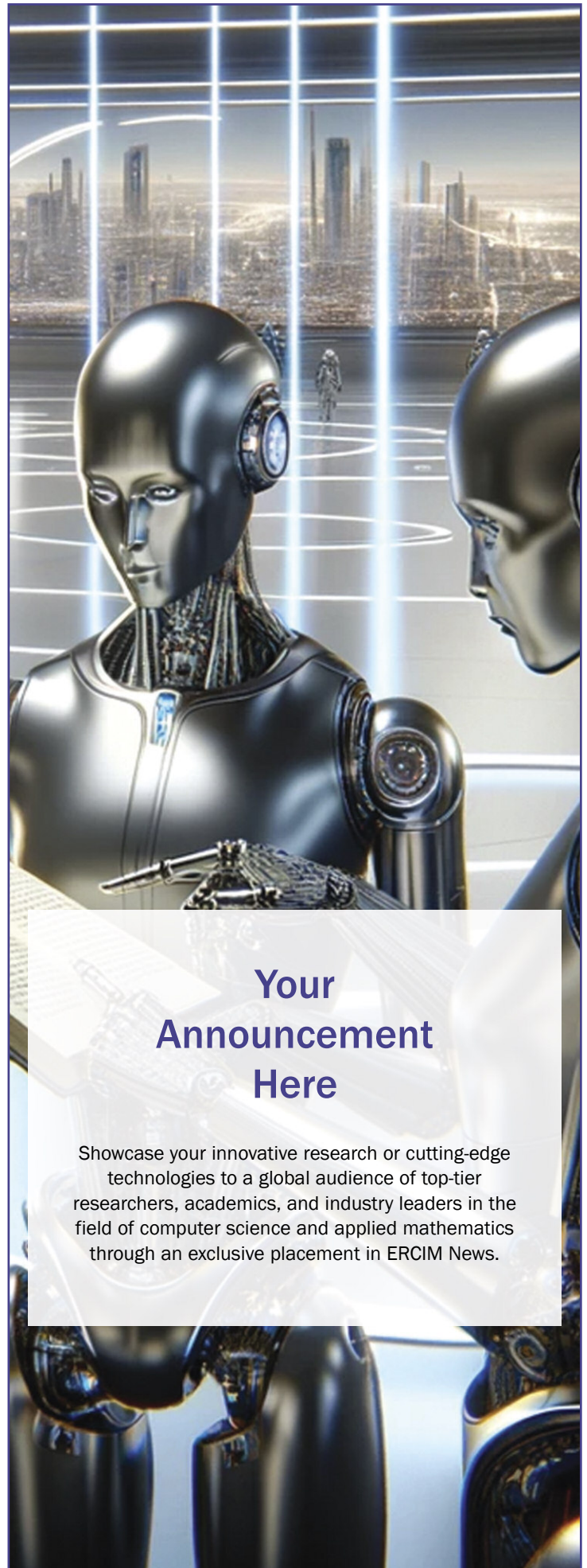
Evaluation of text generation systems is a very important, hard, and yet unsolved challenge. However, we must not forget that it took almost 200 years to define the notion of a metre (which is now based on the speed of light), and it took even longer to master the science of measuring temperature. Thus, it might also take a large amount of effort to measure the quality of general text generation. Our SNF project will provide one piece of the puzzle.

References:

- [1] P. von Däniken et al., “On the effectiveness of automated metrics for text generation systems,” in Findings of the Association for Computational Linguistics: EMNLP 2022. [Online]. Available: <https://aclanthology.org/2022.findings-emnlp.108/>
- [2] J. Deriu et al., “Correction of errors in preference ratings from automated metrics for text generation,” in Findings of the Association for Computational Linguistics: ACL 2023. [Online]. Available: <https://aclanthology.org/2023.findings-acl.404/>

Please contact:

Jan Deriu, Centre for Artificial Intelligence at the Zurich University of Applied Sciences, Switzerland
jan.deriu@zhaw.ch



Your Announcement Here

Showcase your innovative research or cutting-edge technologies to a global audience of top-tier researchers, academics, and industry leaders in the field of computer science and applied mathematics through an exclusive placement in ERCIM News.

Controllable Artificial Intelligence

by Peter Kieseberg, Simon Tjoa (St. Pölten UAS) and Andreas Holzinger (University of Natural Resources and Life Sciences)

The burgeoning landscape of AI legislation and the ubiquitous integration of machine learning into daily computing underscore the imperative for trustworthy AI. Yet, prevailing definitions of this concept often dwell in the realm of the abstract, imposing robust demands for explainability. In light of this, we propose a novel paradigm that mirrors the strategies employed in navigating the opacity of human decision-making. This approach offers a pragmatic and relatable pathway to cultivating trust in AI systems, potentially revolutionising our interaction with these transformative technologies.

During recent years, and especially since the stunning results Chat-GPT brought to the general public, the discussion on how to secure society against drawbacks and dangers resulting from the ubiquitous use of AI has gathered momentum. In addition, the AI Act by the European Union, which has been in the making for a while already, is on the home straight. It will not only set the stage for a common market regarding AI-driven systems in the EU but will also set the rules on putting AI into the market, especially concerning forbidden and high-risk AI applications. In order to mitigate potential risks of these applications, distributing companies will need to be able to conduct risk management and make their AI transparent.

In order to align AI development with these requirements, the term “trustworthy AI” was defined, although different expert groups arrived at different definitions. The two most noteworthy are those of the High-Level Expert Group (HLEG) [L1] of the EC and the NIST [1], which have a lot of overlap, but also some distinct differences. While the HLEG speaks about so-called “key requirements”, the NIST uses the notion of “characteristics”. Figure 1 gives a short overview on both definitions and outlines similarities between them and here you get an overview [2].

Both definitions work in the same direction, while the one devised by the HLEG also includes environmental and societal wellbeing as a requirement, which needs to be discussed, as this requirement is not a technical but a moral/philosophical one, and makes it hard to, e.g. define trustworthy military applications. Still, the major issue in both definitions is the requirement for explainability, as stated explicitly in the fifth characteristic in the NIS definition and several other requirements/characteristics, especially those concerning biases, non-discrimination, accountability and transparency. Also, the issue of providing security is difficult for many machine learning algorithms, as issues of emergence and non-explainability make standard tools for security testing difficult. Thus, developing trustworthy AI-based systems must be considered as difficult to impossible for many applications, especially those sporting large and complex networks with reinforcement components like a doctor in the loop in cancer detection. Still, these are the interesting applications for AI that will be the

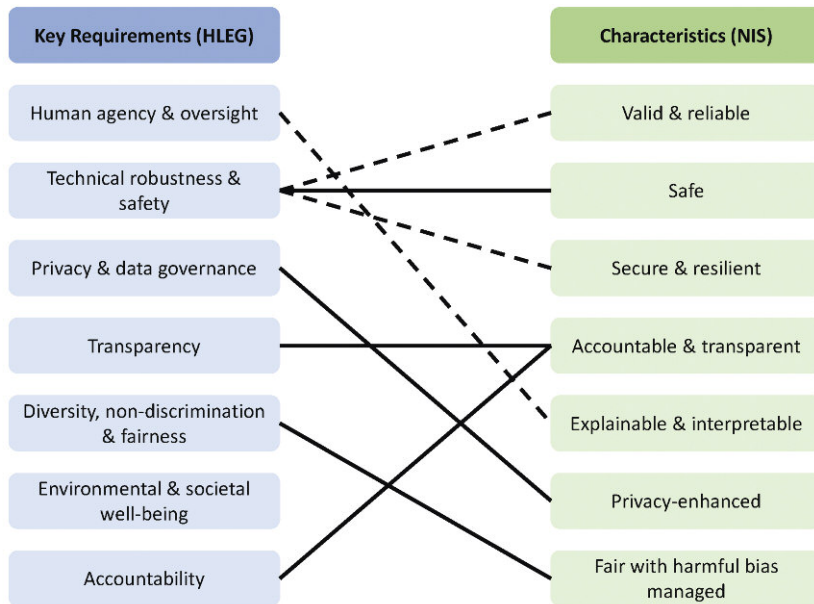


Figure 1: Comparing definitions for Trustworthy AI according to HLEG and NIST.

technological drivers for the coming years, yet the requirements for trustworthy AI are very demanding.

This is in stark contrast to how we deal with human decision-making. Typically, we are not able to explain our own decision-making process, as many decisions are derived subconsciously. Thus, we are actually asking more from AI than from humans, as for the latter, we typically only want to control whether a decision-making process is more or less correct and, if it is not, we want to make sure it is either corrected or the defaulting person faces consequences. This we can easily transform into the notion of “Controllable AI” [3] with the two main requirements:

1. We need to establish controls to detect when an AI system deviates too much from the target/typical behaviour.
2. We need to introduce methods for recalibrating or shutting down the system when such misbehaviour is detected.

Thus, the core idea behind Controllable AI is the assumption that no AI system should be deemed reliable, and it is essential to establish mechanisms for detecting malfunctions and re-asserting control. This definition does not call for explainability, as it is not required to be able to explain how and why a correct or incorrect conclusion was reached by the system; it only requires the detection of conclusions that are too far off the correct behaviour. As a comparison, many functions of the human digestive system are still not understood by science, and even less by the common person, but many severe malfunctions are easily detectable by everybody based on body reaction.

The methods required for achieving Controllable AI are diverse and need to be tailored to the specific system in question. Some examples for techniques that could be facilitated:

- Sanity Checks that bound the correct results and are obtained by knowledge or far simpler, explainable models.
- Critical data surveys that find biases and data poisoning in the training/reinforcement data instead of having to disassemble a complex model.
- Backdoors that allow control over a system that does not correctly react on input.
- Explicitly training the network for neglecting invalid decisions.

- Divine rules used for cross-checking the result of the decision-making process, e.g. rules that drop decisions that result in human harm, even if they are considered to be the best solution by the neural network.
- Non-AI alternatives for a process, or alternatives that, while providing less exact results, can be explained easily and in short time.

Furthermore, it is especially important to put the assumption of failure into the system design itself, i.e. Controllable AI is not a set of techniques that can be slapped onto the finished product, but constitutes a design philosophy for AI systems, much like privacy by design or cyber resilience.

In our future work, we plan to delve deeper into concrete measures and technologies for achieving Controllable AI, as well as providing respective design patterns. Furthermore, the concept will be included in our procurement guide for secure AI systems [L2], which is available for free.

Links:

- [L1] <https://kwz.me/hEH>
- [L2] www.secureai.info

References:

- [1] E. Tabassi, “Artificial Intelligence Risk Management Framework (AI RMF 1.0),” 2023.
- [2] A. Holzinger et al., “Information fusion as an integrative cross-cutting enabler to achieve robust, explainable, and trustworthy medical artificial intelligence,” *Information Fusion*, vol. 79, no. 3, pp. 263–278, 2022. doi:10.1016/j.inffus.2021.10.007
- [3] P. Kieseberg et al., “Controllable AI-An Alternative to Trustworthiness in Complex AI Systems?,” in *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*, pp. 1–12, Springer Nature, 2023.

Please contact:

Peter Kieseberg, St. Pölten University of Applied Sciences, Austria
peter.kieseberg@fhstp.ac.at

From Cultural Heritage Assets to Cultural Linked Data – the Case of the Archaeological Museum of Messara

by Dimitris Angelakis, Lida Charami, Pavlos Fafalios and Christos Georgis (FORTH-ICS)

The new Archaeological Museum of Messara, which opened its doors to visitors on 22 April 2023, is entirely dedicated to the antiquities of the Messara region in Heraklion, Greece. The Centre for Cultural Informatics of FORTH-ICS has undertaken the task to provide a comprehensive data management solution for the scientific and administrative documentation, research and promotion of the new museum’s important assets.

The Archaeological Museum of Messara [L1] is located around 50 kilometres south of the Heraklion city in the island of Crete in Greece. The museum is dedicated to the antiquities of the Messara area, aiming to highlight the unique character of this important region over time, from Prehistory to the Middle Ages. It includes around 900 square metres of exhibition halls, conservation workshops and an inner courtyard used for both exhibition and educational purposes. Its collection includes potteries, sculptures, figurines, metallic, clay and stone objects, inscriptions, seals, jewellery, glasses, frescoes, coins, bones, shells and organic material.

The museum’s data management needs include: a) the scientific documentation of the museum’s assets, which include not only objects but also monuments, archaeological sites, and early books/manuscripts; b) the management of the museum’s administrative material, such as the museum spaces, photographic documentations, condition assessments, and object conservations; and c) the management of presentation material required for dissemination and further data and knowledge ex-

ploitation, i.e. images, drawings, information texts, audio-visual material, etc.

The Data Management Workflow

To cope with these data management needs, the Centre for Cultural Informatics [L2] of FORTH-ICS has defined a workflow, accompanied by a set of IT tools, for the comprehensive management of the museum’s data: from documenting a set of cultural assets and disseminating them through the web, to integrating all data under a standard model and querying the integrated data for research and analysis purposes. The workflow, pictured in Figure 1, consists of the following main activities: i) scientific documentation, performed by the museum curators, ii) data access / dissemination, allowing the public to access the museum’s data, iii) ontology-based data modelling, for representing and integrating the data using a shared vocabulary in order to enable semantic interoperability and facilitate data reuse; iv) data mapping and transformation, for transforming the data to a rich semantic network of linked data that can be effectively queried; v) querying and analysis, performed by researchers and domain experts who study the cultural assets towards specific research objectives. Below we provide more information for each workflow component.

Data Documentation with Synthesis

For the scientific documentation activity, we have configured appropriately the Synthesis system [L3]. Synthesis is an information system for the collaborative documentation of data and knowledge in the cultural heritage and humanities fields. It is highly configurable for use in specific domains and data management needs, supports the parallel use of multiple languages for documentation, allows the recording of provenance information at data-element level, and makes use of standards for data organisation and representation. It supports different user roles (administrator, editor, guest) and utilises XML technology and a multi-layer architecture, offering flexibility and extensibility in terms of data structures and data types. The system has been widely used in different contexts, including cultural heritage [1] and historical research [2,3].

Data Dissemination with the Messara Museum Website

The website of Messara Museum [L1] is used as the main dissemination channel of the museum’s data. The website reads the Synthesis data at real-time, which means that any change

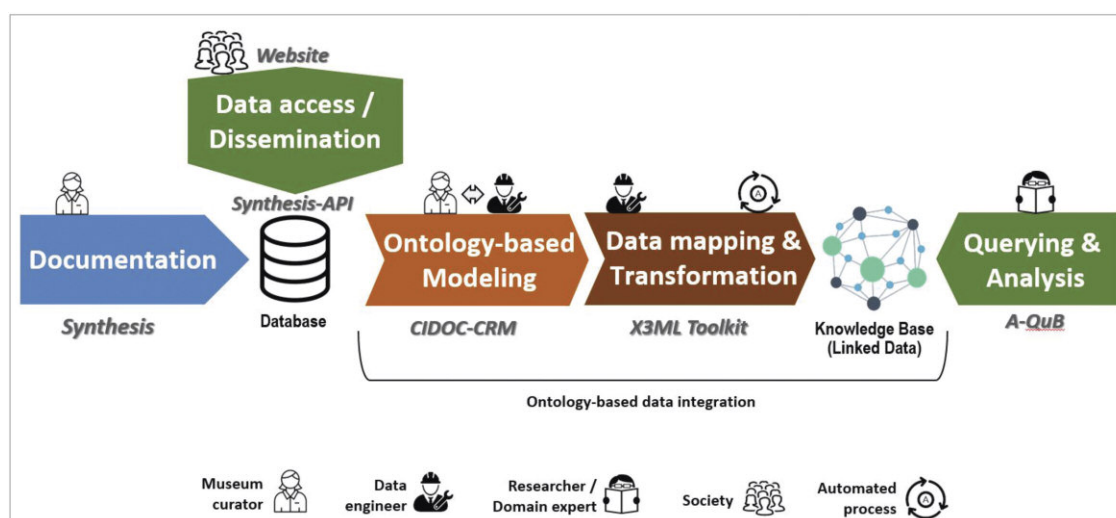


Figure 1: Workflow and IT tools for the holistic management of a museum’s data.

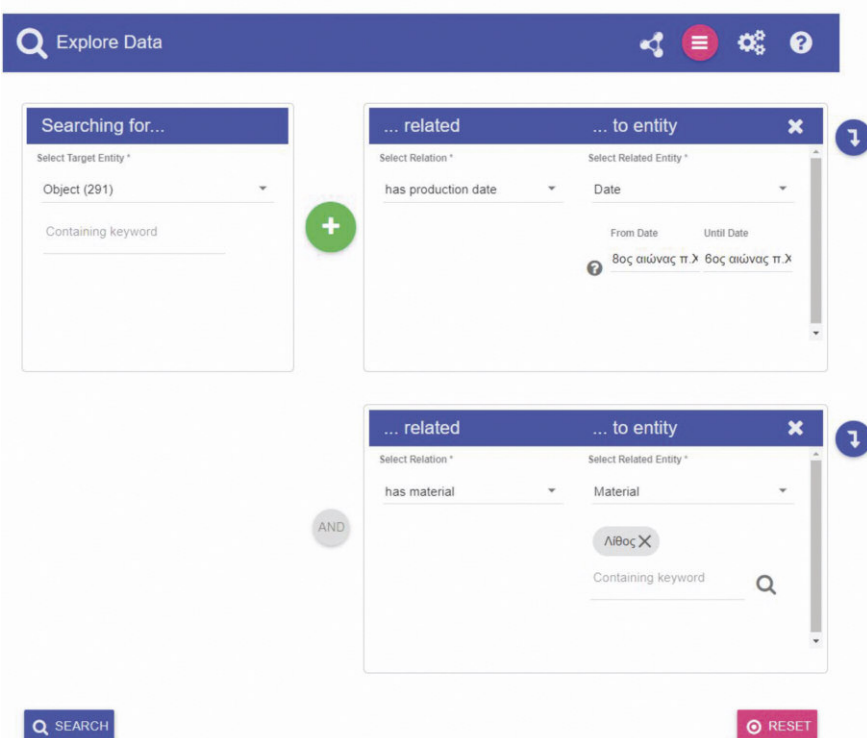


Figure 2: An example of query building in A-QuB-2.

Querying the integrated data with A-QuB-2

The CIDOC CRM-compliant knowledge graph can now be queried and analysed by researchers and domain experts. For this, we have deployed and configured appropriately the A-QuB-2 system [L6], a web application that facilitates the exploration of semantic (RDF) data by plain users. The application offers a user-friendly interface to build and execute complex queries step-by-step, while it also supports i) data filtering by text, date, number, Boolean value (true/false), and location (map area), ii) results browsing, iii) results export in CSV format for further offline data analysis, iv) query saving for future reuse. Figure 2 shows an example of query building in A-QuB-2 in which the user searches for all objects that have a) production date the period 8th-6th century BC, and b) stone as their construction material.

This work has co-financed by Greece and the European Union within the Operational Program “CRETE” (ESPA 2014-2020).

Contributors: Chrysoula Bekiari, Theodosia Bitzou, Maria Chalkiadaki, Konstantina Konsolaki, Kostas Petrakis, Giorgos Samaritakis, Eleni Tsoulouha (FORTH-ICS).

in the Synthesis database is directly reflected in the website. This is achieved through Synthesis-API, an application programming interface that provides a communication channel between the synthesis database and the website.

Data Integration with CIDOC CRM and the X3ML Toolkit

The objective here is to use a standard data model (a domain ontology) for modelling and representing the data as a rich semantic network of linked data (a knowledge graph). This process is important because it adds meaning to the data and increases their value, enables semantic interoperability (the ability of computer systems to exchange data with unambiguous, shared meaning), facilitates integration with similar external datasets that make use of the same data model, and supports advanced data querying, reasoning and inference making. In our case, we have used CIDOC CRM (Conceptual Reference Model) [L4] as the integration data model. CIDOC CRM is a high-level standard ontology (ISO 21127:2014) of human activity, things and events happening in space-time, providing a formal structure for describing concepts and relationships used in cultural heritage documentation. For transforming the Synthesis data to a CIDOC CRM compliant knowledge graph (in RDF format), we use the X3ML Toolkit [L5]. The toolkit offers a declarative language for defining data mappings (X3ML mapping definition language), a user interface for creating these mappings (3M Editor), software for executing the transformation and producing the semantic network (X3ML Engine), as well as a web application for visualising and validating the transformed data (RDF Visualizer).

Links:

- [L1] <https://messaramuseum.gr/>
- [L2] <https://www.ics.forth.gr/isl/centre-cultural-informatics>
- [L3] <https://www.ics.forth.gr/isl/synthesis-core>
- [L4] <https://cidoc-crm.org/>
- [L5] <https://www.ics.forth.gr/isl/x3ml-toolkit>
- [L6] <https://github.com/isl/A-QuB-2>

References:

- [1] C. Bekiari et al., “Building Comprehensive Management Systems for Cultural-historical Information,” CAA, Archaeopress, Oxford, 2012.
- [2] P. Fafalios et al., “Towards Semantic Interoperability in Historical Research: Documenting Research Data and Knowledge with Synthesis,” ISWC, Springer, 2021.
- [3] P. Fafalios et al., “Documenting Data and Knowledge about Feminisms and Politics in the Interwar Balkans with Synthesis-Core,” ERCIM News, no. 133, 2023.

Please contact:

Pavlos Fafalios, Centre for Cultural Informatics, Institute of Computer Science, FORTH-ICS (Heraklion, Greece)
fafalios@ics.forth.gr

Taranis AI: Applying Natural Language Processing for Advanced Open-Source Intelligence Analysis

by Florian Skopik and Benjamin Akhras (Austrian Institute of Technology)

Open-source intelligence (OSINT) provides up-to-date information about new cyber-attack techniques, attacker groups, changes in IT products, updates of policies, recent security events and much more. Often dozens of analysts search a multitude of sources and collect, categorise, cluster, and rank news items from the clear and dark web in order to prepare the most relevant information for decision makers. A tool that supports this job is “Taranis NG” from the Slovakian CERT. This solution ingests information from many types of sources such as websites, RSS feeds, emails and social media channels and makes them searchable. It also supports the creation of reports and daily summaries. However, the number of sources and news items is continuously growing, making it increasingly difficult to search them purely manually. These circumstances call for the application of novel natural language processing (NLP) methods to make OSINT analysis more efficient.

Open-source intelligence is the collection and analysis of data gathered from open sources to produce actionable intelligence [1]. The literature distinguishes at least technical Cyber Threat Intelligence (CTI), from tactical and strategic CTI. While technical CTI mainly includes simple data to configure security systems, such as indicators that are put into SIEMs, domain names for blocklists in name servers or proxies, and execution patterns to block malicious code in endpoint detection and response (EDR) solutions, tactical and strategic CTI is much “softer”. The latter usually consists of higher-level information, presented in natural language on various news sites and

security tickers, and includes quite diverse information about new threat actors, new (features of) security products, news about breaches, incidents, and campaigns, information about vulnerabilities, patches, mitigations, counter-measures, and exploitation. It also includes policy news, such as political and diplomatic initiatives, new EU policy documents, GDPR-related lawsuits, updates on security standards, mergers, acquisitions, failures, or other company-related news.

Gathering this mostly public information is essential to maintain situational awareness and take early actions in security matters. The typical OSINT workflow [1] foresees five phases: (i) collection, (ii) processing, (iii) analysis, (iv) production and dissemination, (v) direction and planning. In short, large organisations, national authorities, and analysis centres collect on a wide scale potentially hundreds of sources with thousands of articles daily, and analyse them for relevant content to create so-called products, which are essentially reports for certain constituencies that support decision-making processes. It is obvious that the quality of these reports highly depends on the level of sophistication of the analysis phase.

However, ingesting, analysing and making use of semantically richer “soft” CTI is much more demanding than ingesting well-structured machine-readable technical CTI. This soft CTI usually comes as unstructured freeform text, containing high-level, often ambiguous strategic information designed for human consumption – and indeed, in course of complex analysis workflows is usually consumed by human analysts only. This is tedious, resource-intensive, and error-prone work. The human element slows down the analysis process and tremendously hinders scalability. As the number of OSINT sources as well as the frequency of published articles rises, we need new analysis techniques to keep pace with these developments and to not miss any critical pieces of information. Luckily, natural language processing (NLP) and Artificial Intelligence (AI) have made tremendous progress in recent years.

In the course of our research, we explore five essential user stories together with our stakeholders from national authorities and CERTs that human OSINT analysts face in their daily work. Supporting these user stories with appropriate technical means is the goal of Taranis AI [L1]:

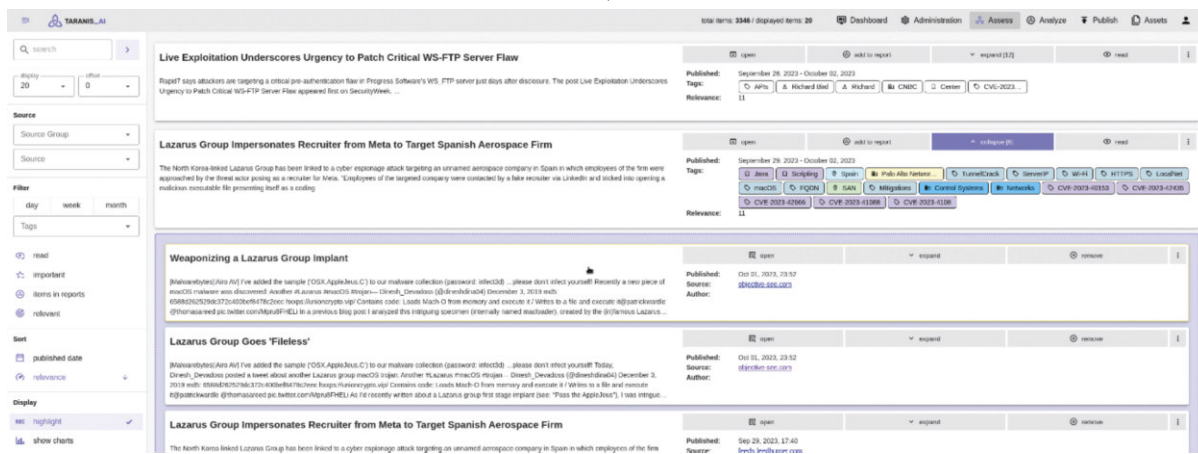


Figure 1: The AWAKE UI applying NLP for named entity recognition of news items, title creation for topic clusters, summary creation and story clustering.

- User Story 1: What was going on in the cyber security domain in the last 24 hours? (“Hot Topics Clustering”)
- User Story 2: What do we know about a specific entity? (E.g. a vulnerability, malware, company, product, person, etc.)
- User Story 3: I’ve read an interesting article. What further related news items exist?
- User Story 4: Which news items are recommended for me based on my recent preferences (collaboratively and AI-assisted)?
- User Story 5: I’d like to build a report for certain clients. How to sum up my findings efficiently?

Project Taranis AI

The CEF project AWAKE, as well as the recently funded research projects NEWSROOM and EUCINF of the European Defence Fund (EDF) assess the application of Taranis NG to improve cyber situational awareness through analysis of information from the clear and dark web. These projects specifically examine the integration of modern NLP methods into Taranis AI [3], a Taranis NG fork, that categorises news items using machine learning, extracts relevant entities, such as locations, people, company names, products, CVEs, attacker groups, and thus indexes and labels the content of items. This is also the basis for identifying relations and grouping news items about the same events – a crucial step to creating “stories” [2]. These stories allow human analysts to capture the most important current “hot topics” even more quickly and massively relieve them from the burden of combining or filtering redundant information from different sources. Additional features such as the automatic creation of summaries of reports and stories, and a collaborative ranking system round off the new features of Taranis AI. A screenshot showing the integration of these features into the Taranis AI user interface is depicted in Figure 1. The project Taranis AI [3] is open source and free to use under the EUPL.

Link:

[1] <https://taranis.ai/>

References:

- [1] Collaborative Cyber Threat Intelligence: Detecting and Responding to Advanced Cyber Attacks at the National Level, F. Skopik, Ed., CRC Press, 2017.
- [2] B. Liu, F. X. Han, D. Niu, L. Kong, K. Lai and Y. Xu, “Story forest: Extracting events and telling stories from breaking news,” ACM Transactions on Knowledge Discovery from Data (TKDD), vol. 14, no. 3, pp. 1–28, 2020.

Please contact:

Florian Skopik, AIT Austrian Institute of Technology, Center for Digital Safety & Security, Austria
florian.skopik@ait.ac.at

Tailoring Decarbonisation and Resilience Strategies to Drive Regional and Local Action

by Christiane Walter (PIK), Luis Costa (PIK) and Sara Dorato (T6)

LOCALISED is a four-year H2020-funded research project (October 2021 – September 2025) that develops, in a co-design process, tailored end-user products and services for local and regional policy-makers, administrations, businesses and citizens, to speed up sub-national decarbonisation processes while considering climate risks and adaptation needs. The flagship outputs are the Decarbonisation Profiler and the Net-Zero Business Consultant, providing information for all NUTS3-regions in Europe, currently under development.

Existing European information platforms supporting a shift to a net-zero energy system have so far been limited in providing information for promoting this transition and only target the national level. For this reason, LOCALISED is designed to close the information gap between national-level decarbonisation plans and local needs for planning and implementing an energy transition towards net-zero while accounting for remaining adaptation. The project develops tailored regional data on energy and climate impacts that are incorporated into information tools for regional governments, municipalities and businesses that support regional decarbonisation and adaptation to climate change. In doing so, the LOCALISED project contributes to the European Green Deal’s goal of a fair and equitable reduction of net greenhouse gas emissions by at least 55% by 2030 (compared to 1990 levels), leaving no one behind. Up to August 2023, LOCALISED has supported cities with information to fill in mandatory data for Sustainable Energy and Climate Action Plans (SECAPs); providing a database of potential mitigation as well as adaptation measures and suitable, related instruments; screening business vulnerabilities and potential technological business solutions; and integrating a strong social perspective in the tools to be developed.

A library of 54 decarbonisation pathways covering all EU Member States has been developed, based on the Long-Term Strategies by 2050 submitted by EU Member States. The library consists of suitable pathways allowing net-zero by 2050 covering multiple aspects of the energy/economy, including: vehicle fleet, residential building renovation, energy demand by carrier, energy production by technology, agriculture or industrial production. To facilitate data access, a database and API have been created, as well as a method for downscaling national-level energy models to NUTS3-regions. This combined methodology and tool will help to close the persistent gap between EU policies, national strategies and regional implementation of climate plans.

In addition, a decision tree methodology has been developed to classify climate actions as options, measures, or instruments: 314 adaptation and/or mitigation measures and 214 in-

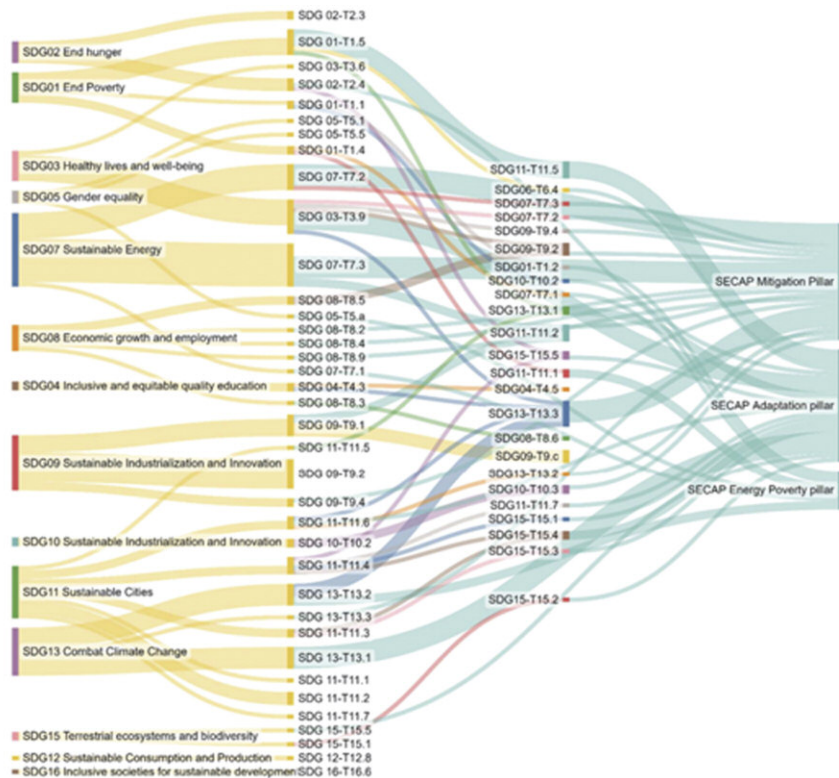


Figure 1: Relation between SDGs targets and SECAPs pillars. Source: own elaboration.

struments are unified in a database, allowing users to set multiple model/pathway constraints and receive suggestions for measures and instruments that are aligned with these constraints. The conceptualisation of a multi-criteria mitigation and adaptation solver was an important step towards making some of the LOCALISED tools a reality.

In parallel, a related set of 93 Sustainable Development Oriented Indicators (SOIs) has been introduced as a robust link between the targets in the Sustainable Development Goals (SDGs) and the requested information to create and monitor Sustainable Energy and Climate Action Plans (SECAPs). These established metrics allow reporting on both initiatives – the Green Deal and the UN SDGs – based on a single set of in-

dicators, optimising resources, allowing benchmarking and providing coherence.

However, many more outputs are envisaged. For example, in order to bring decarbonisation knowledge to end users, LOCALISED will develop two interactive web applications: Decarbonisation Profiler and Net-Zero Business Consultant. The Decarbonisation Profiler will practically support municipalities, regions, and their citizens in setting up and implementing ambitious decarbonisation plans, while the Net-Zero Business Consultant will provide businesses with insights into effective mitigation and adaptation options. The tools are developed in a co-design process with future users and selected experts, making sure the tools will be both effective and user-friendly. After a first phase of expert consultations, the first prototype is now being developed. We invite anyone who is interested in testing, giving feedback, and becoming an early-

adopter of the tools to get in touch with Tobias Gralke from the Climate Media Factory at: gralke@climatemedia.de.

Link:

<https://www.localised-project.eu/>

Please contact:

Christiane Walter, PIK, Potsdam, Germany
christiane.walter@pik-potsdam.de

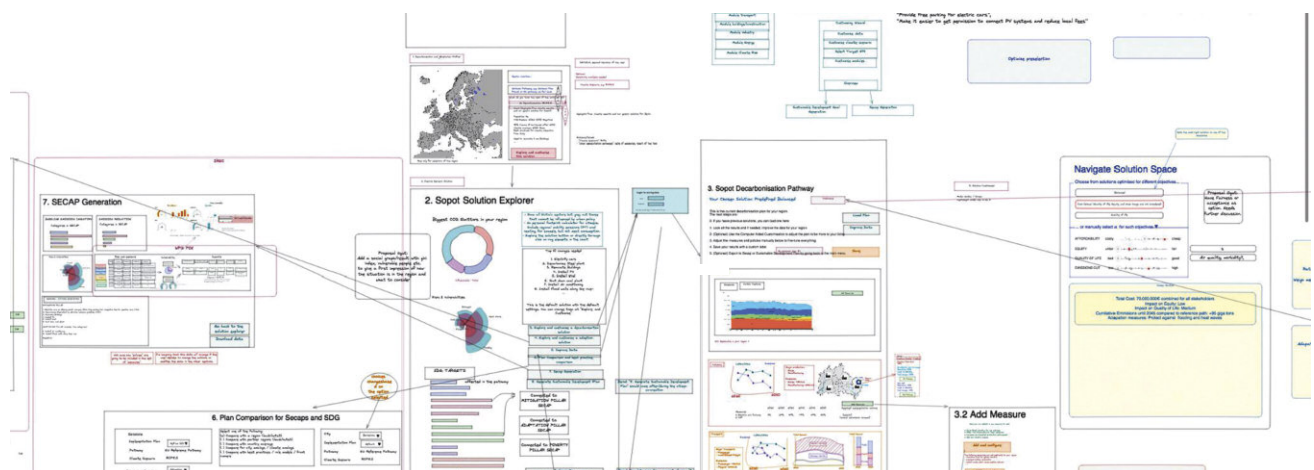


Figure 2: LOCALISED internal working document on Decarbonisation Profiler. Source: own elaboration.

Sponsored Contribution

TERMINET: Pioneering the Future of Smart Interconnected IoT



by Peter Kunz (ERCIM)

January 2024 marks the conclusion of the ambitious European H2020 project, TERMINET, which envisions a revolutionary next-generation IoT architecture. Combining state-of-the-art technologies like SDN, multiple-access edge computing, and virtualization, TERMINET aims to transform the IoT landscape with intelligent devices tailored for low-latency, market-driven use cases.

The project, a collaborative effort of 27 academic and industry partners, attracted third-party involvement in key R&D areas through open calls and a substantial cascading funding scheme. These areas include service deployment across edge computing, IoT nodes, SDN control enablement, security by design models, privacy enhancement through blockchain technologies, and the application of federated models and machine learning algorithms.

A notable achievement of TERMINET is its Integrated Platform, for the complex task of integrating systems in various domains. This integration leverages Continuous Integration/Continuous Delivery (CI/CD) and microservices approaches, ensuring flexibility and independence across services. The project also emphasizes the design of a novel orchestration framework for its MEC-based platform, addressing the convergence of communication and computing networks. This framework promises to optimise computational accuracy, delay, and energy use, leveraging mobile edge computing for next-generation IoT applications.

The project's results are validated through six diverse use cases, each showcasing TERMINET's versatility and impact across sectors. These use cases are:

- User-Centric Devices in Smart Farming
- Pathway of Personalised Healthcare
- Smart, Sustainable, and Efficient Buildings
- Prediction and Forecasting System for Optimising the Supply Chain in Dairy Products
- Group Training Surgery Using VR-enabled IoT Technologies
- Mixed Reality and ML Supported Maintenance and Fault Prediction of IoT-based Critical Infrastructure.

In the agricultural sector, TERMINET's "User Centric Devices in Smart Farming" use case explores IoT applications in live-stock management. By integrating sensors and wearables, this initiative enhances farm productivity, animal welfare, and environmental sustainability. TERMINET's technologies enable farmers to monitor and optimize agricultural processes, leveraging Federated Learning for decision-making.

TERMINET's Pathway of Personalised Healthcare use case showcases the project's impact on healthcare. It involves creating a comprehensive clinical dashboard aggregating data from various hospital departments, coupled with predictive models to assist doctors in decision-making. This initiative is supported by TERMINET's technologies like Edge-X TERMINET and Federated Learning framework, illustrating the synergy of edge computing and AI in healthcare.

Another innovative contribution comes from TERMINET's collaboration with Ericsson, focusing on predictive maintenance in optical modules. Utilizing federated learning, Ericsson could enhance predictive models for optical modules under harsh environmental conditions, demonstrating TERMINET's role in advancing AI applications while respecting data privacy.

TERMINET also addresses challenges like data integration and machine learning applications for predictive maintenance and resource optimization. Its edge computing capabilities play a crucial role in enabling real-time, efficient data processing. All this is relevant for "smart buildings".

The project's edge-based Federated Learning approach signifies a major step in IoT evolution. By embedding intelligence at the edge and utilizing Software-Defined Networking (SDN) and virtualized Multi-access Edge Computing (vMEC), TERMINET fosters rapid, optimized decision-making while ensuring data privacy and security.

TERMINET's mixed reality and machine learning use case, focuses on maintenance and fault prediction in critical infrastructure. This application demonstrates the project's capacity to blend reality with technology, improving operational efficiency and predictive capabilities in industrial settings.

Moreover, TERMINET's venture into blockchain technologies in IoT ecosystems underlines its commitment to data privacy, security, and immutability. The blockchain framework developed within the project ensures secure, private, and immutable transaction records, vital for all use cases involving sensitive data.

TERMINET's approach, spanning healthcare, agriculture, smart buildings, and network maintenance, exemplifies the transformative potential of IoT and AI. Its collaborative framework, integrating advanced technologies and diverse expertise, paves the way for a smarter, more connected future. The project has showcased innovative developments in the IoT sector, signalling a new era filled with interconnected and smart technologies that are transforming a wide range of industries.

ERCIM contributed its experience with the Web of Things as an RDF-based abstraction layer for the IoT, Linked Data, semantic modelling and cognitive approaches to IoT agents using chunks and rules.

More information: <https://terminet-h2020.eu/>

Call for Papers

Interdisciplinary Information Management Talks - IDIMT 2024

Hradec Kralove, Czech Republic, 4-6
September 2024

With over 30 years of history, IDIMT conferences have established themselves as an interdisciplinary international forum for exchanging concepts and visions in the areas of software-intensive systems, management and engineering of information and knowledge, social media, business engineering, and related topics. The conference has been organized since its inception by the University of Economics and Business in Prague and the Johannes Kepler University in Linz, Austria. The proceedings are published by Trauner Verlag, Linz, in the University Edition, Informatics series. Papers are peer-reviewed and indexed by Scopus and Web of Science. Past proceedings can be found at <https://idimt.org/proceedings/>.

The conference features 11 sessions. Session D, Autonomous Vehicles, may be of particular interest to ERCIM. It covers topics around autonomous vehicles, including UAVs; Smart City and Smart Road environments; Sustainable development goals and smart living. This session is co-organized by the ERCIM DES (Dependable Embedded SW-intensive Systems) Working Group, led by Erwin Schoitsch, Austria.

The conference will take place from September 4-6, 2024, in Hradec Kralove, Czech Republic. The deadline for paper submissions is April 14, 2024.

More information: <https://idimt.org/>.

Call for Papers

43rd SAFECOMP 2024, 19th DECSoS Workshop

Florence, Italy, 17-20 September
2024

The renowned international SAFE-COMP Conference will be held this year in Florence, Italy, from September 17-20, with the first day reserved for several parallel workshops (Workshop Day). The key theme is Safety in a cyber-physical interconnected world.

Established in 1979 by the European Workshop on Industrial Computer Systems, Technical Committee 7 on Reliability, Safety, and Security (EWICS TC7), SafeComp has contributed significantly to the progress in the dependable application of computers in safety-related and safety-critical systems. SafeComp is an annual event that covers the state-of-the-art, experiences, and new trends in the areas of safety, security, and reliability of critical computer applications.

The deadline for full paper submissions for the main conference is February 11, 2024. The deadline for workshop papers will be in early May 2024. The workshops will undergo an evaluation process, with details to be announced later on the SAFECOMP website.

Of particular interest to the ERCIM community might be the co-located 19th DECSoS workshop, which has been co-organized for many years by the ERCIM DES (Dependable Embedded SW-intensive Systems) Working Group (Erwin Schoitsch, Amund Skavhaug).

All papers, both for the SAFECOMP main conference and the workshops, are peer-reviewed by at least three independent reviewers and published by Springer in the LNCS series as SAFE-COMP and SAFECOMP Workshop Proceedings (two independent volumes).

More information:
<https://www.safecomp2024.unifi.it/>.



SCHLOSS DAGSTUHL
Leibniz-Zentrum für Informatik

Call for Proposals

Dagstuhl Seminars and Perspectives Workshops

Schloss Dagstuhl – Leibniz-Zentrum für Informatik is accepting proposals for scientific seminars/workshops in all areas of computer science, in particular also in connection with other fields.

If accepted, the event will be hosted in the seclusion of Dagstuhl's well known, own, dedicated facilities in Wadern on the western fringe of Germany. Moreover, the Dagstuhl office will assume most of the organisational/ administrative work, and the Dagstuhl scientific staff will support the organizers in preparing, running, and documenting the event. Thanks to subsidies the costs are very low for participants.

Dagstuhl events are typically proposed by a group of three to four outstanding researchers of different affiliations. This organizer team should represent a range of research communities and reflect Dagstuhl's international orientation. More information, in particular details about event form and setup, as well as the proposal form and the proposing process, can be found on

<https://www.dagstuhl.de/dsproposal>

Schloss Dagstuhl – Leibniz-Zentrum für Informatik is funded by the German federal and state government. It pursues a mission of furthering world class research in computer science by facilitating communication and interaction between researchers.

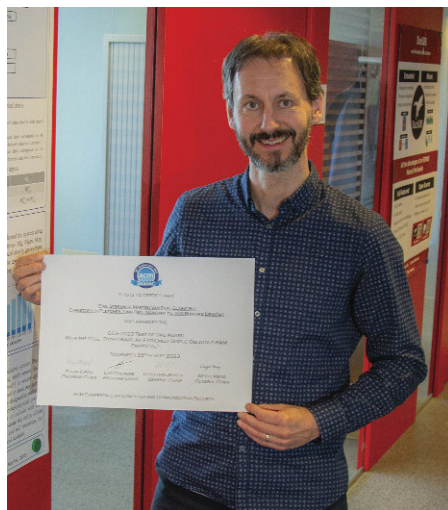
Important Dates

- *Next submission period:*
April 1 to April 15, 2024
- *Seminar dates:*
Between April 2025 and February 2026 (tentative).

CWI Received Two Test of Time Awards

Marten van Dijk, head of CWI's Computer Security group, won the ACM CCS 2023 Test-of-Time Award, and members of CWI's Database Architectures group won the 2024 CIDR Test of Time Award.

Last year, the ACM CCS conference occurred from November 26-30, 2023, in Copenhagen. The SIGSAC CCS Test-of-Time papers are recognized for their lasting impact on systems security and privacy, including opening new research directions, proposing new technologies, making significant discoveries, improving security risk understanding, releasing valuable resources, and more. Marten van Dijk and his co-authors were honored for their paper on Path ORAM, a simple yet effective Oblivious RAM protocol requiring minimal client storage. Due to its simplicity, Path ORAM is currently the most practical ORAM scheme known for small client storage. CWI's Database Architectures group received a Test-of-Time Award for their paper on vectorized execution, a technique that significantly enhances database query performance. Unlike traditional systems where database operators work on individual records, vectorized execution processes small columnar blocks called 'vectors', making analytical systems at least ten times faster. This influential paper, published in 2005, was authored by Peter Boncz, Marcin Zukowski, and Niels Nes.



Marten van Dijk at CWI with the 2023 ACM SIGSAC CCS Test of Time Award certificate.

More information: <https://www.cwi.nl/en/news/>



Peter Boncz, Niels Nes and Marcin Zukowski (2nd, 3rd and 4th from left) with the CIDR Test of Time Award.



Horizon Europe Project Management

A European project can be a richly rewarding tool for pushing your research or innovation activities to the state-of-the-art and beyond. Through ERCIM, our member institutes have participated in more than 100 projects funded by the European Commission in the ICT domain, by carrying out joint research activities while the ERCIM Office successfully manages the complexity of the project administration, finances and outreach.

Horizon Europe: How can you get involved?

The ERCIM Office has recognized expertise in a full range of services, including:

- Identification of funding opportunities
- Recruitment of project partners (within ERCIM and through our networks)
- Proposal writing and project negotiation
- Contractual and consortium management
- Communications and systems support
- Organization of attractive events, from team meetings to large-scale workshops and conferences
- Support for the dissemination of results.

Please contact:

Peter Kunz, ERCIM Office
peter.kunz@ercim.eu



ERCIM – the European Research Consortium for Informatics and Mathematics is an organisation dedicated to the advancement of European research and development in information technology and applied mathematics. Its member institutions aim to foster collaborative work within the European research community and to increase co-operation with European industry.



ERCIM is the European Partner of the World Wide Web Consortium.



Consiglio Nazionale delle Ricerche
Area della Ricerca CNR di Pisa
Via G. Moruzzi 1, 56124 Pisa, Italy
www.iit.cnr.it



Norwegian University of Science and Technology
Faculty of Information Technology, Mathematics and Electrical Engineering, N 7491 Trondheim, Norway
<http://www.ntnu.no/>



Centrum Wiskunde & Informatica

Centrum Wiskunde & Informatica
Science Park 123,
NL-1098 XG Amsterdam, The Netherlands
www.cwi.nl



RISE SICS
Box 1263,
SE-164 29 Kista, Sweden
<http://www.sics.se/>



Fonds National de la Recherche Luxembourg

Fonds National de la Recherche
6, rue Antoine de Saint-Exupéry, B.P. 1777
L-1017 Luxembourg-Kirchberg
www.fnr.lu



SBA Research gGmbH
Floragasse 7, 1040 Wien, Austria
www.sba-research.org/



Foundation for Research and Technology – Hellas
Institute of Computer Science
P.O. Box 1385, GR-71110 Heraklion, Crete, Greece
www.ics.forth.gr



SIMULA
PO Box 134
1325 Lysaker, Norway
www.simula.no



Eötvös Loránd Research Network
Számítástechnikai és Automatizálási Kutató Intézet
P.O. Box 63, H-1518 Budapest, Hungary
www.sztaki.hu/



Fraunhofer ICT Group
Anna-Louisa-Karsch-Str. 2
10178 Berlin, Germany
www.iuk.fraunhofer.de



University of Cyprus
P.O. Box 20537
1678 Nicosia, Cyprus
www.cs.ucy.ac.cy/



INESC
c/o INESC Porto, Campus da FEUP,
Rua Dr. Roberto Frias, n° 378,
4200-465 Porto, Portugal
www.inesc.pt



UNIVERSIDAD DE MÁLAGA

Institute for Software Engineering and Software Technology
“Jose María Troya Linero”, University of Malaga
Calle Arquitecto Francisco Peñalosa, 18, 29010 Málaga
<https://gp.uma.es/itis>



Institut National de Recherche en Informatique
et en Automatique
B.P. 105, F-78153 Le Chesnay, France
www.inria.fr



University of Warsaw
Faculty of Mathematics, Informatics and Mechanics
Banacha 2, 02-097 Warsaw, Poland
www.mimuw.edu.pl/



I.S.I. – Industrial Systems Institute
Patras Science Park building
Platani, Patras, Greece, GR-26504
www.isi.gr



VTT Technical Research Centre of Finland Ltd
PO Box 1000
FIN-02044 VTT, Finland
www.vttresearch.com