Trustworthy AI Summit 2025

TAIS-25

Posters – Book of abstracts

This document gathers all abstracts of posters displayed during the Summit. Most of them are in form of texts, in some cases the missing text is replaced by a copy of the poster.

This is a draft with some known imperfections, for example page numbers are missing and some formatting needs to be done. The final version will be made available on tha HAL open archive platform.

The posters are grouped by topic corresponding to the display area in the hall.

Embedded AI

Foundations of Trustworthy AI

Generative AI

Human-Machine Teaming

Robustness

Verification and Trust

Cybersecurity

A Case Study for Embedded and Certified Aeronautical Systems using AIDGE, an Open and Sovereign ML Tool

Filipo Studzinski Perotto, Anthony Fernandes Pires, Jean-Loup Farges, Youcef Bouchebaba, Mohammed Belcaïd, Benjamin Lesage, Claire Pagetti (ONERA-DTIS)

Abstract
This work presents the first results of the use of AIDGE ML platform on an aeronautical use case: *Airborne Collision Avoidance System for Unmanned Aircraft* (ACAS-Xu), where a neural network is trained as a surrogate model for a large look-up table, comparing performance obtained with other platforms, and sketching the first steps toward the certification of this kind of ML-based solution.

Keywords
Machine Learning, Embedded Systems, Certification, ACAS-Xu.

Introduction

AI was developed in recent years based on giant neural networks and massive data processing, but today the challenge is to transpose these solutions into small components as close as possible to industrial constraints. Two important challenges are frugality and certificability, both necessary to allow the use of AI in embedded critical systems and infrastructures.

The DeepGreen project, funded by the French Research Council, is developing a new open-source software platform for embedded AI called AIDGE[1], designed to learn *Deep Neural Networks* (DNN), transform, and generate optimized code for target hardware architectures, in a completely transparent, open, reproducible, deterministic, controllable, and traceable way, helping to avoid dependence on opaque and non-sovereign software, ensuring good performance, and favoring the cerficability of systems conceived with *Machine Learning* (ML).

In this context, the capabilities of AIDGE to produce an embedded and certifiable version of a Neural Network need to be evaluated (2). Our contribution to the project is defining how the tool can generate the code (in C language) implementing the ML model for a particular use case, the ACAS-Xu system, ensuring to support certification following the latest guidelines provided by EASA. As a preliminary result, we present a comparison between the performance of ACAS-Xu from code generated with different ML platforms.

ACAS-Xu

ACAS-Xu is the new generation of Airborne Collision Avoidance System for fixed-wing unmanned aircraft (1). The function of that system is to generate resolution advisories (evasive actions) in case of imminent risk of collision between two airplanes. The system interrogates the transponders of nearby aircraft to determine the distance, altitude, and bearing of surrounding traffic. The resolution policy has been optimized through dynamic programming, represented within a very large lookup table: 600 million parameters (q-values) for the single horizontal resolution case. In the literature, a strategy to reduce its memory footprint is using ML to approximate the q-function from the table data with a neural network.

However, an important aspect to make possible the use of a DNN version of ACAS-Xu in real embedded systems is the certification of the neural networks (4), (3). In this presentation, we show the preliminary results of performance analyses for different implementations of the ACAS-Xu DNN with the ones obtained with AIDGE. We also introduce the roadmap developed by ONERA and CS-Group within the DeepGreen project in order to use the explicit intermediary representation models that can be extracted anytime from AIDGE at each step of the optimization and transformation of a given neural network to the target code to support certification following the development objectives stated by the EASA concept paper (5).

[1] `https://projects.eclipse.org/projects/technology.aidge`

Certification Qualities and Objectives

The EASA defined a number of guidelines for the certification of ML-based systems in safety-critical systems. While those are not yet regulatory requirements, they highlight a number of desirable qualities from ML-related tooling. Based on the examination of the certification objectives in (5), we identified some basic characteristics necessary for any ML tool aiming to produce certified software: determinism, traceability, reproducibility, formalization.

If the poster is accepted for presentation, the idea is to present how dose general qualities can be enhanced by the ML tool, Aidge. Finally, the EASA document proposes 129 objectives to what a certifiable solution must be compliant. In this work, we will present a first assurance case designed to answer to the following objective: the applicant should describe the MI architecture.

Preliminary Performance Evaluation

In preliminary tests, starting from the same neural network, with 6 fully-connected layers of 50 neurons and ReLU activation, represented as an ONNX file, we ran 30000 inferences using different instances, some compiled from generated code from different platforms, others directly using runtime tools, some for CPU, others using GPU (Nvidia T4). Those initial tests produced interesting results from the version compiled from Aidge generated C++ code.

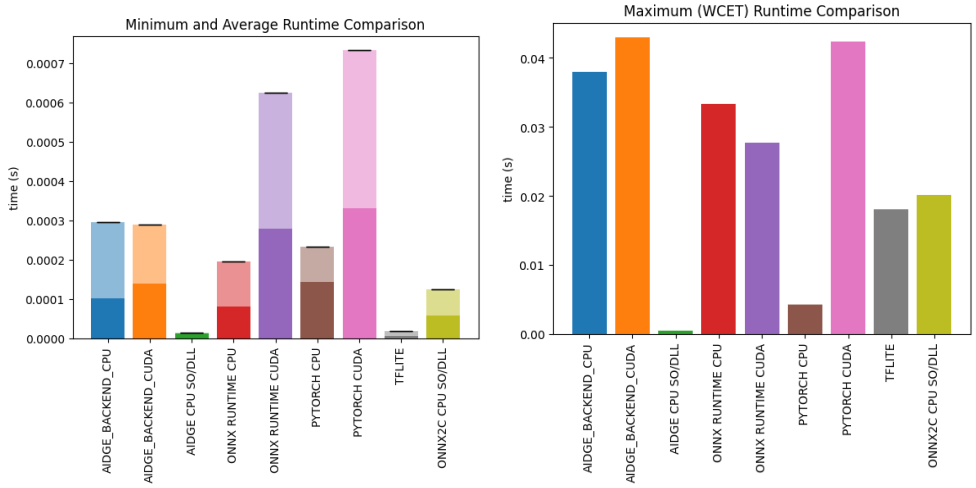


Figure 1: At left, average and best inference times using different code generation or runtime tools, and at right, the worst-case elapsed times.

References:

(1) Perotto. 2024. Using Deep RL to Improve the ACAS-Xu Policy: Concept Paper. In Proc. of the 2nd Int. Conf. on Cognitive Aircraft Systems: ICCAS. INSTICC, SciTePress, 110–117.

(2) Perotto et al. 2024. Thinking the certification process of embedded ML-based aeronautical components using AIDGE, a French open and sovereign AI platform. In Proc. of the 2nd Int. Conf. on Cognitive Aircraft Systems: ICCAS. INSTICC, SciTePress, 64–71.

(3) Gabreau et al. 2022. Toward the certification of safety-related systems using ML techniques: the ACAS-Xu experience. In Proc. of 11th European Congress on Embedded Real Time Software and Systems, ERTS.

(4) Damour et al. (2021). Towards Certification of a Reduced Footprint ACAS-Xu System: a Hybrid ML-based Solution. In Computer Safety, Reliability, and Security (40th SAFECOMP), pages 34–48. Springer.

(5) EASA (2024). Concept Paper: First Usable Guidance for Level 1 and 2 Machine Learning Applications, n.2. Eu- ropean Aviation Safety Agency (EASA), Cologne.

OVERITY-AI: MLOps oriented framework for trustworthy embedded ML

Florian Dupeyron*, Samy Chehade*, Esteban Pustoc'h*, Pierre Romet†‡, Jean-François Béraud‡
*ELSYS Design
†CIAD UMR 7533, Belfort Montbéliard University of Technology, UTBM, Belfort, France
‡Advans Lab
florian.dupeyron@elsys-design.com

***Abstract*—This paper introduces OVERITY-AI, a MLOps oriented framework designed to optimize, deploy, evaluate and compare ML applications on resource-constrained edge devices (EdgeAI/TinyML), targetting industrial and critical systems. We present the key principles and design requirements of the workflow, methodology, and tools employed. Upcoming challenges are also discussed.**

I. Introduction

EdgeAI is currently experiencing unprecedented growth across all critical sectors including healthcare, automotive, and industrial IoT[1]. However, deploying and monitoring these applications on constrained embedded devices in a trustworthy perspective poses a significant technical challenge, as engineering teams shall conciliate performance, precision and cybersecurity requirements, while limiting resource consumption to ensure frugality, as well as being able to assess compliance to regulatory requirements in normative contexts. Faster time-to-market and adoption of Dev(Sec)Ops practices leading to MLOps[2, 3] create the need for automated and integrated solutions. OVERITY-AI (*Optimization, VERification, Integration, Test and Yield of AI infused systems*) is designed to enable trustworthy ML deployment on embedded targets, bridging research-oriented ML practices with production-ready industrial requirements through an end-to-end MLOps oriented workflow to complete existing solutions.

II. Workflow requirements and design

The introduced framework is built around the following essential requirements: (1) *Reproducibility*: the workflow must be capable of capturing essential information — *e.g.* (hyper-)parameters, random seeds, dataset distribution — to reproduce results on a given target. ; (2) *Traceability*: a key component of configuration management in industrial and critical embedded systems is the ability to trace engineering choices back to high-level requirements. All steps, artifacts and intermediate results should be recorded to enable end-to-end traceability on a specific outcome ; (3) *In-situ*: as embedded systems often rely on physical and real-time constraints, the framework should give the necessary tools to assess performance directly on the execution targets, with the ability to monitor both execution metrics as well as physical KPIs — energy and memory consumption, CPU load, specific I/O monitoring ; (4) *Domain-specific*: the workflow should allow the integration of domain-specific methodologies, encompassing *explainable AI* (xAI) and *interpretable machine learning* (iML) stakes ; (5) *Collaboration*: Finally, the developed tool should seamlessly integrate with decentralized practices from the open-source community, promoting open research and facilitating the common capitalization of developed intelligence, while preserving sovereignty and confidentiality of critical assets.

Building upon these requirements and inspiration from existing MLOps frameworks, a three-phase workflow is presented.

a) Training and Optimization (T/O): takes an input model or ML method and produces an optimized version to be deployed on the embedded target. It may use datasets, simulations using digital twins or post-training optimization techniques.

b) Deployment and Measure/Qualification (D/MQ): involves deploying the model on the execution target, via a minimal application referred to as an *inference agent*, and evaluate its performance against a set of user-defined requirements for a given set of physical stimuli (through its testing setup) or dataset. This involves asessing ML metrics (*e.g.* accuracy, F1-score), system specific parameters (*e.g.* inference latency, memory and energy consumption), or application specific KPIs ; these measures are then compared to target values to ensure the model meets desired performance and safety specifications. Using a real test setup allows to assess some niche-case scenario, for instance side-channel attacks on power-supply while using the model.

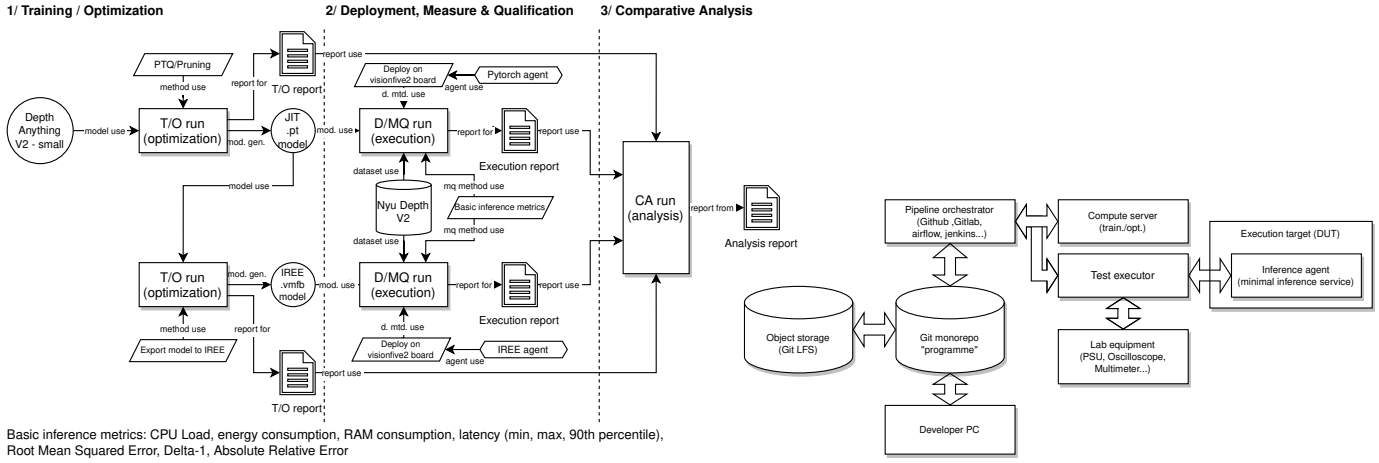
c) Comparative Analysis: allows to compare multiple optimization experiments to select the best solution for a given use-case. This ecompasses various benchmarking scenario — *e.g.* hyper-parameter choices, or testing multiple execution targets for a given model.

Finally, an *experiment* is defined as a sequence of T/O, D/MQ and CA steps, with a fixed set of parameters, to allow reproducibility.

III. Tools and methods

a) Programming language: Python is ubiquitous for ML applications. Therefore, it serves as the main language for this framework. Each step (T/O, D/MQ, CA) leverage python scripts called *methods* as entry-points, with an *Application Programming Interface* (API) tracking parameters, consumed artifacts (datasets, models), and output metrics.

b) Infrastructure: The infrastructure currently in use is shown in Fig. 1b. Heavy computational loads are offloaded to a computation server, while testing is



(a) Example workflow traceability graph

(b) Simplified infrastructure description

Figure 1: Graphical summary of the developed tools

performed directly on the target using a *Hardware-in-the-loop* (HIL) test bench. This setup allows to assess the model considering real-time execution constraints and external I/O, or physical events that may occur on the target computing system.

c) ML Tools: Reusability is a key requirement for OVERITY-AI. Existing code written using mainstream AI and ML tools such as Scikit, PyTorch, Tensorflow or Keras should be used with minimal modification, as well as allowing to include domain-specific tools for xAI/iML. ONNX is the preferred model exchange format.

d) Execution runtimes: Similarly, no custom execution runtime is currently developed, one should be able to use the best fitting one for its target. However, a specific focus is placed on IREE[1], which leverage LLVM's MLIR[4], allowing a fine-grained control over the execution chain, making it especially appealing to resource-constrained applications and for control flow graph based optimizations.

e) Versioning: Versioning is currently managed through *git*, utilizing *git lfs* for large binary files such as models and datasets. A *programme* is a repository workspace containing assets for a project, including *methods* (T/O, D/MQ, CA), *inference agents*, models, datasets, and *reports*. Inspired by the *monorepo* culture[5], each commit thus represents a snapshot of the entire project at any given time. This approach firstly provides a coherent workspace for engineering teams with minimal engineering effort. Future versions will integrate other storage and versioning solutions.

f) Reports and traceability graph: Each step generates a report containing: *context information*, *i.e.* input parameters, random seeds, *etc.* ; *environment information*, *i.e.* used machine, package versions, *etc.* ; *key metrics and log information*. A traceability graph links artifacts with integrity hashes. This enables tracing a ported application back to its original input data and optimization/validation steps. Fig. 1a shows an example traceability structure for a comparative benchmark on different execution runtimes for a pruned and quantized DepthAnything v2 small[6] model, deployed on a StarFive's VisionFive2 RISC-V board, and currently being assesed using OVERITY-AI.

[1] https://iree.dev

IV. Conclusion and upcoming work

The OVERITY-AI framework introduces a workflow suited to port, assess and deploy ML models on constrained targets, leveraging MLOps practices and HIL testing. Its modular approach seamlessly integrate into existing engineering methodologies by allowing usage of mainstream ML tools, but adding required end-to-end traceability to track ported applications.

The primary goal is to validate the tool's effectiveness and refine its design through real-world applications. Subsequent integration into comprehensive MLOps frameworks will facilitate the deployment of a holistic ecosystem, ultimately enhancing the end-to-end integration of trustworthy ML applications in embedded industrial and critical systems.

References

[1] X. Wang et al. "A Survey on Trustworthy Edge Intelligence: From Security and Reliability To Transparency and Sustainability." In: *IEEE Communications Surveys & Tutorials* (2024), pp. 1–1. ISSN: 2373-745X. DOI: 10.1109/comst.2024.3446585.

[2] F. Bayram and B. S. Ahmed. "Towards Trustworthy Machine Learning in Production: An Overview of the Robustness in MLOps Approach." In: *ACM Computing Surveys* 57.5 (Jan. 2025), pp. 1–35. ISSN: 1557-7341. DOI: 10.1145/3708497.

[3] D. Kreuzberger, N. Kühl, and S. Hirschl. *Machine Learning Operations (MLOps): Overview, Definition, and Architecture*. 2022. DOI: 10.48550/ARXIV.2205.02302.

[4] C. Lattner et al. "MLIR: Scaling Compiler Infrastructure for Domain Specific Computation." In: *2021 IEEE/ACM International Symposium on Code Generation and Optimization (CGO)*. IEEE, Feb. 2021, pp. 2–14. DOI: 10.1109/cgo51591.2021.9370308.

[5] R. Potvin and J. Levenberg. "Why Google stores billions of lines of code in a single repository." In: *Communications of the ACM* 59.7 (June 2016), pp. 78–87. ISSN: 1557-7317. DOI: 10.1145/2854146.

[6] L. Yang et al. *Depth Anything V2*. 2024. DOI: 10.48550/ARXIV.2406.09414.

An Agentic Framework for Stable and Interpretable Causal Discovery using Semantic Clustering

Louis Hernandez[1,2], Alessandro Leite[2], Cecilia Zanni-Merk[2], Matthieu Boussard[1]

[1]Craft AI, Paris, France
[2]LITIS, INSA Rouen Normandie, France

Abstract

A key challenge in creating Trustworthy AI is discovering underlying causal mechanisms to ensure systems are robust and their reasoning is interpretable. While Large Language Models (LLMs) show promise for this task, current discovery methods using them often suffer from instability and scale poorly. We propose a novel agentic framework to mitigate these issues by separating stable semantic representation from high-level reasoning. Our system uses sentence-transformer embeddings for hierarchical clustering and deploys an LLM-powered agent to interactively generate robust, high-level causal constraints, providing an efficient and stable path to interpretable causal models.

1 Introduction

Causal graphs are powerful tools for describing data-generating processes and modeling interventions [3]. However, obtaining them is a significant challenge. While traditional causal discovery methods exist, constraint-based approaches struggle with high-dimensional data, and score-based methods can be computationally prohibitive [4].

Recent work has explored using LLMs for causal discovery, but this often involves direct pairwise queries on variables, an approach that scales poorly with $O(n^2)$ complexity and suffers from the inherent stochasticity and instability of LLM outputs [1, 2]. Furthermore, many of these methods depend on large, proprietary models, limiting accessibility.

To address these challenges, we introduce a novel **agentic framework** for causal discovery. Our primary contribution is a system that leverages an LLM not for low-level fact retrieval, but for high-level reasoning over semantically coherent clusters of variables. This layered architecture cleanly separates stable semantic representation, interactive exploration by an agent, and final statistical estimation. By doing so, it prunes the search space for discovery algorithms, reduces reliance on massive LLMs, and increases the overall stability and interpretability of the discovery process.

2 Methodology: An Agent-driven Process

Our framework formalizes causal discovery as a four-stage pipeline, orchestrated by a stateful agent implemented using LangGraph.

2.1 Semantic Representation & Hierarchical Clustering

Given a set of observable variables $V = \{v_1, \ldots, v_n\}$, we first create dense vector representations for each variable based on its name and description using a sentence-transformer embedding model $\varphi : V \to \mathbb{R}^d$. We then apply agglomerative hierarchical clustering to the set of embeddings $\{\varphi(v)\}_{v \in V}$ to produce a dendrogram T. This tree structure provides a multi-resolution grouping of variables based on semantic similarity. Each internal node u in the dendrogram defines a cluster $C(u)$ containing all variables in its subtree.

2.2 The Agentic Interaction Loop

The core of our framework is an agentic loop where an LLM (e.g., Qwen) interactively navigates the cluster hierarchy to generate causal hypotheses. At any step t, the system state is a partition $\mathcal{P}_t$ of the variables into a set of clusters S_t from the dendrogram.

At each step, the agent receives a detailed prompt summarizing the current state, including the variables in each cluster, previously identified causal constraints, and a list of valid actions. Based on this context, the agent invokes one of three tools:

- **`Split(u)`**: If a cluster $u \in S_t$ is deemed too heterogeneous, the agent can split it into its direct children in the dendrogram.
- **`Merge(u_1, u_2)`**: If two clusters $u_1, u_2 \in S_t$ are siblings in the dendrogram and semantically similar, the agent can merge them into their parent node.
- **`TestCausal(u,v)`**: The agent queries the LLM to hypothesize a causal relationship between two clusters $u, v \in S_t$. Positive results are recorded as directed cluster-level relations, $u \to v$.

This process is dynamic; the agent's decisions are not pre-programmed but are made in response to the evolving state of the system, guided by the goal of forming meaningful clusters and identifying causal links between them.

2.3 From High-Level Constraints to Forbidden Edges

The set of cluster-level causal relations, $\mathcal{R}$, generated by the agent are translated into hard constraints for a downstream

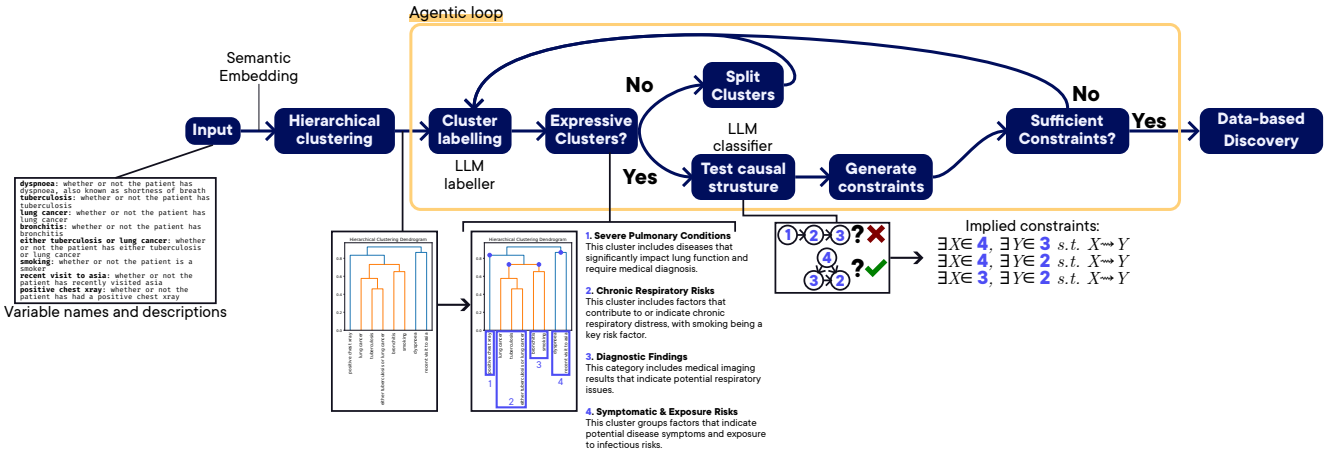


Figure 1: An overview of the proposed frameword

discovery algorithm. For each cluster-level edge $u \to v \in \mathcal{R}$, we assume that no variable in the effect-cluster $C(v)$ can be a cause of any variable in the cause-cluster $C(u)$. This generates a set of forbidden variable-level edges F, where the existence of an edge $x \to y$ is forbidden for all $x \in C(v)$ and $y \in C(u)$.

2.4 Constrained Causal Structure Learning

Finally, we use the observational data matrix $\mathcal{D}$ to learn the structure of the ground truth DAG G^*. We employ DAGMA, a continuous optimization-based discovery algorithm, which solves for the weighted adjacency matrix W of the graph [5]. Our key integration is adding the set of forbidden edges F as hard-zero constraints to the optimization problem:

$$\begin{aligned} \min_{W} \quad & \mathcal{L}(W;\mathcal{D}) + \lambda\|W\|_1 \\ \text{subject to} \quad & h(W) = 0, \\ & W_{ij} = 0 \; \forall (v_i, v_j) \in F. \end{aligned}$$

Here, $\mathcal{L}(W;\mathcal{D})$ is the statistical loss function and $h(W) = 0$ is the acyclicity constraint, also defined in [5].

3 Framework Evaluation

The framework's performance is validated across several key dimensions of causal discovery using standard synthetic benchmarks (ASIA, CHILD) and datasets from the Root Cause Analysis (RCA) literature.

Framework Stability. To assess stability, a core design goal, we evaluate the framework's robustness to minor, semantically equivalent variations in the input variable descriptions. This tests the hypothesis that the combination of deterministic embeddings and high-level LLM reasoning is resilient to superficial changes in the input.

Structural Accuracy. The accuracy of the final learned graph is measured against the ground truth using standard metrics, including Structural Hamming Distance (SHD), Precision, Recall, and F1-score. This evaluates the quality of the constraints generated by the agentic process.

Efficacy in Root Cause Analysis. For scenarios mirroring Root Cause Analysis (RCA), we evaluate the framework's ability to correctly identify the true root cause of simulated system faults on labeled datasets.

4 Discussion

The primary contribution of this work is the agentic framework itself. It offers a path toward more trustworthy AI by enhancing several key aspects of causal discovery.

Interpretability. The process is transparent. The agent reasons about named, meaningful clusters of variables, and the generated constraints are explicit and understandable.

Stability. By design, our architecture mitigates the known instability of LLMs. It uses them for high-level guidance on stable, semantically-defined clusters rather than for noisy, low-level pairwise variable judgments.

Efficiency and Accessibility. The hierarchical pruning of the search space reduces the computational cost of the final discovery phase. The framework is designed to be effective with smaller, more accessible LLMs, avoiding a dependency on large, proprietary models.

Future extensions could adapt the agent's capabilities for more complex scenarios, such as time-series data, by integrating richer forms of constraints.

References

[1] E. Kiciman, R. Ness, A. Sharma, and C. Tan. Causal reasoning and large language models: A survey. *arXiv preprint arXiv:2305.00050*, 2023.

[2] J. Long, T. Schuster, and A. Piché. Can large language models build causal graphs? *arXiv preprint arXiv:2303.05279*, 2023.

[3] J. Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, 1995.

[4] P. Spirtes, C. N. Glymour, and R. Scheines. *Causation, prediction, and search*. MIT press, 2000.

[5] X. Zheng, B. Aragam, P. Ravikumar, and E. P. Xing. Dags with no tears: Continuous optimization for structure learning. In *Advances in neural information processing systems*, volume 31, 2018.

Fairness in intersectional setups : Aggregation choice and some paradoxes

Jeanne Monnier *jeanne.monnier@orange.com*
Orange Research

Thomas George *thomas.george@orange.com*
Orange Research

Abstract

In the quest to design trustworthy AI systems, making fair decisions is a major key. Fairness has to be evaluated in intersectional setup –when individuals can belong to multiple protected subgroups– to reflect reality. Intersectionality requires to make design choices about how we compare and aggregate pairs of subgroups. In this work, we provide a list of possible aggregation choices and clarify their implications. We also demonstrate paradoxes that can occur when optimizing for global fairness results in individual subgroups receiving worse treatment, showing all the complexity of this design issue.

1 Problem definition

For a classifier M learned using a machine learning algorithm, fairness is usually measured by choosing a metric (e.g. the probability of being correctly classifier, or the probability of being assigned the positive outcome) that quantifies how much its prediction differs depending on whether an instance belongs to a protected group or not. As an illustrative case, we focus on the Equalized Odds metric $\mathrm{EO}(M) = P(M(x) = 1|Y = 1, A = 1) - P(M(x) = 1|Y = 1, A = 0)$ that quantifies the difference in the odds of obtaining the positive outcome $M(x) = 1$ provided that the true class Y is 1, depending on the protected attribute A.

The above definition is given in the binary case where $A \in \{0, 1\}$, which does not account for the possibility of being subject to multiple sources of unfair treatment. This is covered in the intersectional setup, where A is now a vector of the size of the number of protected attributes, which can take as many values $A_1, A_2, ...A_k$ as the number of subgroups. This raises the question of *how to aggregate the differences between subgroups* in a global scalar fairness score, our focus in this work. Similarly to how different fairness metrics result in different assessments of whether a model is fair, the choice of aggregation method is essential, and the subtle differences between aggregation methods are often overlooked.

As a first design choice, in the **one-vs-all** approach, we form the vector of differences between all $\frac{k(k-1)}{2}$ possible pairs of subgroups $\mathrm{EO}_{\text{o-v-all}}(M, i, j) := P(M(x) = 1|Y = 1, A = A_i) - P(M(x) = 1|Y = 1, A = A_j)$, whereas in the **one-vs-mean** approach [e.g. in 1], we instead use the k vector of differences to the averaged outcome $\mathrm{EO}_{\text{o-v-mean}}(M, i) := P(M(x) = 1|Y = 1, A = A_i) - P(M(x) = 1|Y = 1)$. Note that the same methodology could be adopted with the vector of ratios instead of differences.

2 Aggregation methods

Once the **one-vs-all** or **one-vs-mean** approach chosen, there are multiple ways to aggregate the multiple differences obtained to produce a single concise score of fairness, independent of the number of subgroups taken into account. This score will be used to decide whether the model can be considered fair or not, regarding a selected threshold; or to compare different models in terms of fairness performances. For simplicity, we will express all aggregation methods with the **one-vs-all** approach.

1. **Worst case based methods**: attention is focused on the fairness metrics applied to the most discriminated subgroup, i.e. $\mathrm{EO}_{\mathrm{worst}}(M) := \min_{i\in\mathbb{N}}\{P(M(x)=1|A=A_i,Y=1)\}$, that accounts for all other subgroups as it is a lower bound over all other subgroups odds by construction. The priority is to avoid a few rare but very poor results being overshadowed by a majority of good results.

   Worst case can be the score itself, or could be contrasted with the best case in a min-max difference or ratio [3, 2]:

$$\mathrm{EO}_{\text{min-max}}(M) := \frac{\min_{i\in\mathbb{N}}\{P(M(x)=1|A=A_i,Y=1)\}}{\max_{i\in\mathbb{N}}\{P(M(x)=1|A=A_i,Y=1)\}}$$

2. In $\gamma-SP$ subgroup fairness [5], the $\mathrm{EO}_{\text{o-v-mean}}$ difference is weighted by the mass of each subgroups before applying a threshold γ. $M(x)$ is $\gamma-SP$ subgroup fair if :

$$\mathrm{EO}_{\gamma-\mathrm{SP}}(M) := \max_{i\in\mathbb{N}} |P(M(x)=1 \mid Y=1) - P(M(x)=1|A=A_i,Y=1)| \times P(A=A_i) \leq \gamma$$

   In practice, pathological cases of subgroups with too little data to obtain a significative probability are eliminated.

3. **Probabilistic methods**: Probabilistic unfairness [6] takes as score the probability $\mathrm{EO}_{\epsilon-\mathrm{proba}}(M) := P(\mathrm{EO}_{\text{o-v-all}}(M,i,j)>\epsilon)$, i.e the probability defined on the set of all pairs of subgroups that $\mathrm{EO}_{\text{o-v-all}}$ exceeds a fixed threshold ϵ. It can be thought of as an opposite to worst case methods, since large differences are allowed, as long as enough subgroup pairs (defined by ϵ) satisfy the criterion.

4. **Norm Evaluation** Taking the vector of differences $\mathrm{EO}_{\text{o-v-all}}(M,i,j)$, the norm of this vector can be calculated to obtain a single score. ℓ_1, ℓ_2 or $\ell_p \forall p \in \mathbb{R}$ norms can be used, each providing a different balance between subgroups.

$$\mathrm{EO}_{\ell_p}(M) := \|\mathrm{EO}_{\text{o-v-all}}\|_p = \left(\sum_{(i,j)} |\mathrm{EO}_{\text{o-v-all}}(M,i,j)|^p\right)^{1/p}$$

5. **Mutual Information**: Coming from information theory, this measure captures the dependence between two random variables. Some works [4, 7] take inspiration of this tool to aggregate the multiple probabilities $P(M(x)=1|Y=1,A=A_i)$ in one score:

$$I_{Y=1}(\mathbf{A};\hat{Y}) = \sum_{\mathbf{a}\in\mathcal{A}}\sum_{\hat{y}\in\hat{\mathcal{Y}}} P_{Y=1}(\mathbf{a},\hat{y}) \log\left(\frac{P_{Y=1}(\mathbf{a},\hat{y})}{P_{Y=1}(\mathbf{a})P_{Y=1}(\hat{y})}\right)$$

   Mutual information between two random variables equals zero if and only if these two random variables are statistically independent which means that $I_{Y=1}(A;\hat{Y}) = 0 \Leftrightarrow P(\hat{Y}|A,Y=1) = P(\hat{Y}|Y=1)$ and $I_{Y=1}(A;\hat{Y})$ reflects fairness results across subgroups.

3 Implications and paradoxes

While they all take as input the same probabilities, the scores derived from the various aforementioned aggregation methods do not carry the same information. Worst case based method might not achieve the best fairness performance reachable on some pairs of subgroups, but they avoid very problematic results. On the contrary, a good score in term of ℓ_1 norm assures high fairness performances across the majority of subgroups couples but might leave some hidden bad performances. Using a ℓ_p norm on the vector of differences, the greater p, the more the bigger values of differences will weight in the score.

Choosing one aggregation method rather than another implies a completely different distribution of results across all the sub-groups considered. Some paradoxes can even happen when reducing the global fairness according to one method leads to increasing "unfairness" towards one or several specific subgroups. A whole study can be conducted to show the impact of the aggregation choice and to highlight different concrete cases where previously mentioned paradoxes happen.

References

[1] Jaewoong Cho, Gyeongjo Hwang, and Changho Suh. "A fair classifier using kernel density estimation". In: *Advances in neural information processing systems* 33 (2020), pp. 15088–15099.

[2] Avijit Ghosh, Lea Genuit, and Mary Reagan. "Characterizing intersectional group fairness with worst-case comparisons". In: *Artificial Intelligence Diversity, Belonging, Equity, and Inclusion*. PMLR. 2021, pp. 22–34.

[3] Rashidul Islam et al. "Differential fairness: an intersectional framework for fair AI". In: *Entropy* 25.4 (2023), p. 660.

[4] Jian Kang et al. "Infofair: Information-theoretic intersectional fairness". In: *2022 IEEE International Conference on Big Data (Big Data)*. IEEE. 2022, pp. 1455–1464.

[5] Michael Kearns et al. "Preventing fairness gerrymandering: Auditing and learning for subgroup fairness". In: *International conference on machine learning*. PMLR. 2018, pp. 2564–2572.

[6] Mathieu Molina and Patrick Loiseau. "Bounding and approximating intersectional fairness through marginal fairness". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 16796–16807.

[7] Jeanne Monnier et al. "A Generic Framework for Bias Evaluation and Mitigation for Fair AI". In: (2024).

Memetic Semantic Boosting for Symbolic Regression

Alessandro Leite[1] & Marc Schoenauer[2]
[1]INSA Rouen Normandie, LITIS UR 4108, France
[2]Inria Saclay, LISN, Université Paris-Saclay, France

Abstract

Symbolic regression (SR) seeks to discover human-readable mathematical expressions that explain observed data. We present a novel boosting framework based on Memetic Semantic Genetic Programming (MSGP) that produces interpretable and accurate models. Our method integrates semantic backpropagation, local search, and linear scaling into weak learners that are assembled via boosting. Experiments on real-world datasets demonstrate that our approach achieves state-of-the-art performance while producing concise and interpretable expressions, making it well-suited for trustworthy AI applications.

1 Introduction

For a given dataset (X, y), symbolic regression (SR) aims to find a function $f(X) : \mathbb{R}^n \mapsto \mathbb{R}$ that represents the underlying relationship between the input features (X) and an output (y). In recent years, Genetic Programming (GP) [1] has attracted increasing attention in machine learning due to its ability to evolve both model structure and parameters without prior assumptions about the data [2, 3]. The symbolic nature and flexible representation of its solutions enable the modeling of complex data relationships. These properties position GP as a potential alternative to neural networks. Traditional GP-based SR approaches rely solely on a program's final output to evaluate performance, often ignoring the semantics of intermediate subtrees [4, 5]. However, incorporating semantic information can guide the search toward generalizable and simpler expressions. Furthermore, small changes in a GP solution may significantly alter fitness, thereby hindering search efficiency. Memetic algorithms (MAs) [2] provide an effective way to compensate for the capability of global exploration of general evolutionary methods with the increased exploitation that can be obtained through local search. They combine population-based evolutionary algorithms and individual-based local search strategies. This work introduces a method, *semantic boosting regression (SBR)* [6], which enhances GP with semantic guidance and memetic local search, and wraps it in a boosting framework to improve robustness and accuracy. Our SBR approach combines a set of semantic learners trying to improve the generalization performance. Experimental results on various real-world benchmark datasets show that our proposed methods can have equal or better performance compared to state-of-the-art (SOTA) non GP-based (e.g., Decision Tree and Random Forest [7]), GP-based methods (e.g., GP-GOMEA [8], `gplearn` [9]), and SyRBO [10], which is a GP-based boosting method.

2 Memetic Semantic Boosting

Although semantic backpropagation (SB) [5] and memetic algorithms have each been applied to symbolic regression, their combination offers enhanced interpretability and generalization. Semantic backpropagation facilitates the generation of programs approximating the desired output, while memetic algorithms refine these programs by exploiting subtree semantics. However, SB often leads to excessive tree growth (bloat), increasing evaluation cost and reducing generalization. We address this by incorporating linear scaling (LS) [11], allowing SB and memetic refinement to focus on structural aspects, while LS optimizes coefficient scaling.

Given a dataset composed of N independent samples (X_i) with m independent input variables $(X_i = [x_{i,1}, x_{i,2}, \ldots, x_{i,m}])$ and a corresponding target output (y_i), the task of symbolic regression is to find a tree $(\mathcal{T}(.))$ that minimizes the distance between its outputs and the target output (y) [13, 14]. Such a tree $\mathcal{T}(.)$ is built from a set of predefined functions and a set of terminals (a.k.a. the input variables and ephemeral constants). For instance, using the mean squared error (MSE) as the distance metric (a.k.a. fitness function) for $\mathcal{T}(.)$, and denoting $\hat{y}$ the outputs of tree $\mathcal{T}$, the task of symbolic regression is to find a tree $\mathcal{T}(.)$ that minimizes $MSE(\mathcal{T})$.

Algorithm 1 Semantic Boosting Regression Training

Require: fitness cases $(\mathbf{X}, y)$
Require: stages {number of boosting stages}
Require: kwargs {arguments of MSGP}

boosters ← []
for $i \leftarrow 1$ to stages **do**
    boosters[i] ← MSGP(kwargs).train(X,y) [12]
    y ← y - boosters[i].predict(X){updates the target values to the remaining residuals [12]}
end for
$\mathcal{T}$ ← join_trees(boosters){concatenates (i.e., sum) the trees of all boosters}
$\mathcal{T}$ ←LS($\mathcal{T}$, y) {computes the coefficients of the tree $\mathcal{T}$}
return $\mathcal{T}$

The SBR algorithm (Alg. 1) includes a set of Memetic Semantic GP for Symbolic Regression (MSGP) models [12] as weak learners. Hence, training a model comprises fitting an MSGP model and updating the target values for the next stage to be the residuals of the current model (Line 4), as usually done by traditional gradient boosting algorithms. The output model is a large tree made of the sum of each individual learner tree after undergoing a global linear scaling phase (Line 7). The prediction phase then simply applies the trained model to the fitness cases.

3 Experimental Results

We evaluate the proposed semantic boosting regression algorithm on different real-world regression dataset benchmarks, having heterogeneous features and samples. Moreover, These datasets are standard in GP literature GP [15, 8, 16] as overfitting the training set occurs either when complex models are learned or when models are built using discontinuous functions. As baselines, we consider both evolutionary and non-evolutionary algorithms: Decision Tree (DT) and Random Forest (RF) [7]) for non-evolutionary approaches, and Gene-pool Optimal Mixing Evolutionary Algorithm (GP-GOMEA) [8] and `gplearn` [9] as the GP-based approaches. Fig. 1 shows that the SBR algorithm consistently outperforms MSGP and Symbolic-regression boosting (SyRBO) across most datasets, particularly with five and ten boosting stages. While SyRBO achieves competitive accuracy, it often generates excessively large expressions (Fig. 2a), likely due to its reliance on `gplearn` as a weak learner. In contrast, SBR produces much more compact symbolic expressions (Fig. 2b and Tab. 1).

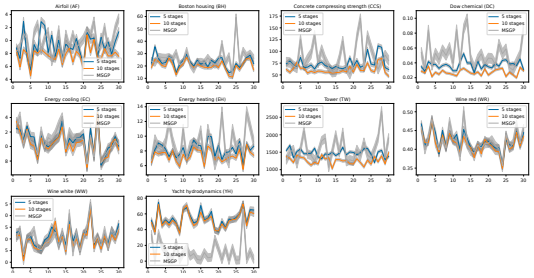
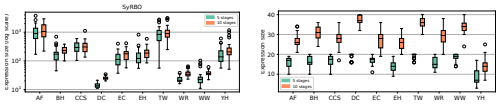


Figure 1: Accuracy of semantic boosting regression on the test set across 30 independent runs, using one MSGP, five, and ten boosters for each benchmark dataset. The shaded area represents the 95% confidence interval.

4 Conclusion

Symbolic regression aims to model data through interpretable mathematical expressions, often using Genetic Programming (GP) due to its flexible, assumption-free nature. However, GP-based methods frequently suffer from overfitting and bloated expressions. This paper presents a memetic semantic boosting algorithm that combines population-based search with semantic-guided strategies to generate compact yet accurate models. Experimental results demonstrate that our method matches or exceeds the performance of both evolutionary and traditional machine learning approaches on diverse real-world datasets.



(a) SyRBO expression sizes (y-axis in logarithmic scale)

(b) SBR expression sizes

Figure 2: Sizes of expressions outputted by SBR and SyRBO for each benchmark dataset.

Table 1: Examples of the best expressions found by SBR for each benchmark dataset.

DS	5 stages	10 stages
AF	$-0.006411 * X1 * X3 - 0.006411 * X1 * X5 + 0.006411 * X2 - 0.006411 * X4 + 127.71$	$-0.006711 * X1 * X3 - 0.006711 * X1 * X5 - 0.006711 * X2 - 0.013422 * X3 \div X4 + 0.006711 * X4 + 127.257606 - 0.006711 * X2 \div X1 - 0.013422 * X3 \div X1$
BH	$5.17 * X1 * X4 + 10.86 + 5.17 * X6 \div X13 + 5.17 * X6 \div X11 + 5.17 * X12 \div X10$	$-0.14 * X1 - 0.14 * X11 \div X4 + 0.14 * X13 - 0.14 * X3 - 0.14 * X5 \div X9 - 0.14 * X8 + 26.14 - 0.14 * X6 \div X3 - 0.14 * X6 \div X11 - 0.14 * X13 \div X10$
CCS	$0.026 * X1 \div X4 + 0.026 * X2 + 0.026 * X4 \div X8 + 0.026 * X5 * X8 + 0.026 * X8 + 25.40 + 0.026 * X8 \div X5$	$0.021 * X1 - 0.01 * X2 + 0.01 * X3 \div X5 + 0.01 * X4 + 0.02 * X4 \div X8 + 0.03 * X5 * X8 + 21.17$
DW	$-0.01 * X16 + 0.01 * X29 + 0.01 * X30 \div X4 + 0.01 * X49 + 0.017203 * X55 - 7.772227 + 0.01 * X55 \div X35$	$-9.3e-5 * X16 \div X7 - 9.3e-5 * X18 - 9.3e-5 * X18 \div X19 - 9.3e-5 * X29 - 9.3e-5 * X32 - 9.3e-5 * X35 - 9.3e-5 * X36 - 9.3e-5 * X36 \div X53 - 9.3e-5 * X49 + 3.16 - 9.3e-5 * X9 \div X37 - 9.3e-5 * X5 \div X30 - 9.3e-5 * X47 \div X16$
EC	$1.55 * X1 * X2 + 1.55 * X1 \div X3 - 762.86 + 1.55 * X5 \div X1$	$0.10 * X1 * X2 + 0.15 * X1 \div X3 + 0.05 * X3 * X7 + 0.05 * X4 * X5 + 0.05 * X4 * X7 - 80.28 + 0.05 * X5 \div X1$
EH	$-0.12 * X1 * X2 - 0.24 * X1 \div X5 + 83.96 - 0.12 * X7 \div X4$	$0.19 * X1 * X2 + 0.19 * X1 \div X3 + 0.09 * X2 \div X4 + 0.09 * X3 * X7 + 0.09 * X5 + 0.09 * X7 - 87.27 + 0.09 * X7 \div X5$
TW	$-5.86 * X1 - 5.86 * X1 \div X23 - 5.86 * X2 \div X25 - 5.86 * X23 \div X5 + 5.86 * X6 + 27.51 - 5.86 * X6 \div X4$	$-2.54 * X1 + 2.54 * X1 \div X3 - 2.54 * X13 + 2.54 * X15 + 2.54 * X24 + 2.54 * X4 \div X7 + 2.54 * X5 + 2.54 * X8 - 76.27 + 2.54 * X8 \div X22 + 2.54 * X16 \div X13 + 2.54 * X2 \div X1 + 2.54 * X6 \div X1$
WR	$0.36 * X10 * X5 + 0.36 * X11 \div X2 + 0.36 * X2 \div X4 + 2.28$	$-0.007 * X10 * X5 - 0.003 * X11 - 0.003 * X11 \div X9 - 0.007 * X3 - 0.003 * X7 * X8 - 0.003 * X9 + 5.90 - 0.003 * X9 \div X2$
WW	$0.01 * X1 \div X4 + 0.01 * X11 * *2 + 0.02 * X2 \div X6 + 4.39 + 0.01 * X2 \div X11$	$0.11 * X10 + 0.11 * X11 + 0.11 * X2 + 0.11 * X2 \div X4 + 0.11 * X4 \div X7 + 0.11 * X5 * X9 + 0.11 * X9 + 3.76 + 0.11 * X9 \div X6 + 0.11 * X2 \div X11 + 0.11 * X6 \div X11 + 0.11 * X2 \div X1$
YT	$85.80 * X6 * *2 + 86.77$	$18.12 * X2 \div X3 + 72.49 * X6^2 + 67.801051$

References

[1] J. R. Koza, *Genetic Programming: On the Programming of Computers by means of Natural Evolution*. Massachusetts: MIT Press, 1992.

[2] Y.-S. Ong, M.-H. Lim, F. Neri, and H. Ishibuchi, "Special issue on emerging trends in soft computing: memetic algorithms," *Soft Computing*, vol. 13, no. 8, pp. 739–740, 2009.

[3] S.-M. Udrescu and M. Tegmark, "AI Feynman: A physics-inspired method for symbolic regression," *Science Advances*, vol. 6, no. 16, p. eaay2631, 2020.

[4] N. F. McPhee, B. Ohs, and T. Hutchison, "Semantic building blocks in genetic programming," in *European Conference on Genetic Programming*, pp. 134–145, 2008.

[5] T. P. Pawlak, B. Wieloch, and K. Krawiec, "Semantic backpropagation for designing search operators in genetic programming," *IEEE Transactions on Evolutionary Computation*, vol. 19, no. 3, pp. 326–340, 2014.

[6] A. Leite and M. Schoenauer, "Memetic semantic boosting for symbolic regression," *Genetic Programming and Evolvable Machines*, vol. 26, no. 1, p. 11, 2025.

[7] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[8] M. Virgolin, T. Alderliesten, C. Witteveen, and P. A. Bosman, "Improving model-based genetic programming for symbolic regression of small expressions," *Evolutionary computation*, vol. 29, no. 2, pp. 211–237, 2021.

[9] T. Stephens, "Genetic programming in python with a scikit-learn inspired API: gplearn." github.com/trevorstephens/gplearn, 2016.

[10] M. Sipper and J. H. Moore, "Symbolic-regression boosting," *Genetic Programming and Evolvable Machines*, vol. 22, pp. 357–381, 2021.

[11] M. Keijzer, "Improving symbolic regression with interval arithmetic and linear scaling," in *European Conference on Genetic Programming*, pp. 70–82, 2003.

[12] A. Leite and M. Schoenauer, "Memetic semantic genetic programming for symbolic regression," in *European Conference on Genetic Programming*, pp. 198–212, 2023.

[13] M. Schmidt and H. Lipson, "Distilling free-form natural laws from experimental data," *Science*, vol. 324, no. 5923, pp. 81–85, 2009.

[14] M. F. Korns, "A baseline symbolic regression algorithm," in *Genetic Programming Theory and Practice X*, pp. 117–137, Springer, 2013.

[15] J. F. B. Martins, L. O. V. Oliveira, L. F. Miranda, F. Casadei, and G. L. Pappa, "Solving the exponential growth of symbolic regression trees in geometric semantic genetic programming," in *Genetic and Evolutionary Computation Conference*, pp. 1151–1158, 2018.

[16] D. Liu, M. Virgolin, T. Alderliesten, and P. A. N. Bosman, "Evolvability degeneration in multi-objective genetic programming for symbolic regression," in *Genetic and Evolutionary Computation Conference*, pp. 973–981, 2022.

- Title :

ULTIMATE: mUlti-Level Trustworthiness to IMprove the Adoption of hybrid arTificial intelligence

- Author:

Michel Barreteau, Thales

- Abstract:

An increasing number of organisations and corporations have incorporated AI technologies for production and other purposes. Unfortunately, due to a lack of validity, ethics and explainability, the wide-scale adoption of AI has been fraught with difficulties despite multiple potential benefits. The EU-funded ULTIMATE project aimed to develop and evaluate industrial-grade hybrid AI models that address these challenges and enable AI to spread even further through the industrial sector. To achieve this, the initiative provided the stakeholders with methods and tools to ensure trustworthiness (for acceptance purposes) all along the hybrid AI model's life cycle in order to improve worker and AI cooperation.

Full text:

AI has entered the business mainstream, opening opportunities to boost productivity and innovation but suffer limitations hindering wider adoption of knowledge-based or data-driven AI algorithms in industrial settings. Both approaches complement each other and form a critical foundation for the adoption of AI in industry. However, hybrid AI does not fully address the issue of trustworthiness (validity, explainability, and ethics).

The ULTIMATE project pioneered the development of industrial-grade hybrid AI based on three stages to ensure trustworthiness:

- relying on interdisciplinary data sources and adhering to physical constraints regarding the design & development of hybrid AI (1st stage),
- as well as the development of tools for explaining, evaluating and validating hybrid AI algorithms and asserting their adherence to ethical and legal regulations (2nd stage);
- these will be exemplified using real-world industrial use cases (3rd stage) in the Robotics (collaboration between human and robots for logistics activities) and Space domains (Failure detection for satellites) to promote the widespread adoption of hybrid AI in industry;
- in addition, ULTIMATE investigated an end-to-end trustworthy AI methodology that considers ethical values together with the usual technical criteria mentioned above.

The breakthrough generic hybrid AI architectures with improved explainability and robustness and the predictive model on trustworthiness developed in ULTIMATE provide industrials with improved shopfloor efficiency (reduction objective of downtime and of operational costs) and empower their staff through trustful human/machine cooperation allowing highly skilled jobs and increasing decision power and safety. This aims at being beneficial to European industry to gain pre-emptive advantage in the market of industrial AI solutions and will increase trustworthiness in the use of hybrid AI components by the wider public.

Trustworthy AI Summit

September 16

Europe's leading event for trustworthy AI

FAIR AI SCRUM

Eliza Hobo, Quirine Smit, Nina van Liebergen, Cor Veenman
eliza.hobo@tno.nl

TNO: The Netherlands Organization for Applied Scientific Research
Data Science Department

Working with Intersectional Fairness: A Handbook for Scrum Teams

INTERSECTIONAL FAIRNESS

Fairness through the intersectional framework understands and addresses the harms experienced by individuals due to the intersecting and often marginalised aspects of their identity.

FAIRNESS INTEGRATED IN SCRUM

Fairness should not be treated as an afterthought, but considered throughout development

This fits well within Scrum because:

- Scrum is iterative & adaptive, and
- there is space for reflection & discussion

FAIR AI CANNOT BE ACHIEVED WITH A CHECKLIST

The Fair AI Scrum Handbook

- offers tools and starting points to support integrating fairness.[1]
- does not guarantee a fair product but guides intentional progress toward it.

Because:

- fairness is not a fixed outcome.
- it requires ongoing dialogue, reflection, and adaptation.
- this process can be difficult and uncomfortable.

1 Approach based on: Steven Vethman, Quirine T. S. Smit, Nina M. van Liebergen, and Cor J. Veenman. 2025. Fairness Beyond the Algorithmic Frame: Actionable Recommendations for an Intersectional Approach. In Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT '25). Association for Computing Machinery, New York, NY, USA, 3276–3290. https://doi.org/10.1145/3715275.3732210

Fairness in Scrum Artefacts, Roles and Events

Product Backlog/ Goal

- **Document** all the risks you come across: in the approach, methods, data, application.
- Define the Product Goal such that it contributes towards **social justice**
- Evaluate whether a **technical or AI solution is necessary**. Would a non-tech solution be better for the problem at hand?
- …

Scrum Master

- Make the participation of community representatives mutually beneficial and **financially sustainable**.
- Facilitate sessions that promote finding a **common language** between all disciplines in the team
- Ensure there is a mechanism for impacted communities to **voice their worries**, passively and proactively
- ...

Sprint Planning

- Invite **community stakeholders** to participate. Recognize that community goals may diverge from those of Developers.
- Consider holding **separate sessions** for community members for an open and honest dialogue
- Ensure there is enough **time to critically test** and evaluate the intended data, models and metrics.
- ...

Want to know more? Download the handbook

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the Culture Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

DIVERSIFAIR

Co-funded by the European Union

Needle in a Patched Haystack: Evaluating Saliency Maps for Vision LLMs

Bastien Zimmermann [1] **Matthieu Boussard** [1]

1. Introduction & Motivation

Retrieval-Augmented Generation (RAG) fuses retrieval with large-language modeling to ground answers in external evidence. In its *multimodal* variant—jointly processing text and page-level imagery—RAG powers document search, medical-image triage, and other high-stakes tasks. Systems such as **ColPali** encode each page into visual tokens, apply a late-interaction matcher to rank pages, and overlay a cosine-similarity heat map as an explanation.

Recent work shows that raw cosine similarity is an unreliable saliency signal. Mapping late-interaction scores back to input patches is ill-posed, and representational overlap can obscure whether the model attends to truly diagnostic features—at odds with the transparency, robustness, and accountability demanded by the EU AI Act. Our analysis confirms that ColPali-style maps exhibit spatial artifacts, modality drift, and lexical cross-talk, undermining trust in the AI systems.

To remidiate this we provide the following contribution:

- **Benchmark**: We release *Needle-in-a-Patched-Haystack*, comprising four synthetic, model-aligned datasets with per-patch metrics for rigorous localization testing.
- **Limit analysis**: We prove that cosine similarity fails as a faithful saliency proxy in late-interaction VLMs.
- **Patch-level dissection**: We propose a lightweight routine that traces evidence accumulation across image and text streams.

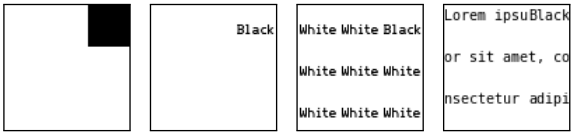
Together, these tools elevate interpretability from a "nice-to-have" visual overlay to a falsifiable property of multimodal RAG systems. In this short paper we introduce the datasets and related benchmark, further results are tackled in (Zimmermann & Boussard, 2025).

2. Benchmark Overview

2.1. Patch-Based Datasets for Vision-Language Models

To probe patch-level localization and text retrieval under controlled conditions, we generate *model-aligned* images whose grid exactly matches each backbone's native patch layout. Every sample contains a single *special patch* that ground-truth saliency should elevate; all other patches act as distractors. We release four increasingly realistic variants:

- **Patch** — a blank grid with one black square; tests pure spatial localisation.
- **Single-Word** — the black square now carries a high-contrast word; evaluates joint visual–text focus.
- **Multi-Words** — every patch contains text; only the target patch bears the *needle* word while distractors show a confounding word. We create *positive* (semantic cosine > 0.7) and *negative* (cosine < 0.1) word-pair subsets by sampling from a 2 000-adjective fastText vocabulary augmented with 10 synthetic non-words.
- **Text** — all non-target patches are filled with unrelated Lorem-Ipsum sentences, stressing the model's ability to ignore dense textual clutter.



(a) Patch (b) Single-word (c) Multi-words (d) Text

Figure 1. Visual representations of datasets used to assess VLMs, with the *special patch* at position (2, 0) inside a 3 × 3 grid.

2.2. Evaluation Metrics for Image Similarity Maps

Let the flattened similarity map be $\mathbf{s} \in \mathbb{R}^n$, with peak index $i_{\max} = \arg\max_i s_i$, and let $\mathcal{I}$ denote the set of ground-truth patch indices (one or more if the *needle* spans several cells). We report four complementary, patch-level scores:

- **Accuracy:** A binary success indicator,
$$\mathrm{Acc} \;=\; \mathbb{1}\big(i_{\max} \in \mathcal{I}\big),$$
- **Score:** The mean similarity of the interesting regions,
$$\mathrm{Score} \;=\; \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} s_i,$$
capturing the absolute response strength the model allocates to the target content.

- **Rank:** The 1-based ordinal position of the interesting patch in the global ordering of similarities:

$$\widehat{\text{Rank}} = \frac{1}{HW} \sum_{j=1}^{HW} \mathbb{1}\big(s_j > \max_{i \in \mathcal{I}} s_i\big),$$

- **Distance (*normalised*):** The Euclidean distance between the max patch and the *nearest* interesting patch:

$$\widehat{\text{Dist}} = \frac{1}{\sqrt{(H-1)^2 + (W-1)^2}} \min_{i \in \mathcal{I}} \big\| \mathbf{p}_{\max} - \mathbf{p}_i \big\|_2,$$

All metrics operate at the patch level, aligning evaluation with the model's internal granularity and avoiding pixel-scale artefacts.

2.3. Needle-in-a-Patched-Haystack Evaluation

For each grid coordinate (x, y) we insert the *special patch*, run the model, and log the four metrics from Section 2.2. The resulting 2-D *grid result map* visualises localisation performance per position. Averaging these maps across random seeds yields the *Needle-in-a-Patched-Haystack* surface, which simultaneously highlights zones of reliable grounding and recurrent failure, offering a succinct diagnostic of patch-level saliency.

2.4. Results at a Glance

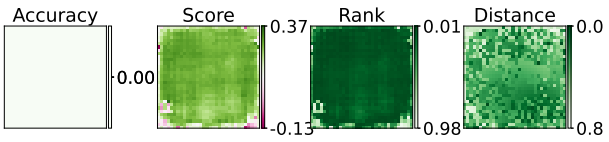


Figure 2. Needle-in-a-Patched-Haystack for the Patch Dataset using the *ColPali* model.

- **Progressive realism helps.** As we move from PATCH to TEXT inputs, all models show rising *Accuracy* and falling *Distance* (Fig. 3), confirming that similarity maps improve when the stimulus resembles the training distribution.

- **Model strengths diverge.** *Gemma* excels on SINGLE-WORD, while the RAG-tuned *ColPali* and *ColQwen* dominate on dense TEXT, indicating that late-interaction fine-tuning favours cluttered layouts.

- **Spatial artefacts persist.** *ColPali* exhibits an "O-shaped" artefact in the bottom left corner (Fig. 2); *ColQwen* shows a bottom-left bias, revealing position-dependent weaknesses invisible to global metrics.

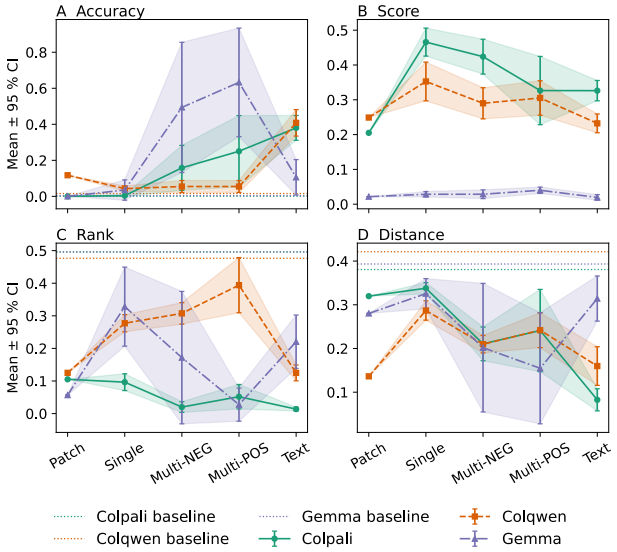


Figure 3. Evolution of four localisation metrics (mean ± 95% CI) as input realism increases. Higher is better for *Accuracy* and *Score*; lower is better for *Rank* and *Distance*. Random Baseline is model patch-grid dependent so there is one baseline per model.

- **Lexical interference is model-dependent.** Semantically related distractors hurt *ColPali* and *ColQwen* but *improve Gemma* ; late-interaction objectives appear to amplify token-level confusion.

- **Well above chance.** All three VLMs outperform a random baseline on every metric, so observed quirks stem from architectural bias rather than noise.

3. Conclusion

Cosine-similarity heat maps rarely reveal how multimodal RAG models truly ground predictions. By pairing a formal analysis of the late-interaction mechanism with the *Needle-in-a-Patched-Haystack* benchmark, we showed that state-of-the-art VLMs suffer from position-dependent artefacts, a persistent image–text modality gap, and limited spatial reasoning. Our open-source datasets, metrics, and diagnostic surfaces provide a lightweight drop-in test that practitioners can run before deployment. Closing the identified gaps—through better positional encodings, bias-aware fine-tuning, and real-document evaluation—remains critical for building transparent and trustworthy multimodal systems.

References

Zimmermann, B. and Boussard, M. Aiguille dans une botte de foin quadrillée : évaluation des cartes de saillance pour les llm de vision. CNIA, 2025.

Towards Trustworthy and Efficient Smart Routing for Large Language Models

ROULE Jule[1], ILHE Paul[2], MOUAYAD Mehdi[1], MAZARS Gilles[2], BARRY Mariam[1]
BNP Paribas[1], Vector 8[2]

Abstract

With the rapid development of numerous Large Language Models (LLMs), the task of selecting the *most suitable* model for a given query has become an important challenge. Smart routing aims to address this by dynamically selecting the model for a query that maximizes a specific performance, cost, or safety criterion. In this paper, we propose a smart routing system tailored to a specific domain that can also generalize to newly released LLMs. Our architecture combines lightweight classifiers with safety-aware filters to route requests based on skill, complexity, and risk. Preliminary results show that learned routers can significantly reduce costs while maintaining output quality, and that integrating guardrails improves robustness against adversarial and sensitive inputs. This work paves the way for LLM routing systems that are not only efficient but also trustworthy and ethically guided.

Goals and Motivation

- **Trustworthiness** Our goal is to build a dynamic router tightly linked to specific domains to be more reliable and transparent.
- **Safety**: We aim to enhance the router's robustness by integrating Guardrails that act as a security layer before any LLM interaction. These filters detect harmful intent, prevent malicious inputs, and reduce exposure to biased or unsafe content. This ensures safer handling of user queries and better alignment with regulatory and ethical standards.
- **Cost efficiency**: High-end models like GPT-4 offer excellent performance but are often overused even for tasks that could be handled by smaller and cheaper models, leading to unnecessary compute costs and carbon footprint.

State of the art

Recent papers propose both a review of the existing dynamic routers and the metrics derived to assess their performance.

	Router	all-strong	all-weak	strong-to-weak
	Oracle r_o	1.39	0.77	0.96
	$r_o(0.5)$	1.55	1.42	1.45
$m=3$	LinearR	1.54	1.54	0.81
	MLPR	1.50	1.52	0.76
	C-RoBERTa	0.93	0.94	0.00
	MLC	1.52	0.34	0.52
	PRknn	1.58	1.56	1.52
	Random	1.59	1.59	1.59
	Oracle r_o	2.09	0.90	1.49
	$r_o(0.5)$	2.27	2.00	2.15
$m=5$	LinearR	2.27	2.28	1.58
	MLPR	2.26	2.25	1.50
	C-RoBERTa	1.53	1.53	0.00
	MLC	2.25	0.03	1.06
	PRknn	2.31	2.30	2.28
	Random	2.32	2.32	2.32

Figure 1:The Entropy (E_p) for different routers from [1]. Some router methods suffer from the classification bias, i.e. when E_p is close to 0.

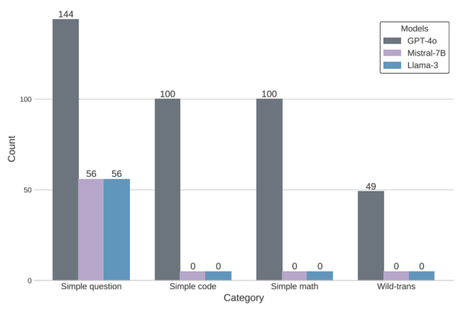


Figure 2:Routing results on Code, Math, and Translation on simple benchmarks [3].

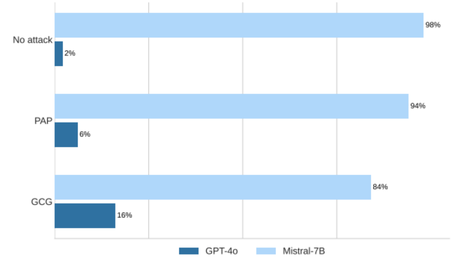


Figure 3:Routing results on the safety benchmark AdvBench, compared against plain harmful text [3].

Proposed Dynamic Routing

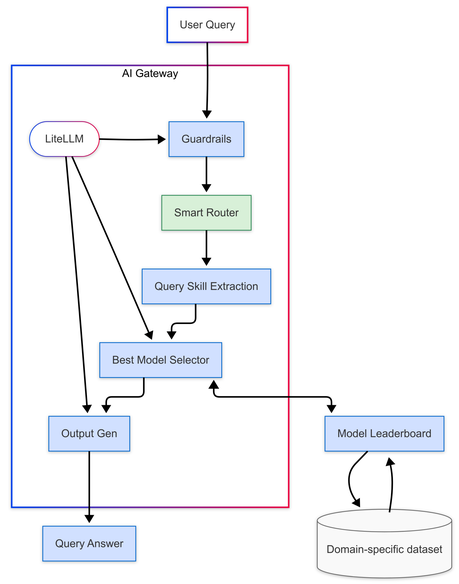


Figure 4:Architecture diagram

Our architecture leverages the functionalities of **LiteLLM** to orchestrate multiple large language models (LLMs) while enforcing safety through integrated guardrails, such as *PII masking* and *prompt injection* detection. In the middle of this system there is an **AI Gateway**, responsible for processing incoming queries: it first assesses whether a request should be accepted or rejected based on predefined safety criteria, then routes it to the most suitable LLM according to a given criteria such as performance/cost ratio or performance/energy ratio. This decision is guided by analyzing the query's complexity and required capabilities, matched against a dynamic model leaderboard.
Building on this foundation, the architecture is an **adaptive and domain-specialized smart router** capable of evolving alongside technological advancements. Unlike static systems, our router can integrate new LLMs without requiring full retraining, ensuring long-term scalability.

For each application domain (e.g., medical, legal, financial), a small fine-tuned language model (SLM) classifies queries and supports routing toward the most efficient model.
This design enables both cost-effective performance and domain-level precision, making the system resilient and optimized for real-world deployments.
With this architecture, several AI Applications can post requests to the Router Service when it needs the result of an ML model, decoupling the application logic from the AI model itself. One immediate benefit is that we can easily switch between models and serve multiple applications from the cumbersome and recurring models' adaptations over time.

Future Work

In the next phase, we aim to strengthen our router with a Responsible AI layer by integrating LLM Guardrails [2]. These safety filters will operate just before the routing, detecting harmful intent, sensitive content, or adversarial inputs before any generation occurs. Additionally, we plan to evaluate our system using a domain-specific benchmark dataset, allowing us to test both its routing accuracy and its ability to enforce safety and ethical constraints in specialized contexts. This will help ensure not only efficient model selection, but also trustworthy, context-aware orchestration of LLMs in real-world applications.

References

[1] Huang, Zhongzhan et al., Routereval: A comprehensive benchmark for routing llms to explore model-level scaling up in llms, (2025)

[2] Dong et al., Safeguarding large language models: A survey, (2024)

[3] Kassem et al., How Robust Are Router-LLMs?(2025).

Trustworthy AI in Air Cargo Compliance: A Small Language Model Approach with Contextual Retrieval and Chain-of-Thought Reasoning

Christopher Enriquez Urban
Fraunhofer IML

Abstract

We present a trustworthy AI system for air cargo regulatory compliance that combines small language models with contextual retrieval and chain-of-thought reasoning to ensure transparent, interpretable decisions. Our approach addresses critical trust requirements in regulatory compliance by providing explicit reasoning chains and auditable decision processes while maintaining accessibility through resource-efficient deployment. The system demonstrates that trustworthy AI can be achieved through interpretable architectures rather than relying solely on larger, opaque models. Current consortium validation focuses on establishing industry confidence in automated compliance decision-making.

1. The Critical Need for Trustworthy AI in Regulatory Compliance

High-Stakes Decision Making: Regulatory compliance errors result in substantial financial penalties and operational disruptions [1]. The consequences of incorrect decisions in this domain extend beyond immediate costs to long-term reputational damage and regulatory scrutiny.

Transparency Requirements: Regulatory authorities and industry stakeholders demand explainable AI decisions for audit and validation purposes [3]. Traditional AI systems must provide clear justification for their recommendations to meet regulatory standards.

Trust Gap: Current AI solutions often operate as ”black boxes,” limiting adoption in compliance-critical environments. The opacity of existing systems creates hesitation among industry professionals who require understanding of decision-making processes.

Industry Impact: Manual compliance verification creates bottlenecks, delays, and human error risks in global supply chains [2]. The scale and complexity of modern air cargo operations necessitate automated solutions that maintain human-level trustworthiness.

2. Our Trustworthy AI Approach: Interpretability by Design

Core Philosophy: Trust through transparency rather than complexity. Our fundamental principle prioritizes explainable decision-making over raw performance metrics, ensuring that stakeholders can understand and validate system outputs.

Architectural Principles:

- Small, specialized models over large general-purpose systems
- Explicit reasoning processes that mirror human decision-making
- Contextual information retrieval that shows ”what the system knows”
- Structured decision outputs that enable validation and review

Trustworthiness Pillars: Our approach is built on four foundational elements: *Interpretability* ensures decisions can be understood; *Auditability* provides clear trails for regulatory review; *Accessibility* enables deployment in resource-constrained environments; and *Reliability* maintains consistent performance across diverse scenarios.

3. System Architecture for Trustworthy Decision-Making

Multi-Stage Reasoning Pipeline:

Contextual Knowledge Retrieval: Transparent selection of relevant regulatory information ensures that the system's knowledge base is visible and verifiable. This stage identifies and prioritizes applicable regulations, standards, and compliance requirements specific to each cargo scenario.

Chain-of-Thought Processing: Step-by-step reasoning visible in "think" tags provides complete transparency into the decision-making process. Each logical step is documented, allowing stakeholders to follow the system's reasoning from initial assessment through final determination.

Structured Decision Output: Clear reasoning chains, explanations, and classification decisions are presented in standardized formats that facilitate review and validation by human experts and regulatory authorities.

Trust Enhancement Features:

Reasoning Transparency: Every decision includes visible logical steps that can be independently verified. This transparency enables stakeholders to identify potential errors or biases in the decision-making process.

Context Provenance: Clear tracking of which regulations inform each decision provides accountability and enables targeted updates when regulatory requirements change.

Uncertainty Handling: Explicit identification of insufficient information cases prevents overconfident decisions and directs attention to areas requiring human intervention.

Local Deployment Capability: Data privacy through on-premises processing addresses security concerns while maintaining full functionality in restricted environments.

Industry Validation Framework: Consortium-based approach for establishing real-world trust involves multiple stakeholders in the validation process, ensuring broad acceptance and reliability across diverse operational contexts.

4. Broader Impact and Future of Trustworthy Regulatory AI

Paradigm Shift: Moving from "trust through performance" to "trust through transparency" represents a fundamental change in how AI systems are evaluated and deployed in critical applications. This shift prioritizes understanding over accuracy metrics alone.

Scalability: Framework applicable across multiple regulatory domains requiring explainable decisions. The architectural principles developed for air cargo compliance can be adapted to other regulatory environments including customs, safety, and environmental compliance.

Industry Transformation: Enabling automated compliance while maintaining human oversight and validation creates new possibilities for efficient, reliable regulatory processes. This balanced approach preserves human authority while leveraging AI capabilities.

Research Contribution: Demonstrating that trustworthy AI can be achieved through thoughtful architecture design rather than computational scale challenges the prevailing emphasis on larger models and highlights the importance of transparency in critical applications.

References

[1] Gai, P., Kemp, M., Sanchez Serrano, A., and Schnabel, I. (2019). Regulatory complexity and the quest for robust regulation. Reports of the Advisory Scientific Committee, No 8, June 2019. European Systemic Risk Board.

[2] World Bank. (2021). Air Cargo Digitalization: From EDI to Community Systems. World Bank Document.

[3] IATA. (2021). Guidance on Compliance with Electronic Advance Cargo Information Requirements. International Air Transport Association.

Title: Empowering Critical Thinking in LLM Use: Designing Support for Risk Mitigation

Abstract:

Large Language Models (LLMs) such as GPT-4 and LLaMA 2 have become widely adopted due to their powerful natural language capabilities and accessible, chat-based interfaces. These systems are increasingly used for tasks like information retrieval, decision support, and creative writing. However, despite their strengths, LLMs are prone to producing outputs that may be inaccurate, inconsistent, incomplete, irrelevant, or biased—risks that are often difficult for users to detect. The fluent and convincing style of LLM-generated text can lead users to overestimate the reliability of the content, resulting in uncritical acceptance and potentially serious real-world consequences, including legal or medical errors.

Current research on mitigating these risks has largely focused on technical solutions such as fine-tuning or fact-checking algorithms. While these strategies may be useful, they are also often resource-intensive, insufficient to fully address the lasting drawbacks inherent to LLMs' functioning, and happen 'behind the scenes'. Therefore, important information that may be pivotal for judging the trustworthiness of the outputs is withheld from the user in such approaches. As such, we argue that a complementary, user-centered strategy is necessary—one that empowers users to critically evaluate and contextualize the outputs they receive from LLMs.

In this work, we explore critical thinking as a key tool for safe and effective LLM use. We define it as a set of reflective and analytical practices that enable users to evaluate claims, identify potential biases, consider alternative perspectives, and question underlying assumptions.

To support this, we developed five **Critical Thinking Support Functions (CTSFs)**—conceptual tools and design features intended to foster user awareness, reflection, and scrutiny during LLM interactions. These functions are a hybrid of technical mitigation and user-facing feedback mechanisms, designed to reduce the cognitive effort needed for critical evaluation while raising awareness of potential pitfalls. Simultaneously, users maintain cognitive agency when interacting with LLMs. The CTSFs include:

1. **Prompt Pro** – Guides users in formulating clearer and more specific prompts, reducing ambiguity and increasing output quality. Aims to mitigate: Inaccuracy, Incompleteness, Irrelevancy, and Inconsistency
2. **Inaccuracy Identifier** – Flags outputs likely to contain factual errors, encouraging users to verify information. Aims to mitigate: Inaccuracy

3. **Consistency Conveyor** – Highlights discrepancies across multiple LLM outputs, fostering awareness of conflicting interpretations. Aims to mitigate: Inconsistency
4. **Tool Teacher** – Informs users when an external tool might be better suited than an LLM for a specific task. Aims to mitigate: Inaccuracy
5. **Bias Buzzer** – Alerts users to potentially biased or harmful content, prompting critical reflection. Aims to mitigate: Bias

These functions were co-developed with experts in the field and evaluated through two participatory workshops involving researchers in NLP, HCI, and data science, as well as professionals from organizations interested in adopting LLMs from domains such as healthcare, finance, ICT, and public services. Interface prototypes of the CTSFs were used in the sessions to illustrate what they might look like in practice. Participants in the first workshop rated the CTSFs on novelty, feasibility, and relevance. The second workshop focused on exploring real-world applicability and added value through activities like storyboarding, scenario analysis, and ranking exercises. Participants applied the CTSFs to their own organizational contexts, using templates to articulate their use cases, expected LLM behavior, and perceived risks.

Across both workshops, **Prompt Pro** was seen as the most broadly applicable and impactful, addressing common prompt formulation issues while indirectly mitigating multiple LLM risks. Participants also saw promise in the **Inaccuracy Identifier** and the **Consistency Conveyor,** but they mentioned that these functions could in part be covered by the **Prompt Pro** as well. The **Tool Teacher** and the **Bias Buzzer** were seen as less generalizable and more context-/culture-dependent.

One key additional insight from this work is the value of intentional “friction” or seamful design—features that momentarily disrupt the interaction to encourage user reflection. This stands in contrast to the trend toward seamless AI interfaces, which may unintentionally suppress user criticality. The Prompt Pro concept, for example, introduces contextual prompt feedback within the chat interface, nudging users to think more deliberately about their inputs and expectations. Striking the right balance between usability and intentional friction will be crucial in future implementations.

By embedding support for critical thinking into LLM experiences, this work helps users better navigate uncertainty, reduce the impact of misleading outputs, and promote more responsible use of generative AI tools. Rather than aiming for (likely impossible) fully automated correctness, we advocate for interaction designs and user support systems that cultivate critical engagement. Future work should focus on testing these support functions in applied settings and developing robust measures to evaluate critical thinking in human-LLM interaction.

Beyond Feature Attribution Explainers: Exploiting Structural Semantics between Features and Outcomes to Explain ML Models

Athina Georgara*, Adarsh Valoor*, Sarvapali D. Ramchurn

University of Southampton, Southampton, UK
{a.georgara, adarsh.valoor, sdr1}@soton.ac.uk

Abstract

This paper discusses an alternative approach to explain ML models, exploiting semantic relations and latent patterns among features and outcomes. In addition, a counterfactual explanation graph is used to support the proposed explanatory technique.

Introduction

One fundamental objective towards Trustworthy AI is the capability of providing explanations regarding the AI-driven decisions. Explanations help users to understand the decisions of AI systems and allow the systems to earn users' trust. This is particularly important in sensitive areas such as healthcare, finance, and legal systems, where the implications of AI decisions can be significant. In this work, we focus on a particular class of AI applications, the machine learning (ML) classifiers.

The most common approach to explain an ML classifier is the so-called *feature attribution* methods. A feature attribution method assigns an importance score (or weight) to each feature, indicating the degree to which each feature influences the final decision. Literature offers many feature attribution methods that are well received by the community. Ribeiro et al. introduced *LIME* (Ribeiro, Singh, and Guestrin 2016), a model-agnostic method for providing local explanations. LIME builds a linear model around the input and uses the weights of this local linear model as indicators to identify the most important input features. Chen and Song (Chen and Song 2018) propose a framework where a feature selection function is trained to identify the most informative subset of features for each individual data instance. Last but not least, Lundberg and Lee put forward *SHAP* (Lundberg and Lee 2017) that utilises a game-theoretic solution concept to compute the marginal contributions of the different input features to the outcome. Notably, the SHAP explainer comes with a plethora of variations (KernelSHAP, TreeSHAP, DeepSHAP, etc.) devised to address different types of models and computational efficiency.

Feature attribution methods provide explanations that illustrate the most influential features identified. Helpful as these techniques might be, in many cases, they might result in 'complex' explanations. For example, listing all the features (with the corresponding importance score) can be confusing and overwhelming. Even if one limits the number of features to the most influential ones, the explanation can still be complex in the sense that these features together may not give a clear justification towards the decision.

Motivation

Feature attribution methods like LIME and SHAP assign importance scores to individual features, often yielding complex explanations that overwhelm users. (Miller 2019) notes users prefer simpler, cognitively salient explanations. ProToMEx addresses this by using probabilistic topic modelling to identify latent feature combinations tied to class labels, offering clear global and local explanations. Extending the explainer with a counterfactual explanation graph identifies alternative decision paths, enhancing interpretability. Analysing path overlaps in instance-specific graphs assesses classifier robustness, promoting trustworthy AI in critical domains like healthcare and finance by providing concise, meaningful insights into decision-making processes.

Related Work

The field of explainable AI has seen significant advancements in developing methods to interpret ML classifiers. Feature attribution methods are among the most prominent. (Smith, Doe, and Brown 2025) enhances LIME by integrating dynamic feature weighting for improved local explanations. (Johnson and Lee 2025) extends SHAP with optimised Shapley value computations, reducing complexity for large datasets. (Zhang, Wang, and Chen 2025) proposes a rule-based explainer using causal inference to identify feature interactions. Unlike these, ProToMEx employs probabilistic topic modelling (LDA) to extract latent feature-class patterns, offering global and local explanations. Its counterfactual graph extension, analysing decision paths, complements these methods, enhancing interpretability and classifier robustness evaluation. ProToMEx innovates by applying LDA to ML classifiers, transforming tabular data into documents to extract topics that represent feature combinations associated with class labels. Followed by a counterfactual explanation graph, which builds on this by con-

*These authors contributed equally.

This work is supported by the Responsible AI UK (EP/Y009800/1) (https://rai.ac.uk/)

structing instance-specific graphs to identify alternative decision paths, complementing existing methods like LIME and SHAP. This approach not only enhances interpretability but also allows for the evaluation of classifier robustness by analysing path overlaps, distinguishing ProToMEx from traditional XAI methods.

Proposed Methodology

Our methodology integrates a probabilistic topic model explainer with a counterfactual explanation graph to enhance Machine Learning classifier interpretability and robustness evaluation. Combining topic modelling for latent feature patterns with graph-based analysis for decision paths, it offers comprehensive explanations and classifier reliability insights. It uses Latent Dirichlet Allocation to uncover latent semantic structures, or topics, representing feature combinations tied to class labels, unlike LIME and SHAP, which focus on individual feature importance. Tabular data is transformed into a textual corpus, with each instance and predicted label forming a document. Input vectors are sampled, labelled by the ML classifier, and converted into documents. An LDA model extracts topics as feature-label distributions. It provides global explanations, showing classifier behaviour, and local explanations for specific instances. A counterfactual explanation graph complements it by constructing instance-specific graphs with nodes as features and labels and edges as their interactions. Paths from features to predicted labels identify feature combination patterns, compared with models' topics for consistency. Path overlap analysis assesses classifier robustness—high overlap indicates reliability, and low overlap suggests instability. This provides counterfactual insights into feature changes altering outcomes, enhancing explanation depth. Preliminary results show effectiveness in uncovering alternative paths, enriching the understanding of classifier behaviour in complex datasets

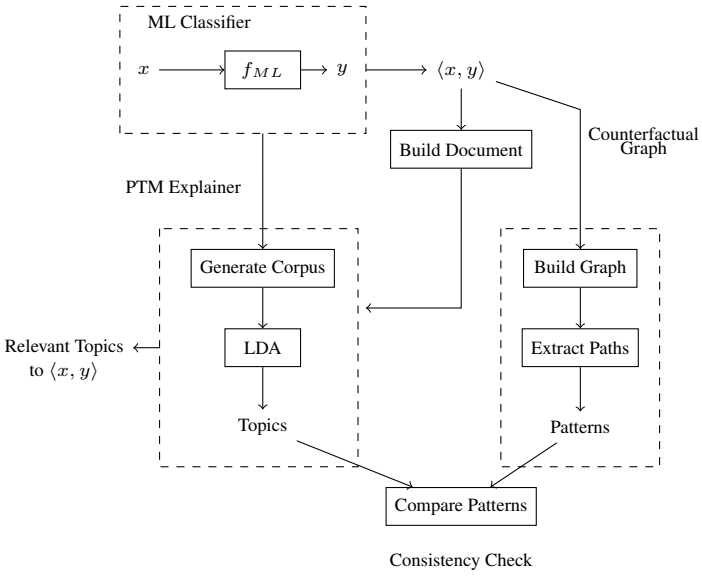


Figure 1: Proposed methodology pipeline

Challenges

Implementing the explainer and its counterfactual graph extension faces challenges: (1) transforming tabular data into a textual corpus must preserve feature semantics to avoid distorted topics; (2) training LDA requires large, representative datasets, which are computationally intensive for high-dimensional data; (3) constructing counterfactual graphs for non-linear classifiers is complex, needing efficient path capture; (4) comparing graph patterns with explainers' topics demands robust similarity metrics; and (5) assessing path overlaps for classifier robustness requires dataset-specific overlap thresholds. These issues necessitate innovative preprocessing, efficient algorithms, and tailored tuning.

Work in Progress and Future Work

Current efforts focus on refining explainers' data transformation process to optimise corpus generation for diverse datasets, improving topic quality. We are also developing scalable algorithms for constructing counterfactual graphs, reducing computational overhead for large feature sets. Preliminary experiments are underway to quantify path overlaps across datasets, aiming to establish robust metrics for classifier evaluation. Future work includes a comprehensive user study to assess the interpretability and acceptance of explanations, particularly with counterfactual graphs, among domain experts in healthcare and finance. We plan to extend it to handle multi-label/class problems more effectively, addressing limitations in current topic modelling applications. Additionally, integrating graph-based metrics with the explainer's topic-based explanations will be explored to create a unified framework for assessing classifier reliability and explanation quality. These advancements aim to enhance its applicability in real-world, high-stakes AI systems.

References

Chen, J.; and Song, L. 2018. Learning to Explain: An Information-Theoretic Perspective on Model Interpretation. *Proceedings of the 35th ICML*, 80: 883–892.

Johnson, R.; and Lee, S. 2025. Optimized Shapley Value Computations for Scalable SHAP. *Proceedings of the 2025 ICML*, 456–467.

Lundberg, S. M.; and Lee, S.-I. 2017. A Unified Approach to Interpreting Model Predictions. *NeurIPS*, 30: 4765–4774.

Miller, T. 2019. Explanation in Artificial Intelligence: Insights from the Social Sciences. *AIJ*, 267: 1–38.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. Why Should I Trust You? Explaining the Predictions of Any Classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on KDD*, 1135–1144.

Smith, J.; Doe, J.; and Brown, A. 2025. Dynamic Feature Weighting for Enhanced LIME Explanations. *Journal of Artificial Intelligence Research*, 82: 123–145.

Zhang, L.; Wang, M.; and Chen, D. 2025. Causal Inference for Rule-Based Explanations in ML Classifiers. *NeurIPS*, 38: 789–802.

CONSUMER LABELS FOR BOOSTING AI TRUSTWORTHINESS

Raphael Fischer
Lamarr Institute for Machine Learning and Artificial Intelligence
TU Dortmund University
raphael.fischer@tu-dortmund.de

ABSTRACT

For establishing transparency, boosting trust, and facilitating sustainable development in the context of artificial intelligence (AI), high-level consumer labels were proposed. In analogy to well-established communication systems, they inform on crucial model properties. By summarizing core concepts and discussing practical implications, this work gives a concise overview of AI labeling techniques.

Keywords transparency · labeling · trustworthy AI · sustainability · explainability · reporting · responsible AI

1 Introduction

While artificial intelligence (AI) has become ubiquitous, the importance of trustworthiness[1] and sustainable development [2] necessitates to establish a comprehensible communication format [3]. In order to bridge knowledge gaps between different stakeholders [4], the concept of high-level *AI labels* was developed [5]. Inspired by well-known systems such as energy labeling [6] or trust sealing [7, 8], these labels aim at informing practitioners (i.e., consumers) about practical model properties at a glance. Several works have applauded the conceptual idea [9], proposed custom variants [10, 11], and evaluated practical implications of AI labels [12]. This paper gives an overview for the current state of AI labeling, for which some examples are shown in Figure 1, and presents guidelines for future refinements.

2 Methods and Results

Early AI documentation approaches such as IBM's *fact sheets* have potential for boosting trust [13], however only few works were found to adequately acknowledge the importance of resource-awareness and non-expert comprehensibility [3]. For this reason, and with an explicit focus on AI users (or consumers), the idea of high-level AI labeling was formulated in analogy to textile *care labels* [5]. Initially conceptualized to convey information on theoretic guarantees and practical implementations of machine learning methods, the focus quickly shifted toward the pressing need for AI sustainability [2]. Drawing inspiration from *energy efficiency labeling*, corresponding labels were thus proposed as a means to inform on the practical trade-off between the resource consumption and prediction quality

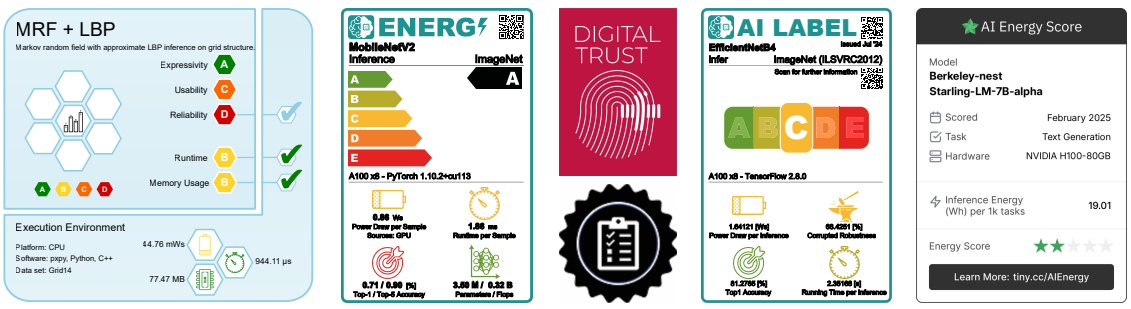


Figure 1: High-level consumer AI labels, proposed in and taken from the respective literature [5, 6, 7, 8, 12, 10].

of AI models [6]. The follow-up generalization introduced the *STREP* framework for sustainable and trustworthy reporting, which constitutes methods for comparing the performance of AI models across various learning tasks and execution environments [3]. With a more explicit focus on trustworthiness, the closely related idea of issuing trust seals was evaluated [7, 8]. However, these studies gave mixed results, which might stem from the hypothetical and simple seal design. This is backed by a recent evaluation study, which found that the labeling authority and balance between simplicity and complexity are prime factors for their trust-boosting effects [12]. From these qualitative findings, design principles for future labeling adaptions were derived, specifically highlighting the need for customizability and the nudging potential for sustainability [14]. This directly relates to recent works proposing to accompany open source AI models with carbon efficiency [11] or energy score [10] labels, in order to advocate sustainable development on platforms like *Hugging Face*.

3 Conclusion

AI labels can potentially foster informed decision-making among various stakeholders and even nudge them toward sustainability. For properly boosting trust, a suitable labeling authority needs to be identified that can adequately balance the trade-off between simplicity and complexity. Future adaptions need to also acknowledge the importance of customizing and certifying the labeling procedure, in order to establish AI labels as a helpful information resource.

References

[1] Raja Chatila, Virginia Dignum, Michael Fisher, et al. Trustworthy AI. *Reflections on Artificial Intelligence for Humanity*, pages 13–39, 2021, DOI 10.1007/978-3-030-69128-8_2.

[2] Aimee van Wynsberghe. Sustainable AI: AI for sustainability and the sustainability of AI. *AI and Ethics*, 1, 2021, DOI 10.1007/s43681-021-00043-6.

[3] Raphael Fischer, Thomas Liebig, and Katharina Morik. Towards more sustainable and trustworthy reporting in machine learning. *Data Mining and Knowledge Discovery*, 2024, DOI 10.1007/s10618-024-01020-3.

[4] David Piorkowski, Soya Park, April Yi Wang, et al. How ai developers overcome communication challenges in a multidisciplinary team: A case study. *Proceedings of the ACM on Human-Computer Interaction*, 2021, DOI https://doi.org/10.1145/3449205.

[5] Katharina Morik, Helena Kotthaus, Raphael Fischer, et al. Yes we care! - Certification for machine learning methods through the care label framework. *Frontiers in Artificial Intelligence*, 2022, DOI 10.3389/frai.2022.975029.

[6] Raphael Fischer, Matthias Jakobs, Sascha Mücke, and Katharina Morik. A unified framework for assessing energy efficiency of machine learning. In *Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, 2022. DOI 10.1007/978-3-031-23618-1_3.

[7] Nicolas Scharowski, Michaela Benk, Swen J. Kühne, et al. Certification Labels for Trustworthy AI: Insights From an Empirical Mixed-Method Study. In *Conference on Fairness, Accountability, and Transparency*, 2023. DOI 10.1145/3593013.3593994.

[8] Magdalena Wischnewski, Nicole Krämer, Christian Janiesch, et al. In seal we trust? Investigating the effect of certifications on perceived trustworthiness of ai systems. *Human-Machine Communication*, 2024, DOI 10.30658/hmc.8.7.

[9] Sergio Genovesi and Julia Maria Mönig. Acknowledging sustainability in the framework of ethical certification for ai. *Sustainability*, 2022, DOI 10.3390/su14074157.

[10] Alexandra Sasha Luccioni, Boris Gamazaychikov, Emma Strubell, et al. *AI Energy Score Documentation*, 2025.

[11] Joel Castaño, Silverio Martínez-Fernández, Xavier Franch, and Justus Bogner. Exploring the carbon footprint of hugging face's ml models: A repository mining study. In *International Symposium on Empirical Software Engineering and Measurement*, 2023. DOI 10.1109/ESEM56168.2023.10304801.

[12] Raphael Fischer, Magdalena Wischnewski, Alexander van der Staay, et al. Bridging the communication gap: Evaluating ai labeling practices for trustworthy ai development. 2025, DOI 10.48550/arXiv.2501.11909.

[13] Matthew Arnold, Rachel KE Bellamy, Michael Hind, Stephanie Houde, Sameep Mehta, et al. Factsheets: Increasing trust in ai services through supplier's declarations of conformity. *IBM Journal of Research and Development*, 2019, DOI 10.1147/jrd.2019.2942288.

[14] Alexander van der Staay, Raphael Fischer, Magdalena Wischnewski, et al. Reflective design theorizing with user interviews: A case study for ai energy labels. In *International Conference on Design Science Research in Information Systems and Technology*, 2025. DOI 10.1007/978-3-031-93979-2_5.

Trustworthy AI Summit Poster Submission

Title : Formal Abductive Explanations for Prototype-Based Networks

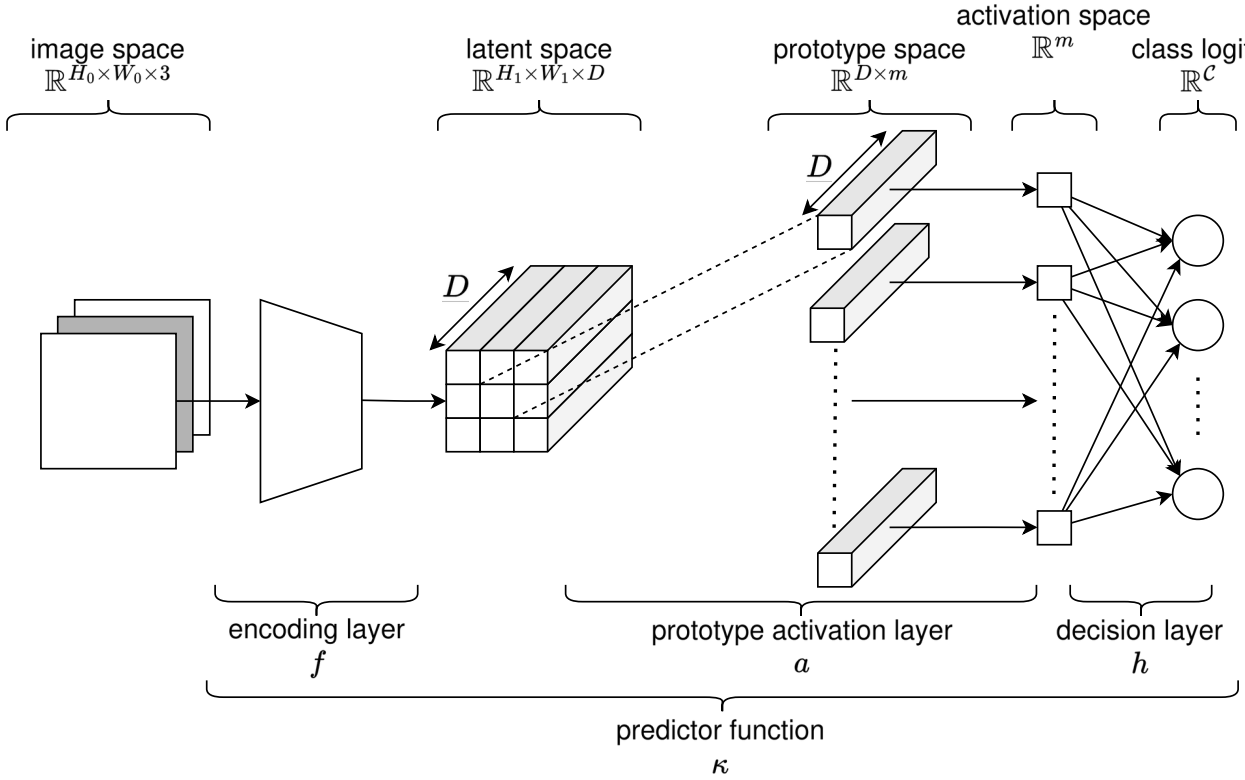
Authors : Jules Soria, Zakaria Chihani, Alban Grastien, Julien Girard-Satabin, Romain Xu-Darme, Daniela Cancila

Affiliation : Université Paris-Saclay, CEA, LIST

Abstract :
We propose a novel approach to Case-based reasoning models, taking advantage of the inherent interpretability in the design of such models, and introducing a new formalism to provide abductive latent explanations. This allows to formally explain images relying on *prototypes* – salient parts of previously seen images, rather than pixel-wise.

Contributions :

In this research project, we aim to bridge the gap between Formal eXplainable Artificial Intelligence (FXAI) and prototype-based learning. We show that it is possible to produce abductive explanations that are correct not only at the pixel-level, but also at the prototype level. We consider that the produced explanation are more suitable for humans than pixel-level ones, while being sufficient to fully justify the model's behaviour.
In particular, we propose a framework to describe Abductive Latent Explanations (ALE) for prototypes. Crucially, our formalism is generic as it can be instantiated given a definition of *feature extractor*, a *prototype* and how *similarity* relates prototypes with current sample.



Concretely, we extend the original definition of formal abductive explanations (AXps); for a given instance **v** and its associated classification *c*, an AXp is a feature subset *X* such that :

$$\forall(\mathbf{x} \in \mathbb{F}). \left[\left(\bigwedge\nolimits_{i \in \mathcal{X}} (x_i = v_i)\right) \rightarrow (\kappa(\mathbf{x}) = c)\right]$$

We adapt this concept to an intermediate latent representation of the instance used during the classification process, which allows us to use prototypes similarities inherently used to reach a decision, and convey a much more human-interpretable explanation. This is akin to relying on the following formula :

$$\forall \mathbf{x} \in \mathbb{F}. \quad \phi_{\mathcal{E}}\big(f(\mathbf{x}), f(\mathbf{v})\big) \Rightarrow \big(\kappa(\mathbf{x}) = c\big).$$

We show that, by using different ϕ experssions, we can produce different formal explanations, that vary in size and shape, and propose three distinct methods to produce them.

We first suggest explaining instances by showing the most activated prototypes by their latent representations, simiilarly to the explanations produced in the original paper introducing prototype-based reasoning networks.

Then, we explore geometric and spatial reasoning to anchor vector that make up the latent representation of the instances, using techniques like the *triangular inequality* of distance metrics, and a *hypersphere intersection approximation* formula.
The intuition behind is that for a vector to be « close » to another (*e.g.* a prototype) also indicates that the vector is close to nearby (with regards to the prototype compared to) vectors, and far away from distant (with regards to the prototype compared to) vectors.

Our work is powered by the CaBRNet library which provides implementations of prototype-based networks.

References :

J. Yu, A. Ignatiev, P. J. Stuckey, N. Narodytska, and J. Marques-Silva. Eliminating the impossible, whatever remains must be true.arXiv preprint arXiv:2206.09551, 2022.

S. Bassan and G. Katz. Towards formal xai: Formally approximate minimal explanations of neural networks. arXiv preprint arXiv:2210.13915, 2022.

J. Marques-Silva and A. Ignatiev. Delivering trustworthy ai through formal xai. In Proceedings of the AAAI Conference on Artificial Intelligence, pages 12342–12350, 2022.

C. Chen, O. Li, D. Tao, A. Barnett, C. Rudin, and J. K. Su. This looks like that: deep learning for interpretable image recognition. Advances in neural information processing systems, 32, 2019.

Trustworthy AI based on Analytical-Model Informed Machine Learning

Frédéric BARBARESCO, THALES ("AI & ALGORITHMS FOR SENSORS" SEGMENT LEADER)
frederic.barbaresco@thalesgroup.com

Abstract

Trustworthy AI for Physics-Informed Neural Networks (PINNs) is an emerging field of research aimed at ensuring that PINNs operate in a reliable, transparent, and robust manner—particularly when employed for critical tasks such as modelling physical systems. PINNs are neural networks trained not only on data but also to respect physical laws, typically expressed as partial differential equations (PDEs). This is achieved by incorporating terms into the loss function that penalise violations of these equations. To render PINNs trustworthy, we propose an extension of the Hamiltonian Neural Network (HNN) by integrating symmetry constraints and Noether invariants within a Symplectic Foliation-Informed Neural Network. This analytically grounded model-informed machine learning framework is further extended to learn dissipative physics with Thermodynamics-Informed Neural Network (TINN), based on metriplectic equations and Souriau's Lie Group Thermodynamics, constrained by transverse Riemannian foliations. Today, we stand at the threshold of the fifth era of science—the artificial scientific intelligence era—in which companies are introducing "AI scientists" based on PINNs and large language models (LLMs) that not only assist in research but actively drive discovery, generate hypotheses, and autonomously test them. These emerging tools present new challenges for the development and assurance of Trustworthy AI.

References:

[1] Barbaresco F (2025), Transverse Symplectic Foliation Structure for Thermodynamics-Informed Neural Network and Lie-Groups Machine Learning, Erwin Schrödinger Institute Seminar, Infinite-dimensional Geometry: Theory and Applications trimester, https://www.esi.ac.at/events/t2213/

[2] Barbaresco, F. (2025) Jean-Marie Souriau's Symplectic Foliation Model of Sadi Carnot's Thermodynamics. Entropy, 27, 509. https://www.mdpi.com/1099-4300/27/5/509

[3] Barbaresco F, Nielsen Frank (2021) Geometric Structures of Statistical Physics, Information Geometry, and Learning: SPIGL'20, Les Houches, France, July 27–31, Springer, https://link.springer.com/book/10.1007/978-3-030-77957-3

[4] European COST network CaLISTA: https://site.unibo.it/calista/en

[5] European HORIZON-MSCA CaLIGOLA: https://site.unibo.it/caligola/en

[6] Nina Miolane, The fifth era of science: Artificial scientific intelligence, https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.3003230

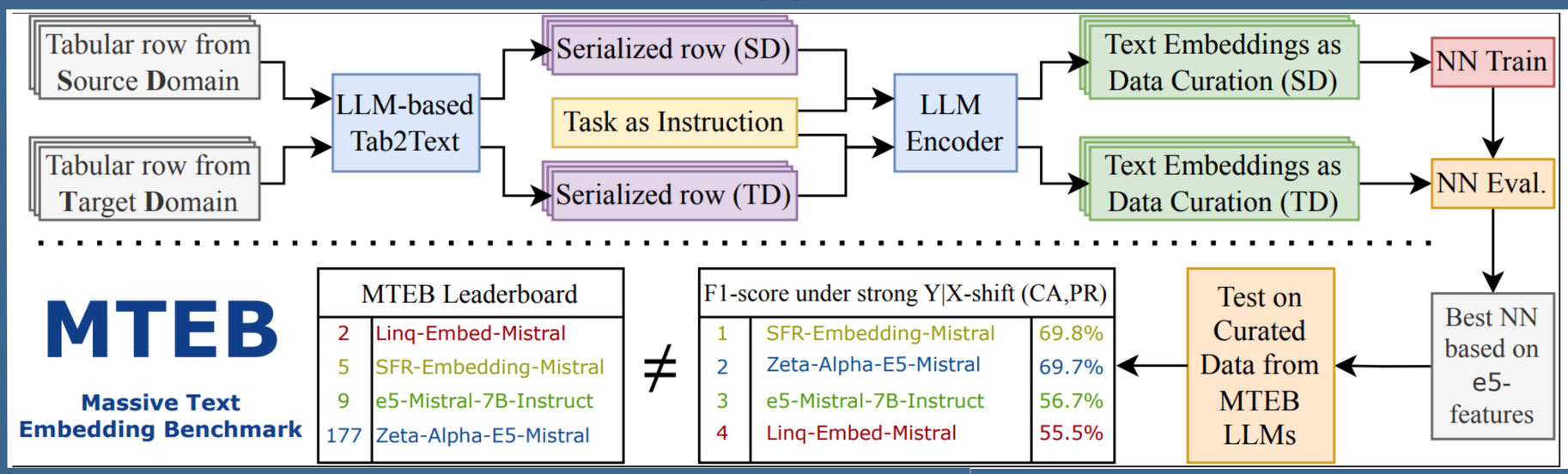
Unraveling OOD Robustness Failure Modes when using LLMs in AI Systems

Lucas Mattioli, Youness Ait-Hadichou, Sabrina Chaouche, Martin Gonzalez

Motivation

- LLM-based text embeddings are increasingly used in ML pipelines/systems.
- There is a need for **Thrustwothy AI Systems** with LLM subcomponents to perform reliably under distribution shifts, especially in critical domains.
- However, uncurated embeddings cause **system collapse** — where predictions degenerate to a single class.

Our Approach



Experimental Setup

Table 1. Name & Ranking of LLMs according to MTEB.

Name	LLM Full Name	Rank
Linq	Linq-Embed-Mistral	2
SFR	SFR-Embedding-Mistral	5
e5	e5-Mistral-7B-Instruct	9
Zeta	Zeta-Alpha-E5-Mistral	177

Table 2. Number of configurations per model family.

Family	MLP	CVaR-DRO	LR	XGB	GBM	RF	SVM
Configs	96	203	100	200	200	101	34

Problem Statement

- Let HP_f: Set of hyperparameters where model f, trained on **raw tabular data** does **not** collapse.
- How many collapse when trained on **LLM embeddings** with **same *HP***?

For test set S, f_c is collapsed if:
$PN^S(f_c) = 0$ $(all\ negative\ prediction)$
or
$PP^S(f_c) = 0$ $(all\ positive\ prediction)$

Collapse ratio:

$$\mathrm{CR}^{S}_{\mathrm{HP}} = \frac{1}{|\mathrm{HP}|}\sum_{c} \mathbb{I}_{\{\mathrm{Cond}^{S}_{T}(c)\}} + \mathbb{I}_{\{\mathrm{Cond}^{S}_{F}(c)\}},$$

Strong collapse ratio:

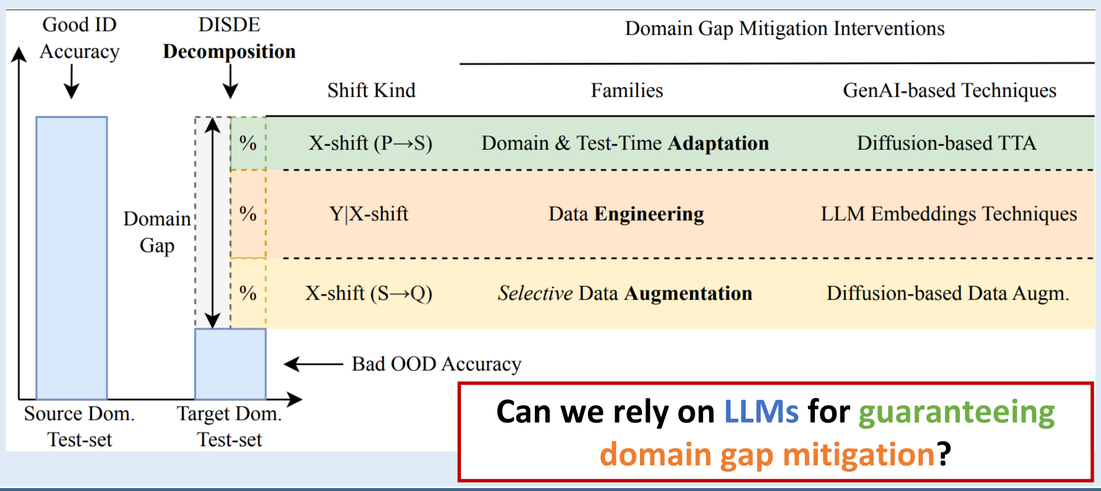
$$\mathrm{CR}^{S,Q}_{s,\mathrm{HP}} = \frac{1}{|\mathrm{HP}|}\sum_{c} \mathbb{I}_{\{\mathrm{Cond}^{S}_{T}(c)\cap\mathrm{Cond}^{Q}_{T}(c)\}}.$$

Projection ratio:

$$\mathrm{CR}^{S,Q}_{p,\mathrm{HP}} = \frac{1}{|\mathrm{HP}|}\sum_{c} \mathbb{I}_{\{\mathrm{Cond}^{S}_{T}(c)\cup\mathrm{Cond}^{Q}_{T}(c)\}}.$$

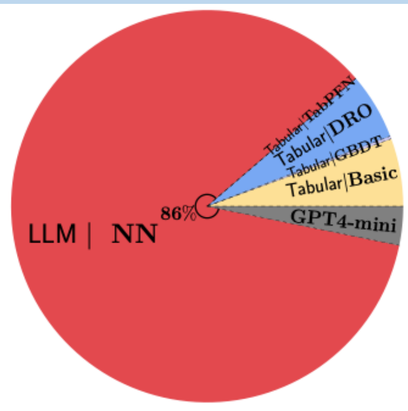
$where\ Cond^S_T(c) = (PN^S(f_c) = 0)$,
and $Cond^S_F(c) = (PP^S(f_c) = 0)$

Domain Gap Dissection Metrics

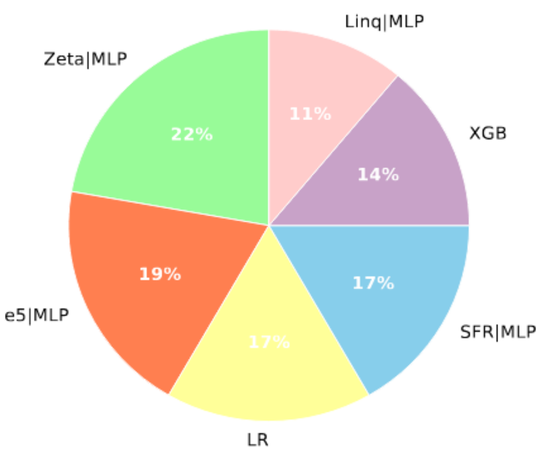


Can we rely on LLMs for guaranteeing domain gap mitigation?

MTEB ranking disagreement

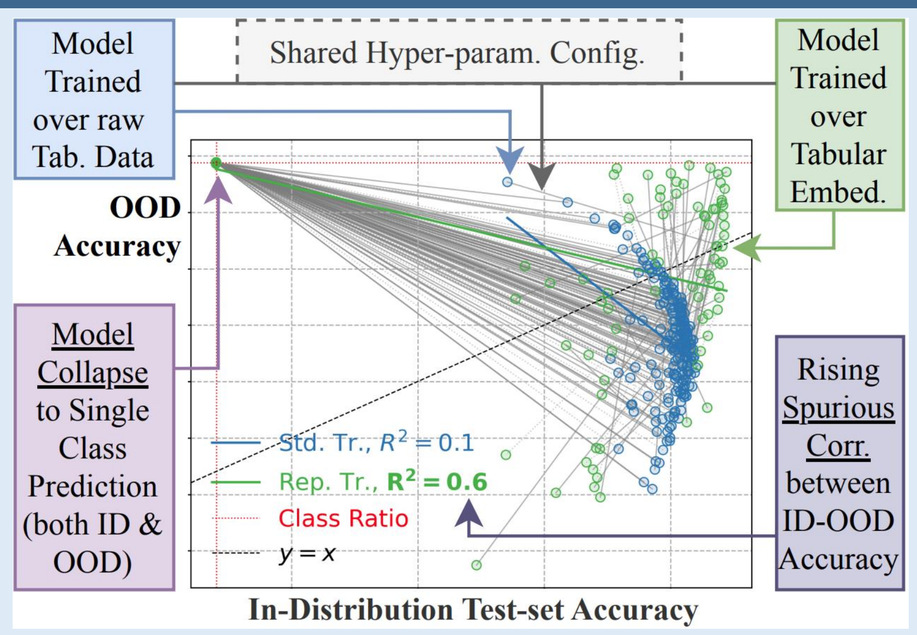


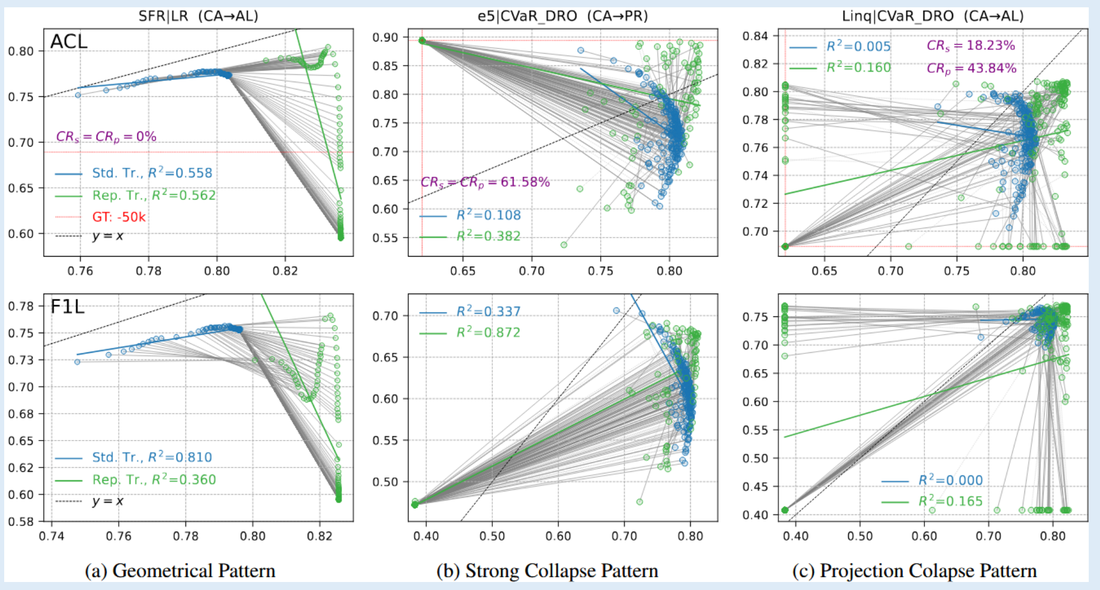
LLM based pipelines outperform single model methods



MTEB LLMs scores and robustness scores are uncorrelated

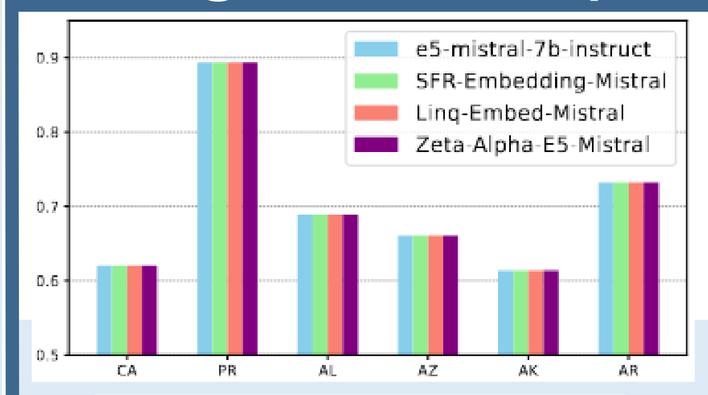
Strong Model Collapse induces spurious ACL



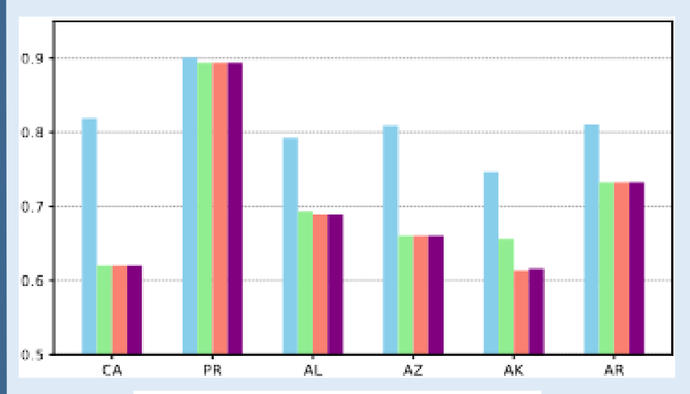


(a) Geometrical Pattern (b) Strong Collapse Pattern (c) Projection Colapse Pattern

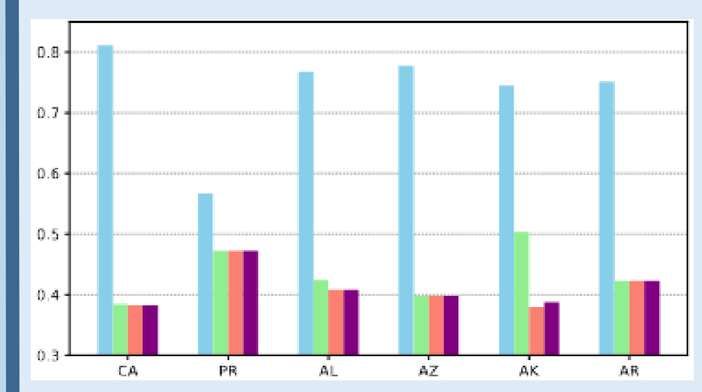
Strong Model Collapse



(a) Single Model Strong Collapse



(b) Optimal e5 HP (Acc.)

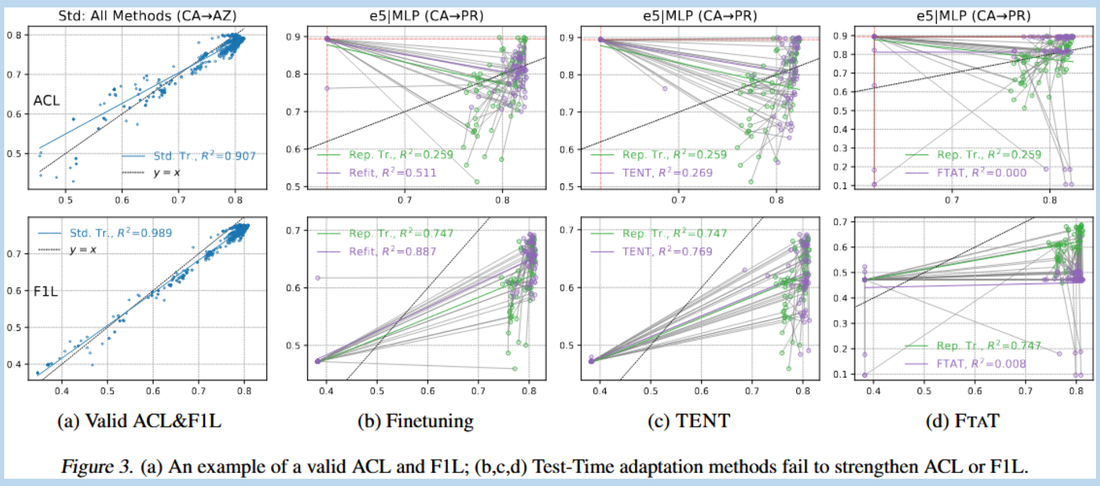


(c) Optimal e5 HP (F1)

Discussion

- LLMs are **not inherently trustworthy**: model collapse can occur even with stable hyperparameters.
- **Performance metrics may be misleading**, especially for ID/OOD robustness.
- **Test-time adaptation (TTA)** fails to restore robustness

TTA methods are ineffective



(a) Valid ACL&F1L (b) Finetuning (c) TENT (d) FTAT

Figure 3. (a) An example of a valid ACL and F1L; (b,c,d) Test-Time adaptation methods fail to strengthen ACL or F1L.

Prospectives

- Integrate OOD generalization as a task into MTEB.
- Need of specialised data curation for text embeddings
- Provide theoretical justification for upper bounds.

Trustworthy AI Summit 2025 - Poster Proposal

Authors

Michele Alberti (CEA LIST), François Bobot (CEA LIST), Zakaria Chihani (CEA LIST), Alban Grastien (CEA LIST), Julien Girard-Satabin (CEA LIST), Aymeric Varasse (Software Heritage)

Abstract

Formal verification of machine learning suffers from a fragmented ecosystem of tools and serious limitations on the scope of verifiable properties. This poster presents approaches to lift those limitations by proposing a specification language and automated graph edition techniques that embed parts of the specification directly within ONNX networks.

Detailed presentation

Extending formal verification of machine learning

The formal specification and verification of machine learning programs saw remarkable progress in less than a decade, leading to a profusion of tools. However, diversity may lead to fragmentation, resulting in tools that are difficult to compare, except for very specific benchmarks. Furthermore, this progress is heavily geared towards the specification and verification of a certain class of property, that is, *local robustness properties*. But while provers are becoming more and more efficient at solving local robustness properties, even slightly more complex properties, involving multiple neural networks for example, cannot be expressed in the input languages of winners of the International Competition of Verification of Neural Networks. In this poster, we present a specification language, suitable for modelling complex properties on neural networks, support vector machines and boosted trees. We show on concrete use-cases how specifications written in this language are automatically translated to queries to state-of-the-art provers, notably by using automated graph editing techniques, making it possible to use their off-the-shelf versions.

Detailed presentation

Automated graph editing

Let nn_1 and nn_2 be NNs with one (resp. two) inputs and one output. Consider the following computation

$$nn_2@@(nn_1@@(x_1), x_1 + \epsilon) + nn_1@@(x_0)$$

nn_1 is evaluated on x_0 and x_1. nn_2 first input is the result of the previously defined nn_1 computation, its second input is $x_1 + \epsilon$. Let H be a valid quantifier-free linear arithmetic formula, defining arithmetic bounds on x_1, x_2, ϵ. Consider now the specification defined in2.

$$\forall x_0, x_1, \epsilon.\ H(x_0, x_1, \epsilon) \rightarrow \tag{1}$$

$$\overbrace{\underbrace{nn_2@@(nn_1@@}_{\text{multiple networks}}(x_1), \quad \underbrace{x_1 + \epsilon}_{\text{operation on the input}})}^{\text{composition of nns}} + nn_1@@(x_0) > 0 \tag{2}$$

This formula exhibit the following features which prevents them to be used in NN verification:

- $x_1 + \epsilon$ describes a computation on the input;
- nn_1 and nn_2 are multiple networks
- nn_2 computes the output of nn_1;

It is however possible to lift those limitations by defining a new NN nn_3 which *embeds part of the specification within its control flow*, following~3.

$$\forall x_0, x_1, \epsilon.\ H(x_0, x_1, \epsilon) \rightarrow nn_3@@(x_0, x_1, \epsilon) > 0 \tag{3}$$

Specifically, nn_3 encodes computation on the input, the composition of NNs and the arithmetic comparison on the output.

The key insight here is that _some expressions from the specification language can be encoded as neural network operators. It is thus possible to write complex specifications and to encode parts of the specification inside of the NN.

Applications

Implemented within the CAISAR open-source platform, that was matured within the Confiance.ai program, those approaches allow to specify and verify properties like δ—equivalence between neural networks, specifying a normalization pipeline and evaluate a composition of neural networks.

Authors:

Wellhöfer, Johannes; Mensch, Maria

4-line-abstract

The "Zertifizierte KI" project (ZKI) establishes standardized testing criteria and validation methods to ensure AI system quality and trustworthiness within a robust quality infrastructure (QI). It supports European and international standardisation efforts by developing technical rules and frameworks and facilitating collaboration between academic experts, industry practitioners, and standardisation organizations. The project focuses on operationalizing specifications for AI quality standards, addressing transparency, explainability, risk identification, uncertainties, and trustworthy AI services - it does not certify AI systems or develop datasets. Specifications are accessible for free, promoting confidence and compliance with AI integration across sectors.

The "Zertifizierte KI" (ZKI, German for "Certified AI") project is focused on establishing standardized testing criteria and methods for ensuring the quality and trustworthiness of AI systems. This includes the development of validation methods and testing tools that can be applied to AI applications. This perspective is integrated in the architectural understanding of a so-called quality infrastructure (QI): A system of

a) technical specification and standardisation of given contextual demands (societal laws, natural laws, etc.) that create rationalisation and coordination benefits,
b) application of these specifications and standards by actors,
c) auditing of the proper application of, and adherance to, specifications and standards by third party auditors, handing out specific certificates, and
d) accreditation and control of third party auditors by an accreditation office.

AI systems shape many aspects of every-day-life and their impact on business activities is progressing rapidly. The QI has been profoundly impacted by the ramifications of this development. The development of AI systems that can effectively adhere to the demands of legal and scientific contexts poses significant challenges, particularly given the nascent state of the technology. In addition, the constant further development of AI systems makes auditing more difficult.
The potential for employing AI systems as a means of auditing other AI systems is currently under evaluation.

QI and its surrounding infrastructure are built on a system that provides trust and security, which is why the above challenges require quick solutions. These solutions must reduce potential disruptions and unlock potential as quickly as possible.

The project ZKI has been created to support the official (European and international) standardisation system.
While the European standardization system prioritizes the development of harmonized European standards in accordance with the European Commission's standardization mandate, the ZKI project establishes crucial supportive structures for QI that would not have been otherwise available.

A central feature of the project is the collaboration between academic experts, industry practitioners, and standardisation organizations. Through this collaboration, the aim of the project is to jointly develop frameworks, testing procedures, and tools that promote trustworthy AI system deployment, leveraging their societal and economic benefits.
One essential part of the integration of community knowledge with the project has been the options for actors to engage with the ZKI project. Specific mention should be given to two integrative measures.
The associated partners – actors that committed to engage in the project and help further its development and adoption.
And secondly the multiple topical user groups each addressing distinct aspects of the ZKI scope – cloud-based AI services; neural network transparency; foundation models; and operationalization of the AI-Act. These user groups each met in multiple workshops to help integrate the ZKI deliverables precisely with community learnings and demands.

A key aspect of the project is its contribution to the implementation of the European AI regulation (mainly the European AI Act). This includes operationalizing specifications for quality standards and ensuring compliance with legal requirements. The project is the

conduit for transferring identified standardisation needs into formal specifications, procedures and international standards.

It is important to emphasize that the project itself does not certify AI systems or AI models. It does not develop data sets for AI training or AI test tools. Instead, it focuses on identifying, understanding, interpreting, communicating and coordinating requirements within the QI in relation to AI systems. It develops technical rules for the development of reference and test datasets, determines the context for the trustworthy implementation of AI systems, and how to perform testing of AI models – to give some examples. We realize this contribution through the development of various DIN specifications[1]. These specifications define technical rules and promote common technical process understanding.

These specifications were published as part of the project and address various aspects such as the explainability of AI systems, risk identification and analysis during the lifecycle of AI systems, and the quantification of uncertainties in machine learning. The specifications are designed in such a way that they can be integrated into the existing QI – their makeup and logic slotting into best practice and know structures of points a) through c) of the QI description listed above. These specifications are available free of charge to promote their use and quick adoption.

Another key aspect of the project is transparency in neural networks and trustworthy AI cloud services, reflecting the broader context of AI standardisation efforts. Workshops and standardisation projects are a fundamental part of ZKI. They identify stakeholder needs and facilitate active participation from various sectors.

ZKI's contribution is undeniable. It plays a pivotal role in establishing a robust QI, providing the foundation for the certification of AI systems according to established standards. This, in turn, fosters trust in AI technology.

[1] Some examples of DIN specifications (DIN SPECs) from the ZKI project:
1) DIN SPEC 92001-3: Life cycle processes and quality requirements - Part 3: Explainability;
2) DIN SPEC 92005: Artificial intelligence - quantifying uncertainties in machine learning;
3) DIN SPEC 92006: Artificial intelligence - Requirements for AI testing tools

Trustproofer: assisting operationalised AI System trustworthiness

Mattheos Fikardos, Yiannos Paranomos, Katerina Lepenioti, Dimitris Apostolou and Gregoris Mentzas

Abstract: Trustproofer is an innovative neuro-symbolic framework that combines AI agents with human oversight to automate and enhance the development of trustworthy AI systems. By leveraging a Knowledge Graph (KG) constructed from documentation cards, agents assist in identifying, assessing, exploring, and enhancing trustworthiness across AI models and datasets.

While AI's rapid advancements drive widespread adoption across critical areas, the increasing number of harmful incidents has prompted regulatory responses and an academic emphasis on fostering trust in AI systems. The existing literature offers numerous surveys, reviews and frameworks that provide guidelines, methodologies and technical solutions to increase the trustworthiness of AI [1-4]. Moreover, organisations have compiled extensive catalogues of tools designed to aid practitioners in implementing TAI principles. For instance, RAND Europe lists 233 tools, the OECD catalogues over 900 tools, and Confiance AI provides more than 180 guidelines and tools. Furthermore, some approaches advocate the documentation of AI system information in the form of structured cards (e.g. Model/Data cards) [5-7]. This approach aims to report related data and increase transparency, aligned with the EU's AI Act requirement to document high-risk systems [8]. Despite the abundance of legislation, methodologies, and tools, a considerable gap exists between theory and practice. Many practitioners find themselves overwhelmed by the available resources and technical expertise required to operationalise trustworthiness. Our current work seeks to bridge this gap and empower practitioners to effectively utilise available resources to increase their AI systems' trustworthiness. Building on our previous work, specifically the Trustworthiness Optimization Process (TOP) [9], we introduce Trustproofer, an agentic framework designed to assist practitioners in creating trustworthy AI systems. TOP is structured around four main stages that integrate related approaches from the literature (e.g. documentation cards, risk management, metrics, enhancement methods) to assess and enhance AI systems in terms of trustworthiness.

Having TOP as the foundation, Trustproofer employs multiple agents that collaborate with a human supervisor to execute and automate key aspects of TOP. The architecture of Trustproofer is neuro-symbolic, leveraging symbolic representations and neural (AI) agents. Specifically, a Knowledge Graph (KG) is constructed based on the documentation cards, which represent the AI system and the available methods for assessing and enhancing trustworthiness. The AI agents, realised by a Large Language Model (LLM), leverage tools to interact with the KG and the AI system to perform actions across the stages of TOP. The agents are built with open-source models (e.g. Mistral, Deepseek) and libraries such as CrewAI and Langchain, while Neo4j was selected for the KG. Initially, in the Identify stage, agents can engage in dialogue with the system's stakeholders, asking questions to record the documentation cards. The documented information is essential for the subsequent stages, providing the necessary context to the agents and the assessment and enhancement actions. During the Assess stage, metrics are used to quantify the performance of AI Models and datasets, while the agents can use external tools, such as a risk management framework, to derive the corresponding risks. The results of these assessments are compiled into reports for review by the human supervisor, providing a detailed

overview of potential shortcomings and ensuring transparency. In the Explore stage, agents query the KG to discover applicable enhancement methods based on relationships between system characteristics and available techniques. These methods are then grouped into solution sets—combinations of multiple approaches—that are applied to a non-production version of the AI system. After applying these solutions, the system is reassessed to evaluate their impact. A multi-criteria decision-making approach is employed to compare solution sets, taking into account user preferences for specific trustworthiness characteristics and addressing potential conflicts between them. This comparison and the results are presented to the human supervisor to select a solution to be applied, enabling informed decision-making. Finally, in the Enhance stage, the selected solution is applied to the AI system. This triggers updates to the documentation cards and ongoing monitoring of the system to track its performance and ensure sustained and acceptable levels of trustworthiness.

While Trustproofer automates many aspects of the process, it maintains a human-in-the-loop approach, allowing the supervisor to oversee agent activities, provide guidance, and intervene as needed. This balance between automation and human oversight ensures that Trustproofer remains both efficient and aligned with user needs, ultimately supporting practitioners in their efforts to build trustworthy AI systems.

[1] Li, B.; Qi, P.; Liu, B.; Di, S.; Liu, J.; Pei, J.; Yi, J.; Zhou, B. Trustworthy AI: From Principles to Practices. ACM Comput. Surv. 2023, 55, 1–46.

[2] Díaz-Rodríguez, N.; Del Ser, J.; Coeckelbergh, M.; de Prado, M.L.; Herrera-Viedma, E.; Herrera, F. Connecting the dots in trustworthy Artificial Intelligence: From AI principles, ethics, and key requirements to responsible AI systems and regulation. Inf. Fusion 2023, 99, 101896.

[3] Kaur, D.; Uslu, S.; Rittichier, K.J.; Durresi, A. Trustworthy Artificial Intelligence: A Review. ACM Comput. Surv. 2022, 55, 1–38.

[4] Hupont, I.; Fernández-Llorca, D.; Baldassarri, S.; Gómez, E. Use Case Cards: A Use Case Reporting Framework Inspired by the European AI Act. arXiv 2023, arXiv:2306.13701.

[5] Golpayegani, D.; Hupont, I.; Panigutti, C.; Pandit, H.J.; Schade, S.; O'Sullivan, D.; Lewis, D. AI Cards: Towards an Applied Framework for Machine-Readable AI and Risk Documentation Inspired by the EU AI Act. In Privacy Technologies and Policy; Jensen, M., Lauradoux, C., Rannenberg, K., Eds.; Lecture Notes in Computer Science; Springer Nature Switzerland: Cham, Switzerland, 2024; Volume 14831, pp. 48–72.

[6] Mitchell, M.; Wu, S.; Zaldivar, A.; Barnes, P.; Vasserman, L.; Hutchinson, B.; Spitzer, E.; Raji, I.D.; Gebru, T. Model Cards for Model Reporting. In Proceedings of the Conference on Fairness, Accountability, and Transparency, Atlanta, GA, USA, 29–31 January 2019; ACM: Atlanta, GA, USA, 2019; pp. 220–229.

[7] Pushkarna, M.; Zaldivar, A.; Kjartansson, O. Data Cards: Purposeful and Transparent Dataset Documentation for Responsible AI. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, Seoul, Republic of Korea, 21–24 June 2022; pp. 1776–1826.

[8] EU AI Act: First Regulation on Artificial Intelligence. Topics|European Parliament. 8 June 2023. Available online: https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence (accessed on 2 June 2025).

[9] Fikardos, M., Lepenioti, K., Apostolou, D., & Mentzas, G. (2025). Trustworthiness Optimisation Process: A Methodology for Assessing and Enhancing Trust in AI Systems. *Electronics*, *14*(7), 1454.

Towards Compliance with the EU AI Act:

Insights from the Confiance.ai Program and Beyond

Author

Romane Vernhes, Consultante Conformité IA, Practice AI for Business Sopra Steria Next

Guillemette Jahn, Consultante senior Space, Defence and Security, Sopra Steria Next

(potentially: in partnership with Numeum)

Abstract

This poster explores how the EU AI Act can be implemented in practice, based on lessons learned from the French Confiance.ai program — providing valuable insights, methods, and technical assets. These contributions now support broader efforts to build governance and processes for trustworthy and compliant AI solutions.

Context: The Challenge of Operational Compliance

The AI Act creates a legal framework that governs the use of artificial intelligence in the EU, especially for high-risk systems. It defines detailed obligations related to transparency, risk management, technical robustness, data quality, documentation, and human oversight. Complying with this regulation requires more than legal interpretation. It calls for structured implementation processes and governance models that can translate regulatory principles into day-to-day practices — across diverse sectors and organisations.

What the Confiance.ai Program Contributes to

Confiance.ai, a French national program launched in 2021 federating industrial, academic and scientific players, was designed to accelerate the adoption of trustworthy AI systems, specially into critical applications. Over three years, the program produced a body of knowledge, tools, and demonstrators focused on topics that now directly echo AI Act requirements. These contributions represent useful reference points and a toolbox for organisations shaping their AI Act compliance strategies.

A Step Toward Operational Readiness

Although the Confiance.ai program was not designed for regulatory purposes, it laid key foundations that are now proved relevant for compliance. Its outputs can support the creation of early governance models and structured documentation. To move from foundation to implementation, organisations must also address broader dimensions such as operational governance, coordination between teams, tooling, and integration. These practices are critical to ensure that trustworthy AI can be deployed at scale, and in line with the AI Act's expectations.

Continuity Through Standards

With technical standards now being drafted by CEN-CENELEC JTC21 to complement the AI Act, ensuring continuity with existing work becomes essential. The experience accumulated and the expertise within Confiance.ai — and now shared via the European Trustworthy AI Foundation (ETF) — can contribute to these standards and make sure they reflect real industrial needs and practices. This connection will help organisations to comply to the regulation more easily, with shared tools and clear expectations.

Conclusion

The path to compliance with the AI Act requires more than new rules — it requires practical foundations. The results of Confiance.ai offer a structured base to build upon, supporting the development of trustworthy, compliant AI systems. By connecting regulatory requirements with field-tested approaches, these insights contribute to a broader effort to implement responsible AI at scale.

Privacy Amplification Through Synthetic Data: Insights from Linear Regression

Clément Pierquin [1 2] **Aurélien Bellet** [3] **Marc Tommasi** [2] **Matthieu Boussard** [1]

Abstract

Synthetic data inherits the differential privacy guarantees of the model used to generate it. Additionally, synthetic data may benefit from *privacy amplification* when the generative model is kept hidden. While empirical studies suggest this phenomenon, a rigorous theoretical understanding is still lacking. In this paper, we investigate this question through the well-understood framework of linear regression. First, we establish negative results showing that if an adversary controls the seed of the generative model, a single synthetic data point can leak as much information as releasing the model itself. Conversely, we show that when synthetic data is generated from random inputs, releasing a limited number of synthetic data points amplifies privacy beyond the model's inherent guarantees. We believe our findings in linear regression can serve as a foundation for deriving more general bounds in the future.

1. Introduction

Trustworthy AI aims to ensure that artificial intelligence systems are reliable, fair, transparent, and aligned with ethical and legal standards. A core component of this trustworthiness is the protection of individuals' privacy, especially when AI models are trained on sensitive personal data.

Differential privacy (DP) (Dwork and Roth, 2014) has become the gold standard for privacy-preserving data analysis. Training machine learning models with DP guarantees can be achieved through various techniques: *output perturbation* (Chaudhuri et al., 2011; Zhang et al., 2017b; Lowy and Razaviyayn, 2024), which adds noise to the non-private model; *objective perturbation* (Chaudhuri et al., 2011; Kifer et al., 2012; Redberg et al., 2023), which introduces noise into the objective function; and *gradient perturbation*, which injects noise into the optimization process, as in DP-SGD (Song et al., 2013; Bassily et al., 2014; Abadi et al., 2016; Feldman et al., 2018). Once trained, the model can be safely released, with its privacy guarantees extending to all subsequent uses thanks to the post-processing property of DP. This is particularly relevant for *differentially private generative models* (Zhang et al., 2017a; Xie et al., 2018; McKenna et al., 2019; Jordon et al., 2019; McKenna et al., 2021; Lee et al., 2022; Dockhorn et al., 2023; Bie et al., 2023), where the synthetic data they produce inherits the same privacy guarantees as the model itself.

Empirical studies, however, suggest that synthetic data may offer even stronger privacy protection than the theoretical guarantees provided by the model (Annamalai et al., 2024). This suggests that certain structural properties of the data or the generative process itself could contribute to an implicit privacy amplification effect. One possible intuition is that the privacy leakage might be reduced when the number of released synthetic data points is "small" relative to the complexity of the generative model. However, to the best of our knowledge, no existing work has formally established the existence of such a privacy amplification effect, and a rigorous quantification of differential privacy in synthetic data release remains an open question.

To address this gap, this paper takes an initial step towards developing a theoretical framework for quantifying privacy in synthetic data release. We focus on the well-studied setting of (high-dimensional) linear regression trained via a least-squares objective as a simple case study. This model has the advantage of being analytically tractable but sufficiently expressive to capture phenomena observed in more complex models—such as double descent in overparameterized regimes (Hastie et al., 2022) and, more recently, model collapse in generative AI (Dohmatob et al., 2024; Gerstgrasser et al., 2024).

We rely on the f-Differential Privacy (f-DP) framework (Dong et al., 2022), which provides a flexible and robust approach to privacy analysis, allowing precise characterizations of privacy guarantees through trade-off functions. When these trade-offs functions are difficult to interpret, we also express privacy guarantees in the Rényi differential privacy (RDP) framework (Mironov, 2017).

[1]Craft AI, Paris, France [2]Université de Lille, Inria, CNRS, Centrale Lille, UMR 9189 CRIStAL, F-59000 Lille, France [3] Inria, Université de Montpellier, INSERM. Correspondence to: Clément Pierquin <clement.pierquin@craft-ai.fr>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

Our results are two-fold. First, we present negative results in scenarios where an adversary controls the seed of the synthetic data generation process. Specifically, we show that the adversary can leverage this control to achieve privacy leakage equivalent to the bound imposed by post-processing the model using only a single synthetic sample. Second, we analyze the privacy guarantees when synthetic data is generated from random inputs to a private regression model obtained via output perturbation. We demonstrate that privacy amplification is possible in this setting, depending on the model size and the number of released synthetic samples.

Our findings highlight the critical role of the randomness given as input to the model, which must remain concealed from the adversary in order to enable privacy amplification. While the practical impact of our results is limited, we believe they offer valuable insights and lay the groundwork for a deeper understanding of synthetic data privacy in more complex machine learning models.

Related work. To the best of our knowledge, existing methods for differentially private synthetic data generation rely on learning a differentially private generative model (Hu et al., 2024). Early approaches focused on marginal-based techniques for tabular data, where a graphical model—such as a Bayesian network—is privately estimated from data and then used to generate new samples (Zhang et al., 2017a; McKenna et al., 2019; 2021). More recent methods extend to other data modalities, leveraging expressive neural network-based generative models—like GANs and diffusion models—trained with differentially privacy (Xie et al., 2018; Jordon et al., 2019; Lee et al., 2022; Dockhorn et al., 2023; Bie et al., 2023). A key advantage of neural networks is the availability of general differentially private training algorithms, such as DP-SGD (Abadi et al., 2016) and PATE (Papernot et al., 2017), which can be applied across various generative models. Crucially, all these methods rely on the post-processing theorem to ensure the privacy guarantees of the generated synthetic data—but it remains unclear whether this guarantee is tight or potentially overly conservative.

In principle, one could deviate from this dominant approach by adding noise directly to the data generated by a (non-private) generative model. In such cases, the overall privacy loss would scale with the number of released data points due to the composition property of differential privacy. However, this approach would require strong and often unrealistic assumptions about the data. Most critically, it would lead to significant utility loss—particularly for high-dimensional perceptual data such as images, where even small perturbations can severely degrade semantic content and downstream performance. To our knowledge, no successful applications of this approach have been demonstrated in practice.

Interestingly, our results suggest that differentially private generative models may offer the best of both worlds: the post-processing guarantee, which strictly bounds the privacy leakage when releasing a large number of samples, *and simultaneously* a privacy guarantee that scales with the number of released data points, which is more favorable when only a few samples are released.

Our results relate to the concept of *privacy amplification* (Balle et al., 2018; Feldman et al., 2018; Erlingsson et al., 2019; Cyffers and Bellet, 2022), which leverages the non-disclosure of certain intermediate computations to strengthen the privacy guarantees of existing mechanisms. We note that the form of amplification we study in the context of synthetic data release differs from privacy amplification by iteration (Feldman et al., 2018). In that setting, the final model is released after private training. In contrast, our approach withholds the model entirely and releases only synthetic data generated from random inputs to the model, introducing an additional layer of privacy protection.

We conclude our discussion of related work by mentioning a recent study that shows synthetic data can satisfy differential privacy guarantees without formal guarantees for the generative model itself (Neunhoeffer et al., 2024). However, this work is limited to a simple model where the private training data is one-dimensional, and the synthetic data is generated from a Gaussian distribution with mean and variance estimated from the private data. In contrast, our paper addresses a different, more complex problem: we investigate the privacy guarantees associated with releasing the output of a differentially private model, specifically linear regression. In our case, we directly model the distribution of the output of linear regression for a random seed, which corresponds to a product of Gaussian matrices.

2. Conclusion & Perspectives

We have shown that there exists a privacy amplification phenomenon for synthetic data in the context of linear regression. However, there is no amplification when the adversary has control over the seed of the synthesizer.

This negative result could inform the development of tighter privacy auditing strategies for synthetic data release (Annamalai et al., 2024). By quantifying the degradation in privacy guarantees, our findings offer insights that can help design more robust auditing methods in adversarial settings.

Several important directions remain for future work, including deriving general privacy amplification bounds that hold in more general settings. Extending our analysis to more complex models such as neural networks is a crucial step toward making these theoretical results applicable to real-world scenarios.

References

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., and Zhang, L. (2016). Deep learning with differential privacy. In *CCS*.

Annamalai, M. S. M. S., Ganev, G., and Cristofaro, E. D. (2024). "what do you want from theory alone?" experimenting with tight auditing of differentially private synthetic data generation. In *USENIX Security 24*, pages 4855–4871.

Balle, B., Barthe, G., and Gaboardi, M. (2018). Privacy amplification by subsampling: Tight analyses via couplings and divergences. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Bassily, R., Smith, A. D., and Thakurta, A. (2014). Private Empirical Risk Minimization: Efficient Algorithms and Tight Error Bounds. In *FOCS*.

Bie, A., Kamath, G., and Zhang, G. (2023). Private GANs, revisited. *Transactions on Machine Learning Research*. Survey Certification.

Chaudhuri, K., Monteleoni, C., and Sarwate, A. D. (2011). Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12:1069–1109.

Cyffers, E. and Bellet, A. (2022). Privacy amplification by decentralization. In *AISTATS*.

Dockhorn, T., Cao, T., Vahdat, A., and Kreis, K. (2023). Differentially Private Diffusion Models. *Transactions on Machine Learning Research*.

Dohmatob, E., Feng, Y., and Kempe, J. (2024). Model collapse demystified: The case of regression. In *Advances in Neural Information Processing Systems*, volume 37, pages 46979–47013. Curran Associates, Inc.

Dong, J., Roth, A., and Su, W. J. (2022). Gaussian differential privacy. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(1):3–37.

Dwork, C. and Roth, A. (2014). The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3–4):211–407.

Erlingsson, U., Feldman, V., Mironov, I., Raghunathan, A., Talwar, K., and Thakurta, A. (2019). Amplification by shuffling: From local to central differential privacy via anonymity. In *SODA*.

Feldman, V., Mironov, I., Talwar, K., and Thakurta, A. (2018). Privacy amplification by iteration. *2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 521–532.

Gerstgrasser, M., Schaeffer, R., Dey, A., Rafailov, R., Sleight, H., Hughes, J., Korbak, T., Agrawal, R., Pai, D., Gromov, A., Roberts, D. A., Yang, D., Donoho, D. L., and Koyejo, S. (2024). Is model collapse inevitable? breaking the curse of recursion by accumulating real and synthetic data. *CoRR*, abs/2404.01413.

Hastie, T., Montanari, A., Rosset, S., and Tibshirani, R. J. (2022). Surprises in high-dimensional ridgeless least squares interpolation. *The Annals of Statistics*, 50(2):949 – 986.

Hu, Y., Wu, F., Li, Q., Long, Y., Garrido, G. M., Ge, C., Ding, B., Forsyth, D., Li, B., and Song, D. (2024). SoK: Privacy-Preserving Data Synthesis . In *2024 IEEE Symposium on Security and Privacy (SP)*, pages 4696–4713, Los Alamitos, CA, USA. IEEE Computer Society.

Jordon, J., Yoon, J., and van der Schaar, M. (2019). Pategan: Generating synthetic data with differential privacy guarantees. In *International Conference on Learning Representations (ICLR)*.

Kifer, D., Smith, A. D., and Thakurta, A. (2012). Private convex optimization for empirical risk minimization with applications to high-dimensional regression. In *COLT*, volume 23, pages 25.1–25.40.

Lee, J., Kim, M., Jeong, Y., and Ro, Y. (2022). Differentially private normalizing flows for synthetic tabular data generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 7345–7353.

Lowy, A. and Razaviyayn, M. (2024). Output perturbation for differentially private convex optimization: Faster and more general. arXiv:2102.04704.

McKenna, R., Miklau, G., and Sheldon, D. (2021). Winning the nist contest: A scalable and general approach to differentially private synthetic data. *Journal of Privacy and Confidentiality*, 11(3).

McKenna, R., Sheldon, D., and Miklau, G. (2019). Graphical-model based estimation and inference for differential privacy. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pages 4435–4444.

Mironov, I. (2017). Rényi differential privacy. In *2017 IEEE 30th Computer Security Foundations Symposium (CSF)*, pages 263–275.

Neunhoeffer, M., Latner, J., and Drechsler, J. (2024). On the formal privacy guarantees of synthetic data. In *Data Privacy Protection and the Conduct of Applied Research: Methods, Approaches and their Consequences*. National Bureau of Economic Research.

Papernot, N., Abadi, M., Erlingsson, U., Goodfellow, I. J., and Talwar, K. (2017). Semi-supervised knowledge transfer for deep learning from private training data. In *ICLR*.

Redberg, R., Koskela, A., and Wang, Y. (2023). Improving the privacy and practicality of objective perturbation for differentially private linear learners. In *NeurIPS*.

Song, S., Chaudhuri, K., and Sarwate, A. D. (2013). Stochastic gradient descent with differentially private updates. In *2013 IEEE Global Conference on Signal and Information Processing*.

Xie, L., Lin, K., Wang, S., Wang, F., and Zhou, J. (2018). Differentially private generative adversarial network. arXiv:1802.06739.

Zhang, J., Cormode, G., Procopiuc, C. M., Srivastava, D., and Xiao, X. (2017a). Privbayes: Private data release via bayesian networks. *ACM Trans. Database Syst.*, 42(4).

Zhang, J., Zheng, K., Mou, W., and Wang, L. (2017b). Efficient private ERM for smooth objectives. In *IJCAI*.

Feder: Privacy-Preserving Federated Learning Across Enterprises

Timon Sachweh
TU Dortmund University

Helen Kuhlmann
TU Dortmund University

Thomas Liebig
Lamarr Institute, Tapekuna UG

Abstract

We propose *Feder*, a privacy-preserving framework for cross-enterprise Federated Machine Learning (FML) that ensures GDPR compliance. Feder integrates Homomorphic Encryption for secure model aggregation and Federated Unlearning to support data removal rights. By addressing key legal and technical challenges, *Feder* enables scalable, decentralized machine learning across organizations without compromising data privacy or regulatory obligations.

I Introduction

As data-driven business processes become increasingly prevalent, the rise of decentralized data across enterprises creates both opportunities and challenges for Machine Learning (ML). Traditional centralized ML approaches, which aggregate data into a single repository, often conflict with regulations like the European Union's General Data Protection Regulation (GDPR), which restrict data sharing and storage. This makes cross-organizational model training legally and operationally difficult. Federated Machine Learning (FML) has emerged as a promising alternative, enabling decentralized model training without transferring raw data [1].

FML is particularly suited to cross-enterprise scenarios involving sensitive, siloed datasets — common in sectors like healthcare, finance, and manufacturing. By exchanging only model updates, FML supports GDPR principles such as data minimization and locality [6]. However, it still faces privacy challenges, including model inversion and inference attacks [5]. Moreover, GDPR's "right to be forgotten" introduces additional complexity that FML alone does not fully address.

To meet these challenges, recent research has introduced Homomorphic Encryption (HE) for secure computation on encrypted model parameters [3, 2], and Federated Unlearning to enable compliant data removal from models after training [4].

In this paper, we present *Feder* - a framework for Federated Encryption and Deletion Machine Learning for European Regulation-Integrated Systems. *Feder* integrates HE and federated unlearning to enable secure, GDPR-compliant FML.

We begin with an overview of related work, detail the *Feder* framework and conclude with an evaluation of its practical applicability in enterprise environments.

II Related Work

To address all previously described issues with GDPR-conformity is very challenging. Therefore, we build upon well established libraries and methodologies, which solve parts of the data privacy issues.

We will use typical Federated Learning methods as described in [1], but integrate the Multi-Party-Homomorphic Encryption (MHE) based on the Cheon-Kim-Kim-Song (CKKS) Scheme. The CKKS scheme is superior to other schemes like Brakerski, Gentry, Vaikuntanathan (BGV) because CKKS allow encrypting vectors of floating point numbers. A Go implementation called Lattigo is well suited and used in our framework to cover HE [2].

To be GDPR compliant, the option of deleting personal data must also be available. This also applies to derived information contained in ML models. There is a novel approach called FedSSU, which addresses unlearning of neural networks in a federated learning setting [4]. This approach will be used for the unlearning task in *Feder*.

III Proposed Framework: Feder

With our approach, we provide a single framework that addresses all problems for FML caused by GDPR regulations. An exemplary usage model of the *Feder* framework is shown in Figure 1. There are two companies

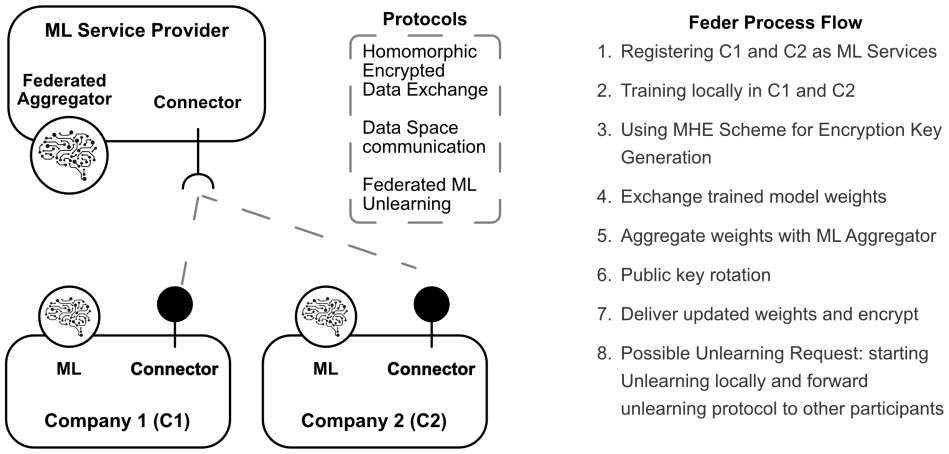


Figure 1: Exemplary production setup for running Feder in Cross Enterprises

C1 and *C2*, which have local data to train on. Additionally, there is the Federated Aggregator, which does the aggregation step in typical FML approaches. The *ML* component at the companies works as a plugin system with defined interfaces, such that the rest of the *Feder* framework can be used in a standardized manner for different ML architectures. *Feder* itself works similar to Eclipse Dataspace Connectors (EDC), with a standardized data protocol for the unlearning and model weight exchange.

Feder prioritizes the privacy of personal data throughout the entire federated learning process. As shown in the right-hand process flow, it begins with the registration of *C1* and *C2* at the *ML Service Provider*. Both parties (*C1*, *C2*) train ML models locally using a shared architecture suitable for FML. Simultaneously, each company's *Connector* generates cryptographic keys for the HE schema. The MHE schema establishes a shared public key.

Trained local model weights are encrypted using this key and sent to the *Federated Aggregator* for secure aggregation. Following aggregation, keys are rotated using MHE techniques to enable decryption on *C1* and *C2*, after which the updated global model is deployed back to them. This iterative cycle continues from step 2 as new data becomes available.

The unlearning process is triggered only upon a data deletion request, which is handled by the company that owns the data. Using the FedSSU method, the data is removed locally, and corresponding updates are made to the global model. The revised model is then redistributed to all participants, ensuring GDPR-compliant unlearning.

IV Conclusion

This paper introduced *Feder*, a privacy-centric framework that enables GDPR-compliant FML across enterprise boundaries by integrating HE and Federated Unlearning. Its modular connector-based architecture simplifies integration with existing systems, making privacy-preserving, decentralized machine learning both practical and scalable for real-world enterprise environments.

References

[1] Peter Kairouz et al. "Advances and open problems in federated learning". In: *Foundations and trends® in machine learning* 14.1–2 (2021), pp. 1–210.

[2] *Lattigo v6*. Online: `https://github.com/tuneinsight/lattigo`. EPFL-LDS, Tune Insight SA. Aug. 2024.

[3] Joon-Woo Lee et al. "Privacy-preserving machine learning with fully homomorphic encryption for deep neural network". In: *iEEE Access* 10 (2022), pp. 30039–30054.

[4] Yuhe Leng et al. "FedSSU: flexible and efficient decentralized unlearning for federated learning". In: *The Journal of Supercomputing* 81.8 (2025), pp. 1–26.

[5] Milad Nasr, Reza Shokri, and Amir Houmansadr. "Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning". In: *2019 IEEE symposium on security and privacy (SP)*. IEEE. 2019, pp. 739–753.

[6] Nicola Rieke et al. "The future of digital health with federated learning". In: *NPJ digital medicine* 3.1 (2020), p. 119.

Augmenting Security Operation Center With Artificial Intelligence and Machine Learning

Elies Gherbi

July 10, 2025

Abstract

The deployment of AI accelerated the detection process and provided deeper insight into malicious scripts underlying intent and attack vectors. In this document, we investigate practical uses of AI in cybersecurity workflows such as its application in Security Operations Centers (SOCs), with a primary focus on augmenting intrusion detection and automating alert triage. In addition, the application of AI extends beyond malware analysis, offering transformative potential in areas such as threat intelligence, incident response, and vulnerability management, thus reinforcing the overall cybersecurity posture against evolving threats. This document delves into these advancements, highlighting the role of AI in addressing critical cybersecurity challenges and paving the way for more resilient defense mechanisms.

1 Introduction

A Security Operations Center (SOC) is a centralized facility where dedicated security professionals build, maintain, and operate the architecture that monitors, detects, analyzes, and responds to cybersecurity incidents. Unifies people, processes, and technologies to maintain vigilance over an organization's networks, systems, and applications, thus strengthening its security.

SOCs are critical command centers that continuously monitor, detect, and respond to cyber threats. However, the deluge of raw logs and alerts can overwhelm human analysts, leading to high false-positive rates and slow reaction times. Augmenting SOC workflows with AI and machine learning enhances traditional intrusion detection by incorporating explainable AI (XAI) and uncertainty quantification, enabling transparent and reliable decision support. AI-driven alert triage further reduces noise and prioritizes genuine threats, cutting false positives, and focusing analyst effort on high-impact incidents. Finally, integrating Large Language Models (LLMs) automates and accelerates incident analysis, playbook generation, and contextualization, significantly reducing Mean Time To Acknowledge (MTTA) and Mean Time To Respond (MTTR).

The typical SOC workflow begins with the collection and normalization of logs from diverse sources (network devices, endpoints, applications), followed by event correlation and initial alert generation using SIEM or similar platforms. Analysts then triage these alerts classifying, prioritizing, and routing them for further investigation. High-confidence incidents trigger incident response playbooks, leading to containment, eradication, and recovery actions. Finally, post-incident reviews and continuous feedback loops refine detection rules and response procedures.

A central theme of this document is the critical role of Explainable AI in bolstering the trustworthiness of AI systems in cybersecurity. As AI models increasingly operate as black boxes, the lack of transparency and interpretability can undermine user confidence and limit the effectiveness of these technologies in real-world scenarios. By incorporating XAI techniques, security professionals can gain deeper insights into AI-driven decisions, enabling more informed, transparent, and trustworthy security measures.

In this work we aim to provide a structured approach to understand the current SOCs workflow and identify the usage and limits of AI applications, thus addressing the associated challenges, and proposing future research directions. By achieving these objectives, we aim to enhance the development and implementation of more resilient, transparent, and effective AI-based SOCs solutions.

2 AI Applications for modern SOC

2.1 AI based detection

Modern intrusion detection systems (IDS) leverage machine learning (ML) to enhance both signature and anomaly based detection. Supervised models (e.g., decision trees, SVMs, neural networks) achieve over 95 % accuracy on known attack classes in benchmark datasets, though performance often degrades on real-world traffic. Unsupervised techniques address this gap by modeling normal behavior and flagging deviations, enabling detection of zero-day attacks but at the cost of higher false-positive rates. Continuous research focuses on balancing detection efficacy with reduction of spurious alerts. However, as noted in a recent survey, these results are often achieved in benchmark data sets that may not fully represent real-world network traffic, which means that the effectiveness of a model can drop when facing truly novel attacks. This is where unsupervised learning becomes valuable: By clustering or learning patterns in unlabeled data, anomaly detection techniques can identify suspicious behavior that does not match any known profile, thus catching zero-day attacks. The trade-off is that unsupervised IDS tend to generate more false positives (benign anomalies mistaken for attacks) because anything unusual is flagged [PHDG23, Okd24, ASAAF25].

In practice, AI-powered IDS are a cornerstone of many organizations' security architecture today. They serve as early warning systems, raising alerts for security analysts when something suspicious is spotted. The role of AI in IDS is to sift through massive volumes of network logs and traffic in real time, intelligently differentiate benign anomalies from malicious ones, and continuously adapt to new attack tactics. Studies underscore that AI-based IDS can detect complex multi-stage attacks (like APTs) by correlating subtle deviations over time[NBTI23]

Moreover, by reducing false positives compared to naive anomaly detection, AI helps ensure that IDS alerts are meaningful and actionable. There are still challenges to address – for instance, ensuring IDS models remain effective as networks evolve, and defending the IDS itself from adversarial evasion (attackers crafting traffic to fool the model). Nonetheless, the consensus in recent research is that AI has become an indispensable component of high-performance IDS and will continue to drive improvements in intrusion detection capabilities.

2.2 AI Techniques for Alert Triage

Large organizations often receive thousands of security alerts per day from intrusion detection systems, antivirus, firewalls, and other monitoring tools. Security Operations Center analysts face the daunting task of triage these alerts i.e., determining which are true threats that merit immediate investigation and which are false positives or low priority. This process can be extremely time consuming and prone to errors due to the sheer volume of alerts (a problem known as 'alert fatigue'). AI techniques have been introduced to SOC workflows to automate and assist with alert triage, helping analysts focus on the most critical incidents [BDH16].

Supervised learning can rank alerts by threat likelihood [NBTI23] applies learning to rank algorithms so analysts focus on top-ranked, most malicious alerts. In the Active Learning for Alert Triage (ALAT) system, a model trained on analyst-labeled past alerts is continuously retrained with new feedback, steadily improving precision. By emulating expert decision patterns from historical incident data, AI-driven ranking markedly reduces the risk of overlooking critical alerts.

Unsupervised and hybrid triage methods group similar alerts or flag outliers without labels using Isolation Forest sifts alert logs to highlight stealthy anomalies and enabling novel attack detection but risking excess false positives. Probabilistic fusion combines multisource alerts for incident estimation, and provenance or knowledge-graph approaches assign threat scores by contextualizing alerts within broader attack chains [NBTI23, BDH16].

AI-assisted alert triage markedly reduces human review workload. The system automated over 50 % of daily alerts with high confidence. Thereby lowering analyst cognitive load and mean time to detect and respond. Operating continuously at machine speed, AI correlates logs and security feeds beyond human capacity. Such systems can automatically gather context such us related logs, threat intelligence and summarize why an alert is suspicious, further accelerating investigations.

3 Conclusion

In conclusion, AI-driven detection and triage systems have demonstrably transformed SOC operations by automating the first-pass filtering of alerts, prioritizing high-risk incidents, and significantly reducing analyst workload and mean time to response. Yet to fully realize their promise in live environments, these systems must be both accurate and trustworthy. Integrating explainable AI techniques ensures that analysts can inspect and validate model decisions based on the specific features, log entries, or graph structures that underpin each alert. While uncertainty quantification provides calibrated confidence scores that help distinguish genuine threats from benign anomalies and curb false-positive rates. Moreover, embedding large language models into SOC workflows enables natural language summarization of complex attack sequences, automated report generation, and interactive reasoning about threat scenarios. By marrying XAI, uncertainty metrics, and LLM-based reasoning, next-generation SOC platforms will not only enhance detection efficacy and triage precision but also foster analyst trust, facilitate continuous learning, and deliver actionable intelligence making AI augmentation both practical and indispensable in modern security operations.

Acknowledgement

This work has been supported by the French government under the "France 2030" program, as part of the SystemX Technological Research Institute R&D Program CYBELIA. The CYBELIA Program is co-funded by Airbus Protect and Réseau de Transport d' Electricité (RTE)

References

[ASAAF25] Abdullah Al Siam, Moutaz Alazab, Albara Awajan, and Nuruzzaman Faruqui. A comprehensive review of ai's current impact and future prospects in cybersecurity. *IEEE Access*, 13:14029–14050, 2025.

[BDH16] Michael Bierma, Justin E. Doak, and Corey M. Hudson. Learning to rank for alert triage. In *Proc. of the 11th IEEE International Conference on Cyber Security (ICS) Workshop*, 2016. Sandia National Laboratories Report SAND2016-3373C.

[NBTI23] Samuel Ndichu, Tao Ban, Takeshi Takahashi, and Daisuke Inoue. AI-assisted security alert data analysis with imbalanced learning methods. *Applied Sciences*, 13(3):1977, 2023.

[Okd24] Selcuk Okdem. Artificial intelligence in cybersecurity: A review and a case study. *Applied Sciences*, 14(22):10487, 2024.

[PHDG23] Andrea Pinto, Luis-Carlos Herrera, Yezid Donoso, and Jairo A. Gutierrez. Survey on intrusion detection systems based on machine learning techniques for the protection of critical infrastructure. *Sensors*, 23(5):2415, 2023.