

BUKHARA STATE UNIVERSITY · UZBEKISTAN

Reliable AI beyond the benchmark

Mohammad Sabokrou

Associate Professor, New Uzbekistan University · Visiting Researcher, OIST, Japan

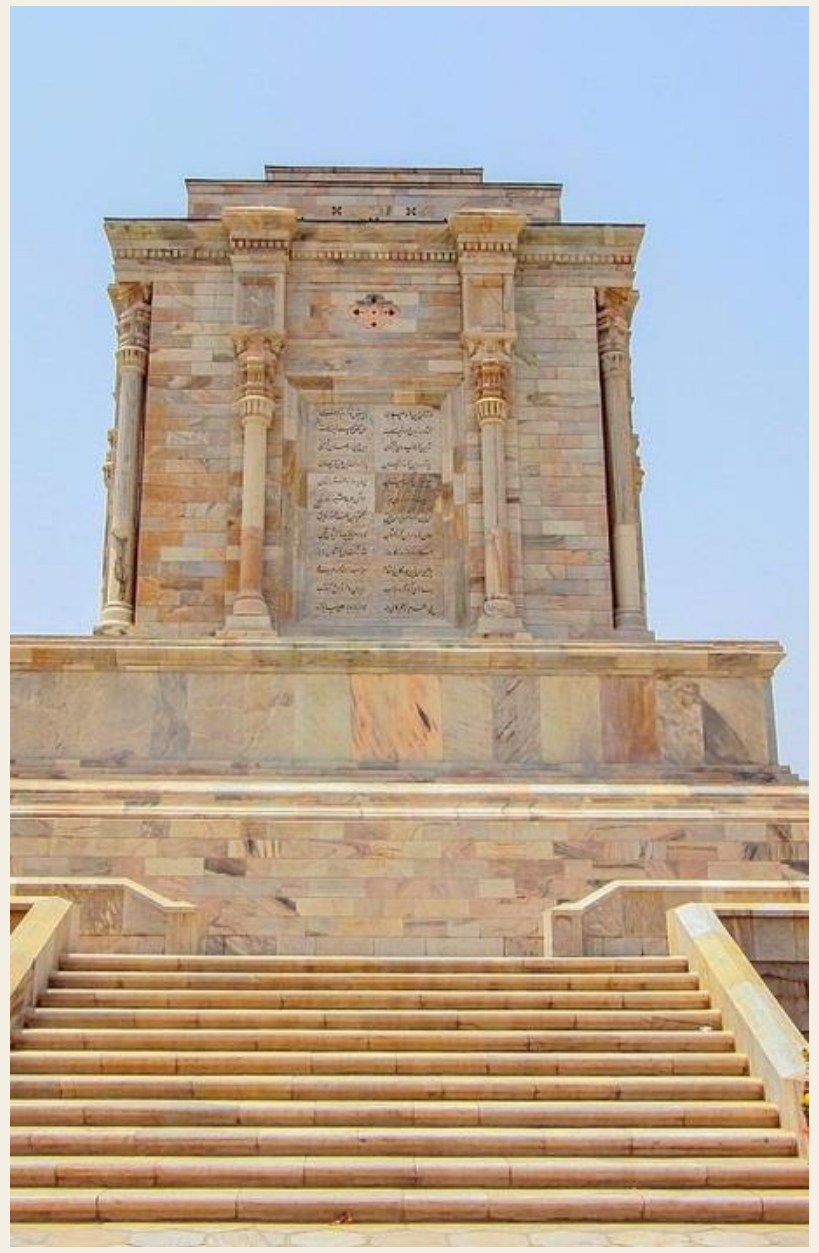
Photo: Po-i-Kalyan complex, Bukhara · Erik Albers, CC BY-SA 4.0, cropped

A PERSONAL CONNECTION

From Khorasan to Bukhara

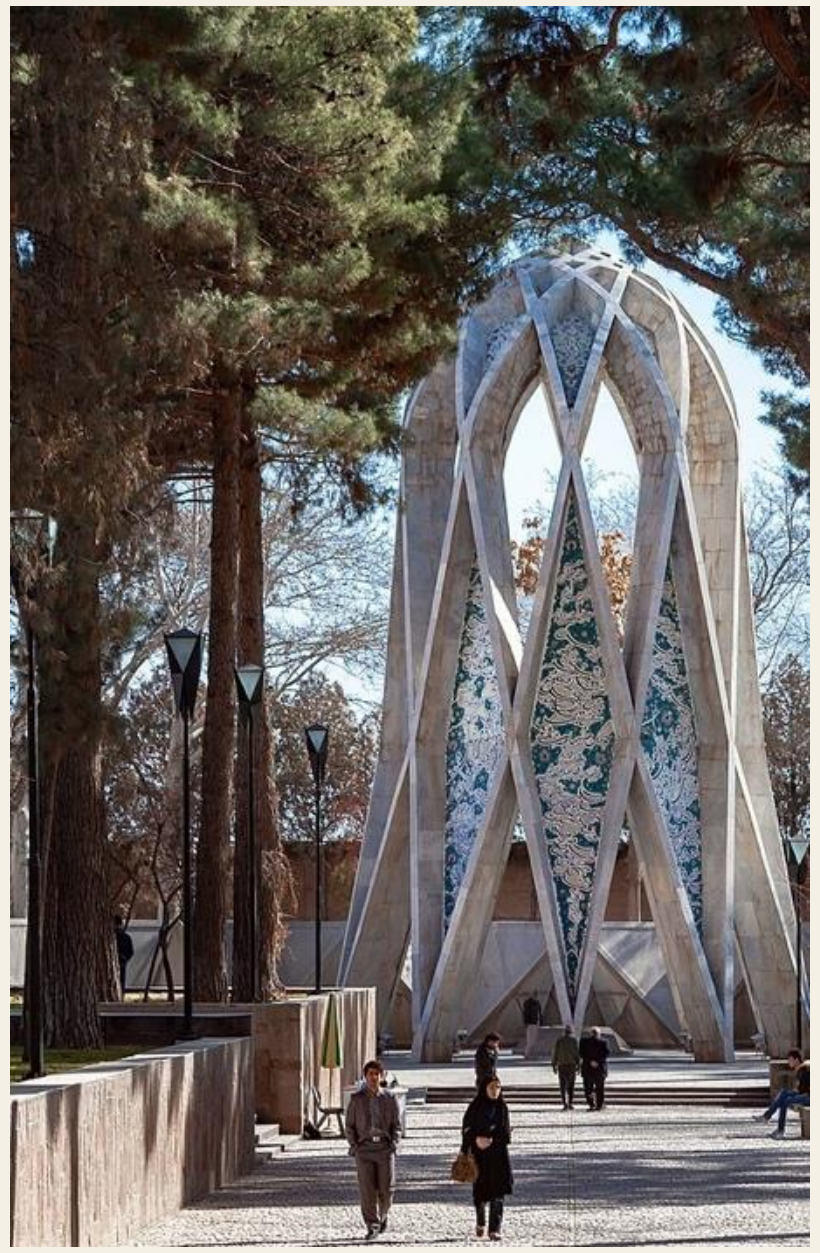


Gonabad, Khorasan, Iran



Ferdowsi

Tus



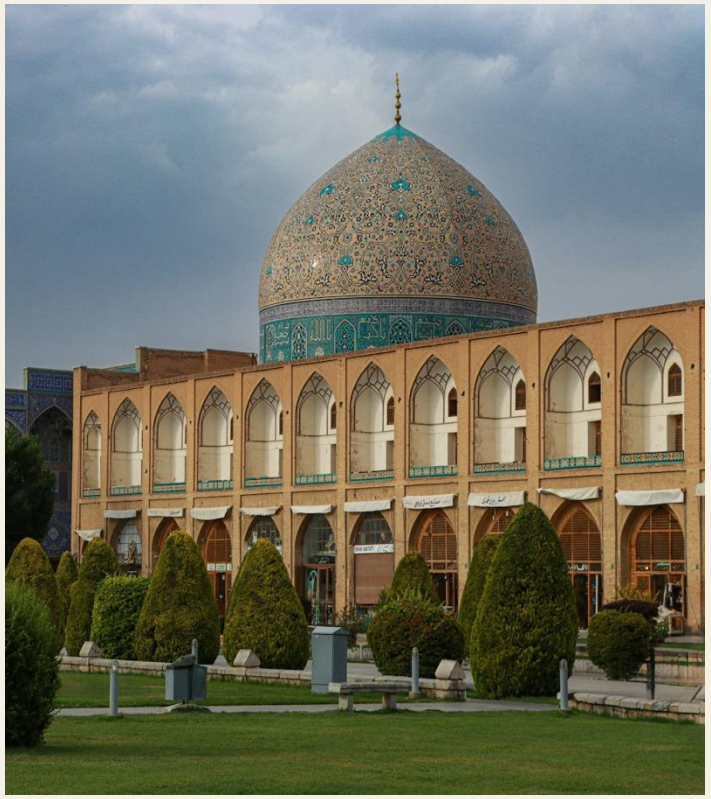
Khayyam

Nishapur

Map: © OpenStreetMap contributors. Photos: Mostafameraji (CC BY-SA 4.0), Ninara (CC BY 2.0).

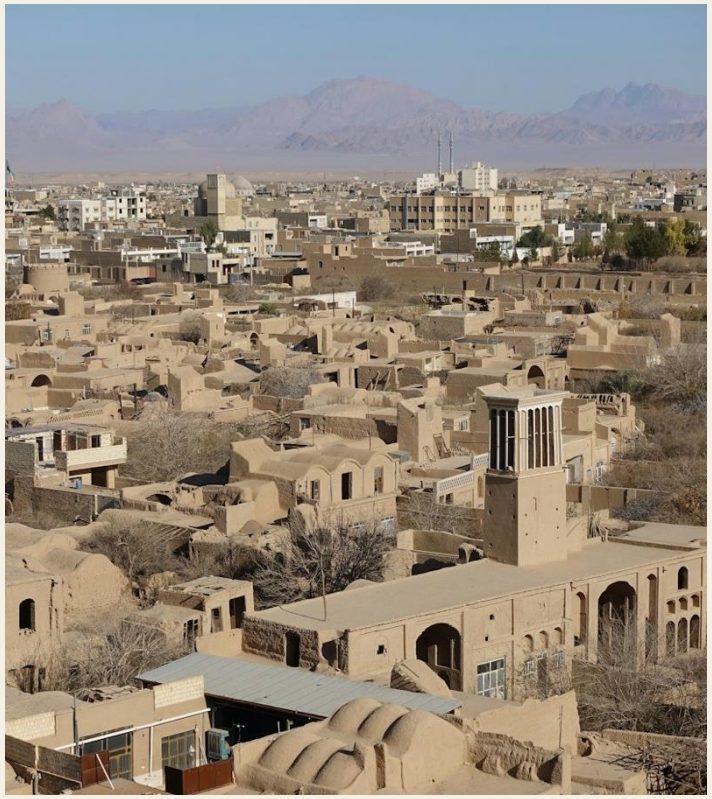
A PERSONAL CONNECTION

Iran, past and present



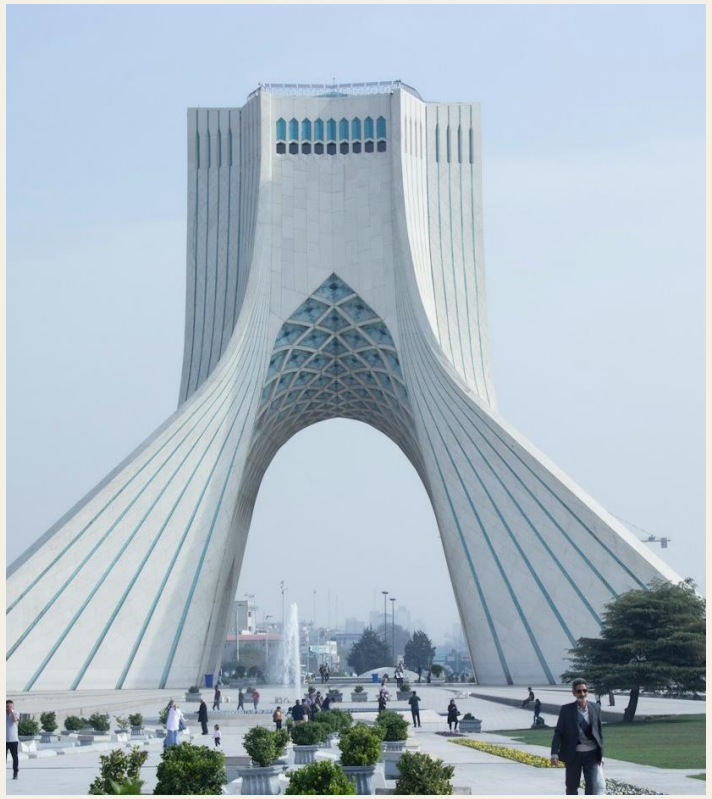
Isfahan

Tilework and public squares



Yazd

The historic desert city



Tehran

Azadi Tower, a modern landmark

Photos: Ali HSFT, Sofiya Popovych and Elham Abdi / Unsplash.

POETRY AND BUKHARA

Rudaki and Bukhara

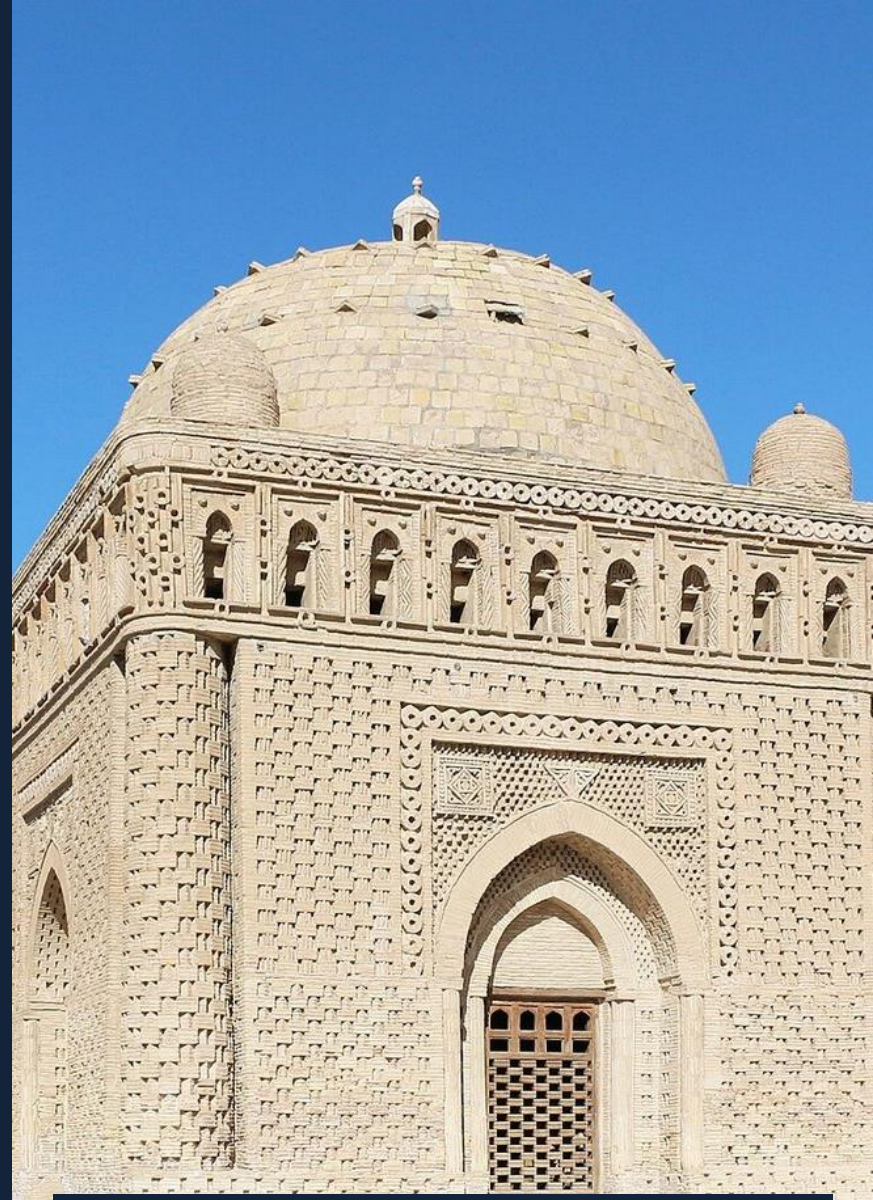
بوی جوی مولیان آید همی
یاد یار مهربان آید همی

The scent of Muliyan's stream returns,
and the memory of a dear friend.

ای بخارا شاد باش و دیر زی
میر زی تو شادمان آید همی

Thank you for welcoming me
to Bukhara.

Rudaki (public domain). Photo: Bgag, CC0, cropped.



Samanid Mausoleum, Bukhara
Shared literary traditions across
Central Asia and Iran

INTRODUCTION

Mohammad Sabokrou



I study what happens when a model meets conditions it was not trained to handle.

New Uzbekistan University

Associate Professor, Department of AI, School of Computing, since August 2026

OIST, Japan

Visiting Researcher; previously Staff Scientist, 2023–July 2026

Research focus and service

Trustworthy machine learning and computer vision · Area Chair, ICLR, NeurIPS and BMVC

ACADEMIC EXPERIENCE

Research and teaching across six countries

1	Iran	**IPM** PhD in AI, 2017 · AI research leadership
2	France	**Troyes University of Technology** Postdoctoral Researcher
3	Finland	**University of Oulu, CMVS** Senior Research Scientist
4	Japan	**OIST, Okinawa** Staff Scientist, now Visiting Researcher
5	Türkiye	**Istanbul Technical University** Visiting Faculty · Deep Learning, 2025
6	Uzbekistan	**New Uzbekistan University** Associate Professor, 2026 onwards



Highlighted: the six countries where I have studied, researched or taught. Markers show the cities.

Also taught at IPM (2019–2021) and Iran University of Science and Technology (2019) · sabokrou.github.io/cv

JAPAN

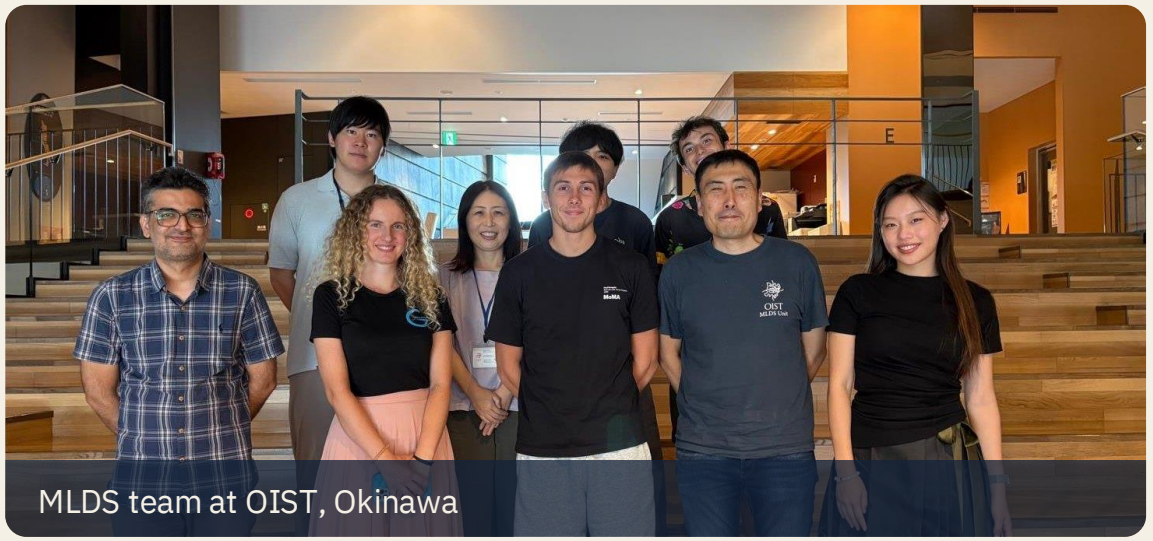
OIST in Okinawa

2011
Founded by the Government of Japan

5-year
PhD in science; no academic departments

English
Language of all teaching and research

>50%
Faculty and students from outside Japan


MLDS team at OIST, Okinawa







> **MLDS unit, led by Prof. Makoto Yamada**
> Staff Scientist, 2023–July 2026; now Visiting Researcher, continuing the joint work from Uzbekistan

Facts: oist.jp/about · team and campus photos from the presenter · bridge photo © OIST, CC BY 4.0

SUPPORTED BY JAPAN

Research funded by Japan

CURRENT · 2026–2030

¥18.33M

Breaking Boundaries: Robust, Domain-General Anomaly Detection with VLMs

JSPS KAKENHI Scientific Research (B) · PI · grant 26K02988

PREVIOUS · 2024–2026

¥4.68M

Investigating the Trustworthiness of Deep Pre-trained and Self-supervised Learned Model

JSPS KAKENHI Early-Career Scientists · PI · grant 24K20806

My travel to Bukhara is supported by Japan, with sincere thanks.

Official KAKEN records · amounts include direct and indirect costs · both grants hosted at OIST

UZBEKISTAN

New Uzbekistan University



2021
Founded in Tashkent by presidential decree

English
Main language of teaching and research

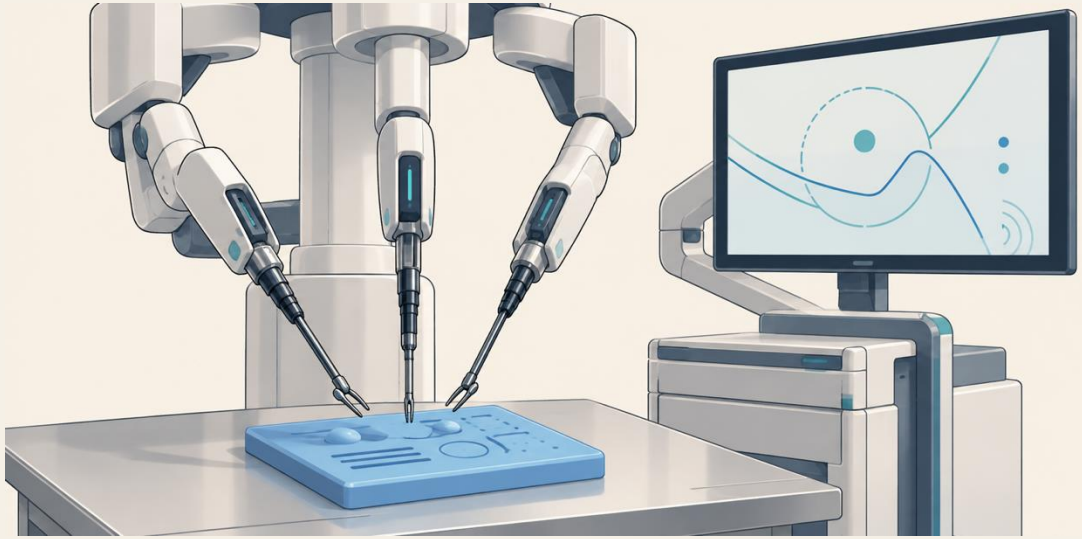
About the university
Uzbekistan’s first comprehensive research university, with partners such as MIT J-WEL and TUM International.

Aim
Integrate advanced research with transformative education, to develop the next generation of leaders, thinkers and innovators.

My role
Associate Professor, Department of AI, School of Computing; teaching Computer Vision and Data Science Fundamentals

Sources: newuu.uz/en/about · founding decree, June 2021 · NewUU faculty profile · photo: New Uzbekistan University

High accuracy. But can we trust it?



Autonomous surgery research

SRT-H: 8/8 successful ex vivo trials
A laboratory result requiring clinical validation



Autonomous driving

Driverless vehicles operate on public roads
What happens in a rare, unfamiliar situation?

High test performance leaves a gap: unfamiliar conditions, rare failures and safe recovery.

Robotics illustration: conceptual, AI-generated. Study: Johns Hopkins, Science Robotics 2025. Driving photo: Waymo.

RESEARCH INTERESTS

Trustworthy machine learning

Anomaly & OOD detection

Novelty, open-set and zero-shot anomaly detection

Robustness & security

Adversarial attacks, backdoors and trojan scanning

Privacy

Knowledge leakage, membership inference, memorization

Fairness & bias

Dataset bias, spurious correlations, shortcut learning

LLM & GenAI safety

Hallucination, generated-content detection, alignment

AI for safety

Video understanding for surveillance, driving and care

From classical deep networks (CNNs, RNNs) to large-scale models: LLMs, vision-language and diffusion models

Failures across generative AI systems

Hallucination

A paid government report cited papers that do not exist

Deloitte Australia · Oct 2025

Invented policy

An airline was held liable for a refund policy its chatbot made up

Moffatt v. Air Canada · 2024

Deepfake fraud

A video call with a deepfaked CFO led to a US$25 million transfer

Hong Kong · Feb 2024

Data leakage

A simple prompt made ChatGPT leak memorized training data

Nasr et al. · 2023

Prompt injection

Hidden text in a web page or email can take over an LLM agent

Greshake et al. · 2023

Hidden backdoors

Backdoored LLMs kept their trigger behavior through safety training

Hubinger et al. · 2024

Top row: incidents in deployed systems · bottom row: research demonstrations

THE QUESTION BEHIND MY RESEARCH

How can AI remain reliable outside the conditions it learned?

01

Recognizing the unknown

When should a model reject an input?

Unified survey, TMLR 2022
ALOCC, CVPR 2018

02

Generalizing to new domains

What should transfer when the domain changes?

UNODE, CVPR 2024
CRANE, BMVC 2026

03

Staying robust and secure

What happens under manipulation or hidden faults?

CBS, TMLR 2024
TRODO, NeurIPS 2024

01

Recognizing the unknown

A confident prediction can still concern an unfamiliar input.

FEATURED · UNIFIED SURVEY, TMLR 2022 · ALOCC, CVPR 2018

Reliability under distribution shift

A matched benchmark split

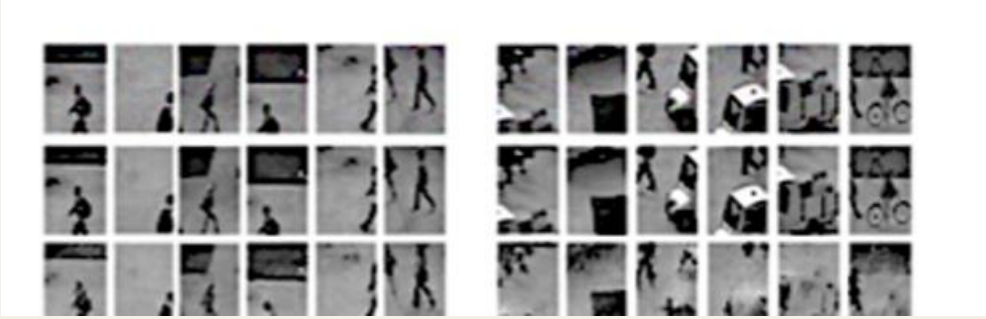


$$P_{\text{train}}(X, Y)$$

$$\approx P_{\text{test}}(X, Y)$$

Known classes
Similar statistics
A controlled split

Deployment changes the data



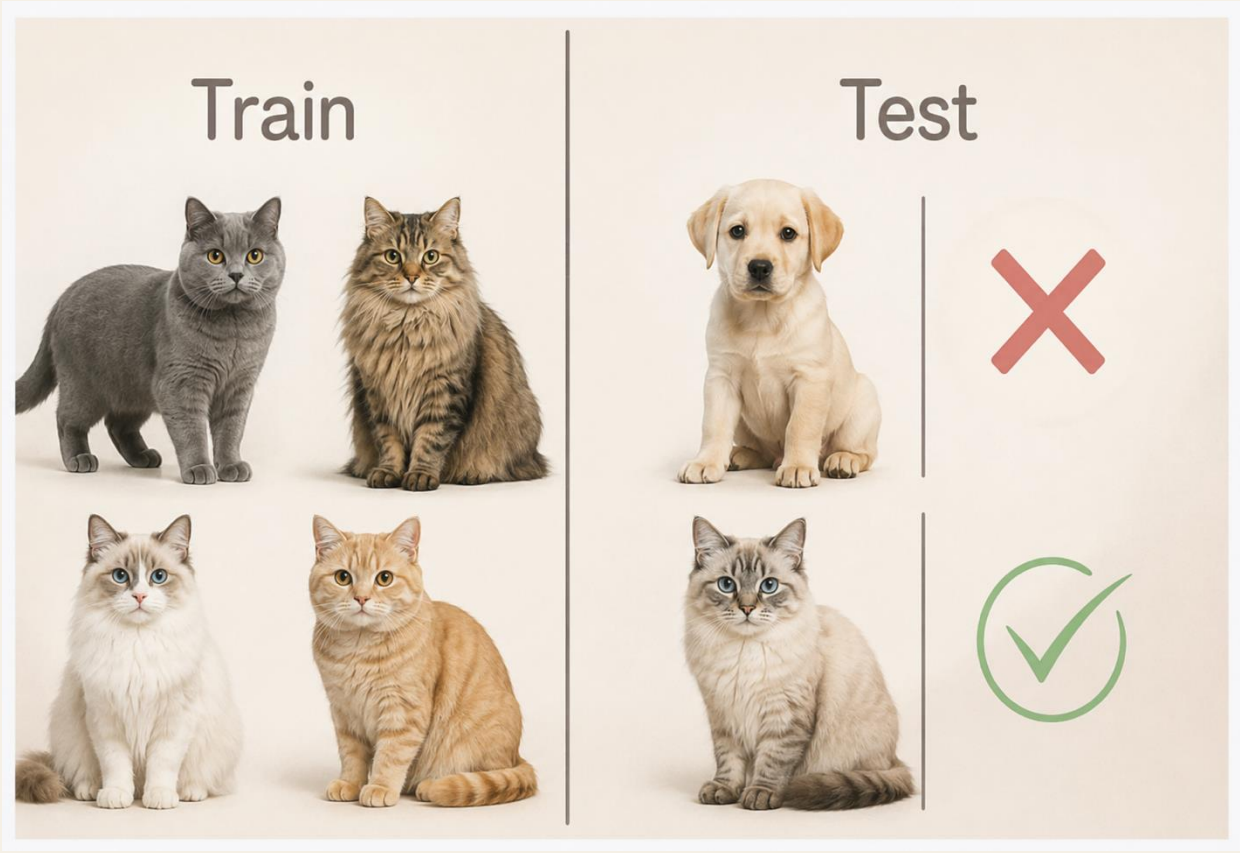
$$P_{\text{train}}(X, Y) \neq P_{\text{deploy}}(X, Y)$$

New cameras or lighting can shift inputs.
New objects can appear after training.

A test score describes the test distribution. Deployment may ask a different question.

RECOGNIZING THE UNKNOWN

Novelty and anomaly detection



The normal set depends on the task. A dog is “unknown” only in this cats-only illustration.

Learn what is normal

Training data describe the normal or known set. Here, that set contains cats.

Score a new input

Accept a familiar input. Flag an unfamiliar one for rejection or review.

The hard part

Real anomalies may be missing or very rare during training.

Related tasks, different assumptions

Task	Typical training setup	Required test-time decision
Open-set recognition	Labeled examples of known classes	Assign a known class or reject an unknown
Novelty detection	A clean normal training set	Distinguish normal from previously unseen patterns
OOD detection	A defined in-distribution reference set	Decide whether an input belongs to that distribution
Anomaly detection	Mostly normal data; contamination may occur	Detect unusual events or defects under a task definition

These families overlap. State the training labels, contamination assumptions and test split.

Two baseline scores for unfamiliar inputs

Mahalanobis distance

How far is the feature from normal data?

$$D_M(x) = \sqrt{(z-\mu)^\top \Sigma^{-1}(z-\mu)}$$

$$z = f(x)$$

Fit the mean μ and covariance Σ using normal training features.

Reject when the distance is large

$$D_M(x) > \tau_D$$

Maximum softmax probability

How confident is the known-class classifier?

$$c(x) = \max_k p_\theta(y = k \mid x)$$

Use the largest class probability from a classifier trained on known classes.

Reject when confidence is low

$$c(x) < \tau_c$$

Both are baselines. The representation and validation threshold determine what they can detect.

Why normal-only learning is difficult

Training contains normal examples only. Real anomalies may be rare, diverse or unknown.

Ordinary binary classification

If every training label is “normal”, the loss rewards assigning “normal” to every input.

The missing signal

The data do not show where “normal” should end. A low training loss can hide this problem.

$$\mathcal{L}(\theta) = -\frac{1}{n}\sum_{i=1}^{n} \log p_\theta(y = 1 \mid x_i)$$

CNNs can learn normal-only features. They need an objective that makes unfamiliar inputs distinguishable.

One-class learning changes the learning objective. Adversarial reconstruction is one possible approach.

Generative adversarial networks (GANs)

Generator G

Turns noise into synthetic examples. It learns to make the discriminator mistake them for real data.

Discriminator D

Compares real and generated examples. It learns to tell which source produced each input.

$$\min_G \max_D \quad \mathbb{E}_{x\sim p_{\text{data}}}[\log D(x)] + \mathbb{E}_{z\sim p_z}[\log(1 - D(G(z)))]$$

Training alternates between the two networks. ALOCC uses a reconstructor in the generator's role.

Goodfellow et al., Generative Adversarial Nets, NeurIPS 2014.

ALOCC: learning from normal examples

R refines a noisy normal image

$$\widetilde{x} = x + \eta, \qquad x' = R(\widetilde{x})$$

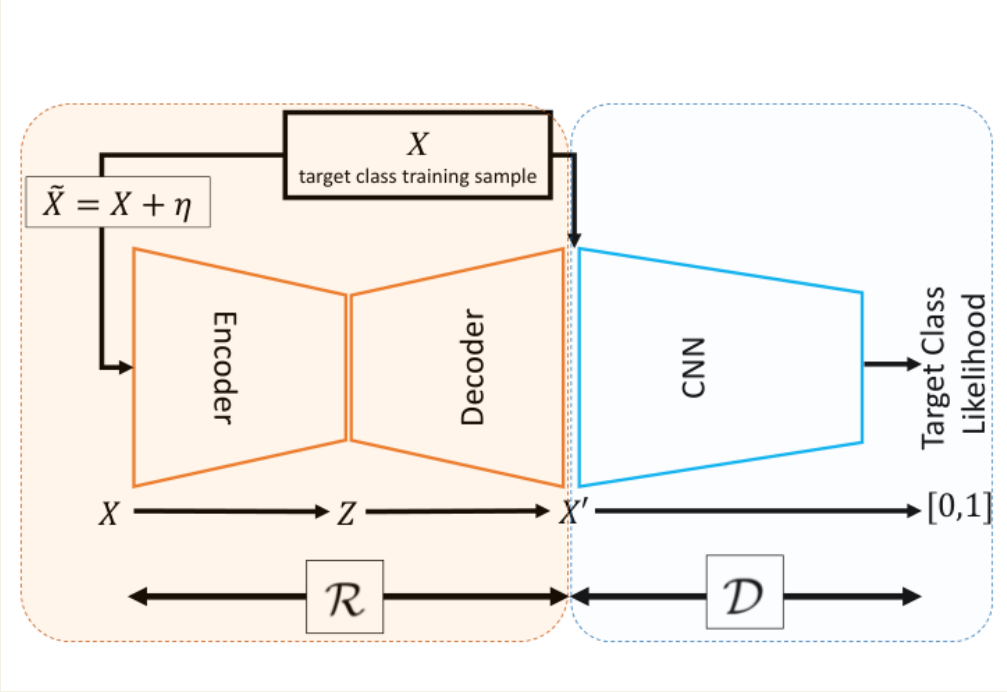
D learns from the contrast

Compare original normal images with R's reconstructions during training.

At test time, accept as normal when

$$D(R(x)) \geq \tau$$

$$\min_R \max_D \quad \mathbb{E}_x[\log D(x)] + \mathbb{E}_{\widetilde{x}}[\log(1 - D(R(\widetilde{x})))] + \lambda\, \mathbb{E}_{x,\eta}[\|x - R(\widetilde{x})\|_2]$$



Sabokrou et al., CVPR 2018, Fig. 2 and Eqs. 3–7. Adversarial objective plus reconstruction regularization.

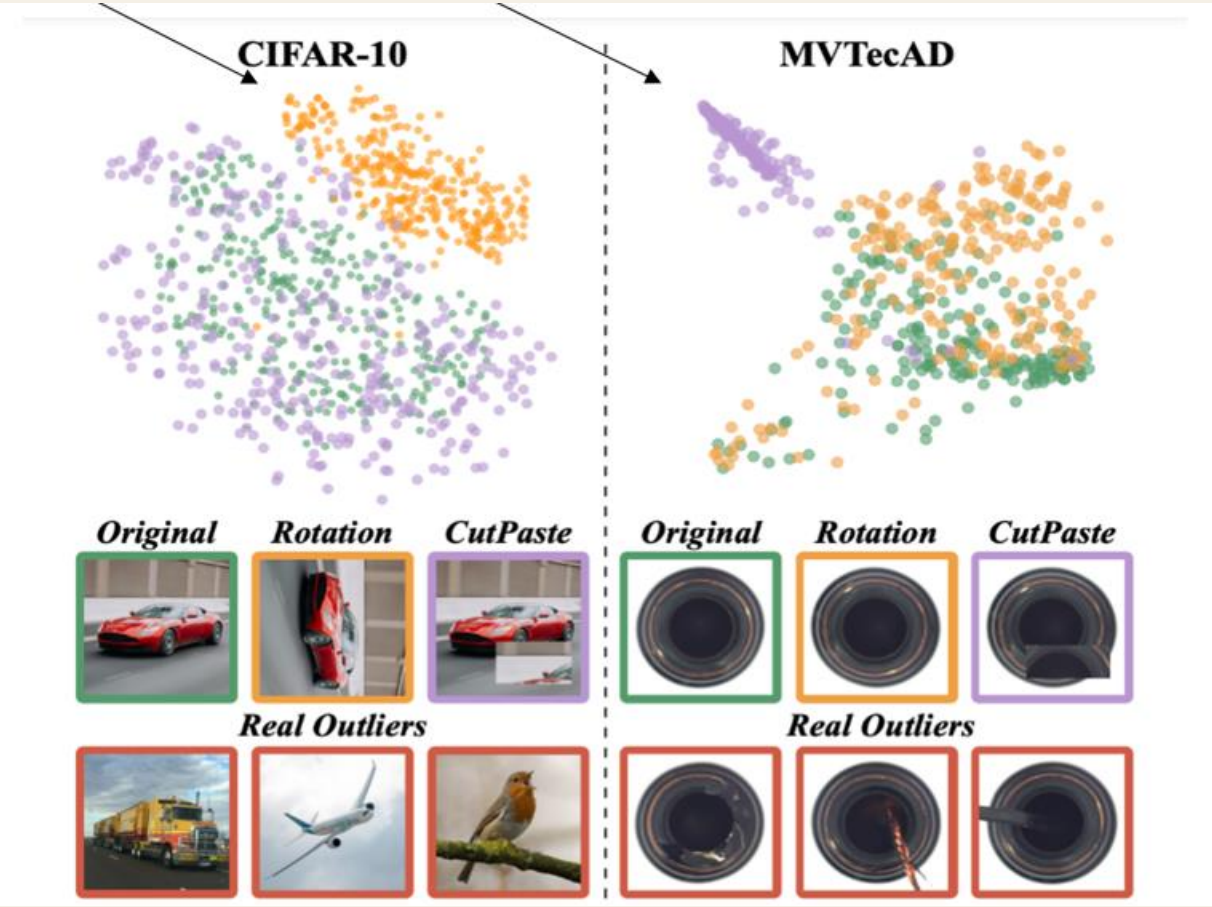
02

Generalizing to new domains

A detector tuned on one dataset can fail on the next.

FEATURED · UNODE, CVPR 2024 · CRANE, BMVC 2026

Which synthetic negatives fit the data?



A useful negative depends on the domain

A rotation changes object appearance, but may leave a symmetric industrial object looking normal.

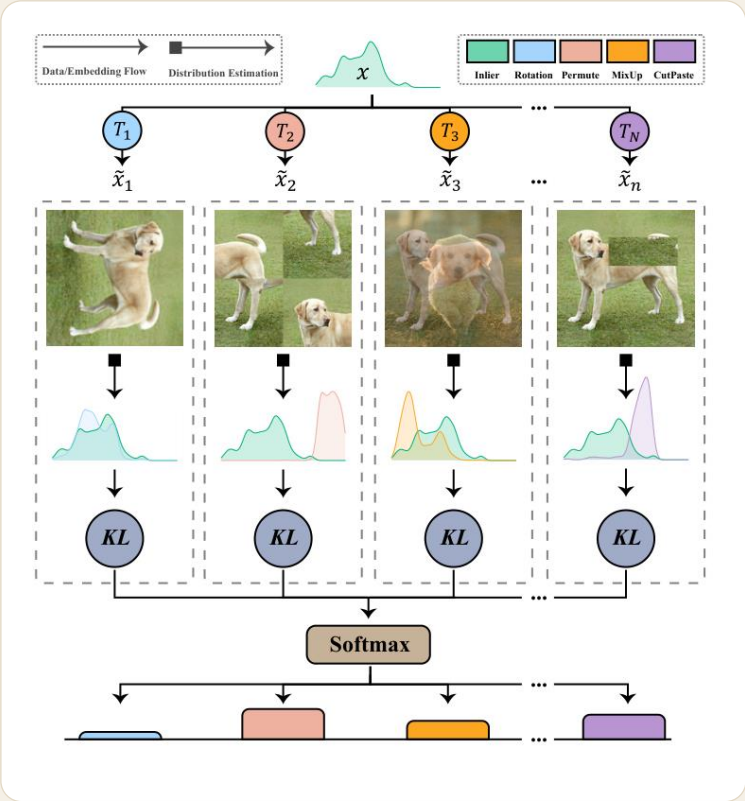
The shortcut risk

The model may detect the artificial transformation instead of the anomaly that matters.

UNODE adapts the negative transformations to the dataset.

Mirzaei et al., Universal Novelty Detection Through Adaptive Contrastive Learning, CVPR 2024.

UNODE: negatives that adapt to the dataset



1 · Candidate negatives

Hard transformations produce possible negative examples from the normal data.

2 · Adaptive selection

AutoAugOOD uses feature-distribution differences to choose transformations for each dataset.

3 · Contrastive loss plus classification

$$\ell_{\text{con}}(x) = -\frac{1}{|\mathcal{P}_x|} \log \frac{\sum_{x_i \in \mathcal{P}_x} \exp(\langle \hat{z}(x), \hat{z}(x_i) \rangle)}{\sum_{x_j \in \mathcal{P}_x \cup \mathcal{N}_x} \exp(\langle \hat{z}(x), \hat{z}(x_j) \rangle)}$$

P: positive views. N: adaptive negatives.
A binary head also separates inliers from negatives.

Mirzaei et al., Universal Novelty Detection Through Adaptive Contrastive Learning, CVPR 2024, Fig. 2

GENERALIZING TO NEW DOMAINS · RESULTS

UNODE: one recipe across ten datasets

92.5

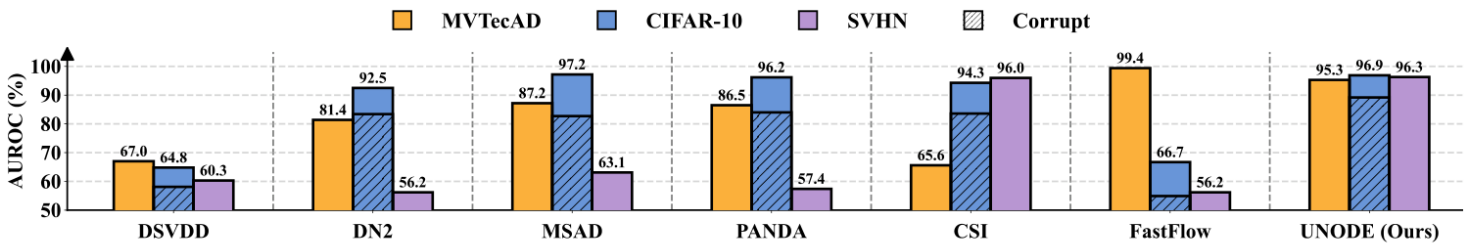
Mean AUROC over 10 datasets, trained from scratch (best prior: 82.3)

5.8

Spread across datasets (STD); other from-scratch methods: 13.6–14.9

92.2

Mean AUROC with pretrained features (best prior: 85.8)

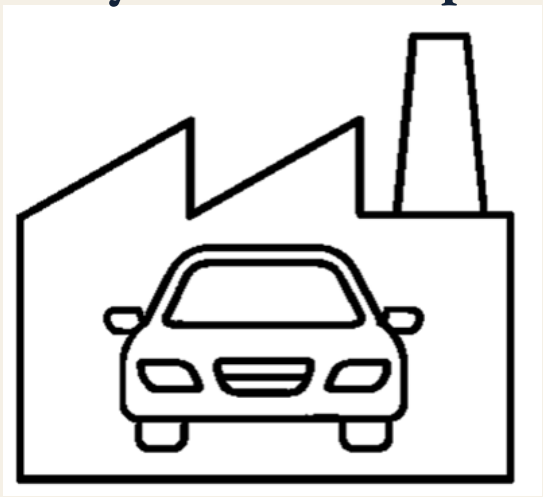


AUROC as reported on MVTec AD, CIFAR-10 and SVHN; hatched bars show corrupted test data

Mirzaei et al., CVPR 2024 · CIFAR-10/100, MNIST, EMNIST, Fashion-MNIST, SVHN, MVTec AD, Head CT, FGVC, ISIC 2018

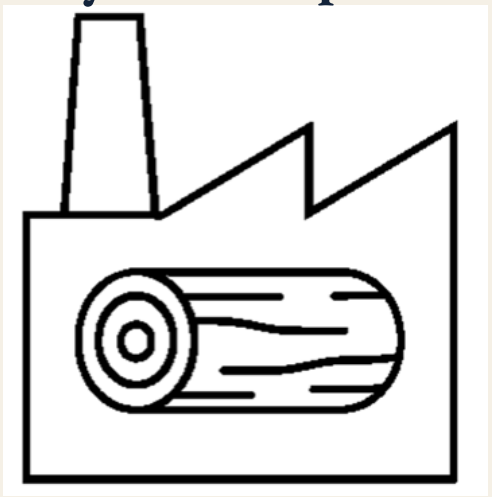
A detector moves to a new factory

Factory A · familiar production



Training and validation match this factory's products and cameras.

Factory B · new production



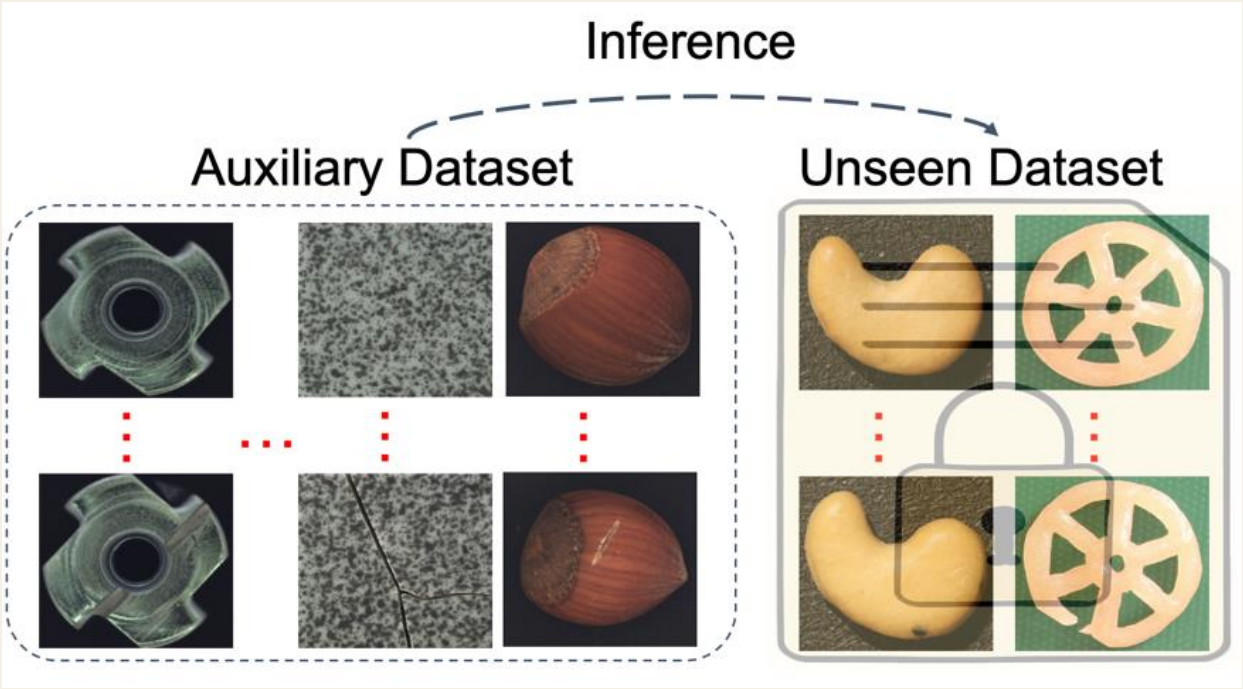
New products, lighting or defect types can break that match.

Zero-shot goal: inspect the new domain without collecting target-domain training images.

Conceptual transfer example. Performance after transfer must be measured; no numerical result is implied.

Zero-shot anomaly detection

No training images from the target domain



Auxiliary data

Learn generic normal and abnormal cues from source datasets.

An unseen target

Compare image features with normal and damaged text prompts, then localize suspicious regions.

The remaining gap: global image–text alignment does not directly train local region–word alignment.

CLIP as an anomaly detector



Input image

Text prompts define the comparison

“A photo of a normal hazelnut”

“A photo of a damaged hazelnut”

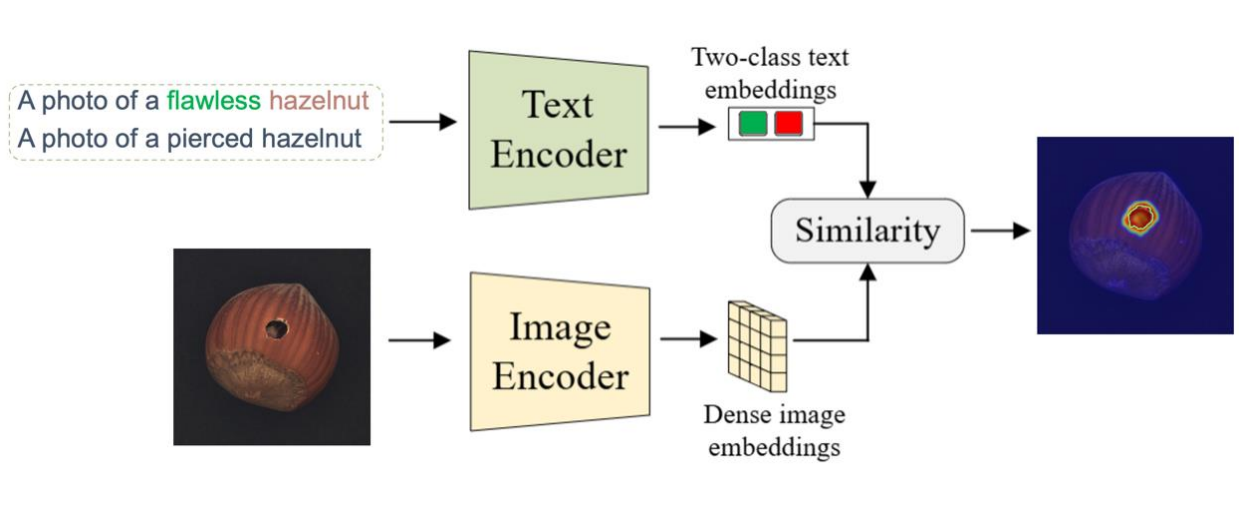
CLIP maps images and text into a shared feature space.
Compare the image with both text embeddings.

$$s(x) = \mathrm{sim}(f_I(x), t_{\mathrm{abn}}) - \mathrm{sim}(f_I(x), t_{\mathrm{norm}})$$

A higher abnormal-versus-normal similarity is a candidate anomaly score.

CLIP: Radford et al., ICML 2021. This is an illustrative CLIP-based scoring rule, before specialized adaptation.

Localization requires patch-level evidence



$$M_{uv} = \mathrm{sim}(z_{uv}, t_{\mathrm{abn}}) - \mathrm{sim}(z_{uv}, t_{\mathrm{norm}})$$

Where is the defect?

Compare each image patch with the normal and abnormal text embeddings to build a spatial score map.

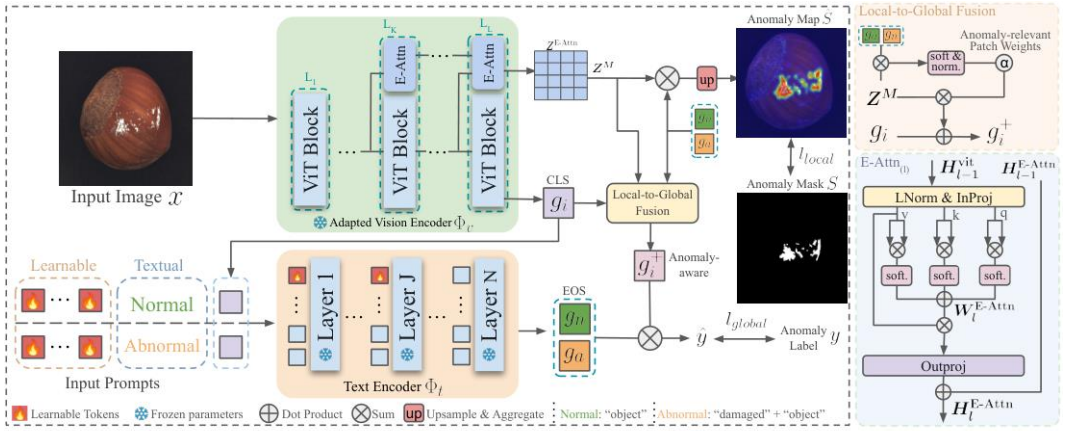
Why plain CLIP can struggle

Image-level caption training does not directly teach fine defect boundaries.

CRANE develops context-guided prompts, local attention and patch-to-global fusion.

Illustration: supplied technical lecture, slide 29. CLIP, AnomalyCLIP and CRANE motivate the local-alignment problem.

CRANE: three changes to a frozen CLIP



Backbone weights stay frozen. Prompt and deep text tokens are learned.

1 · Context-guided prompts

During training, learned normal/abnormal prompts receive the image's global [CLS] context.

2 · E-Attn in the last layers

Self-correlation weights retain local structure:

$$A(E) = \mathrm{softmax}\left(\frac{EE^{\mathsf{T}}}{\sqrt{d}}\right)$$

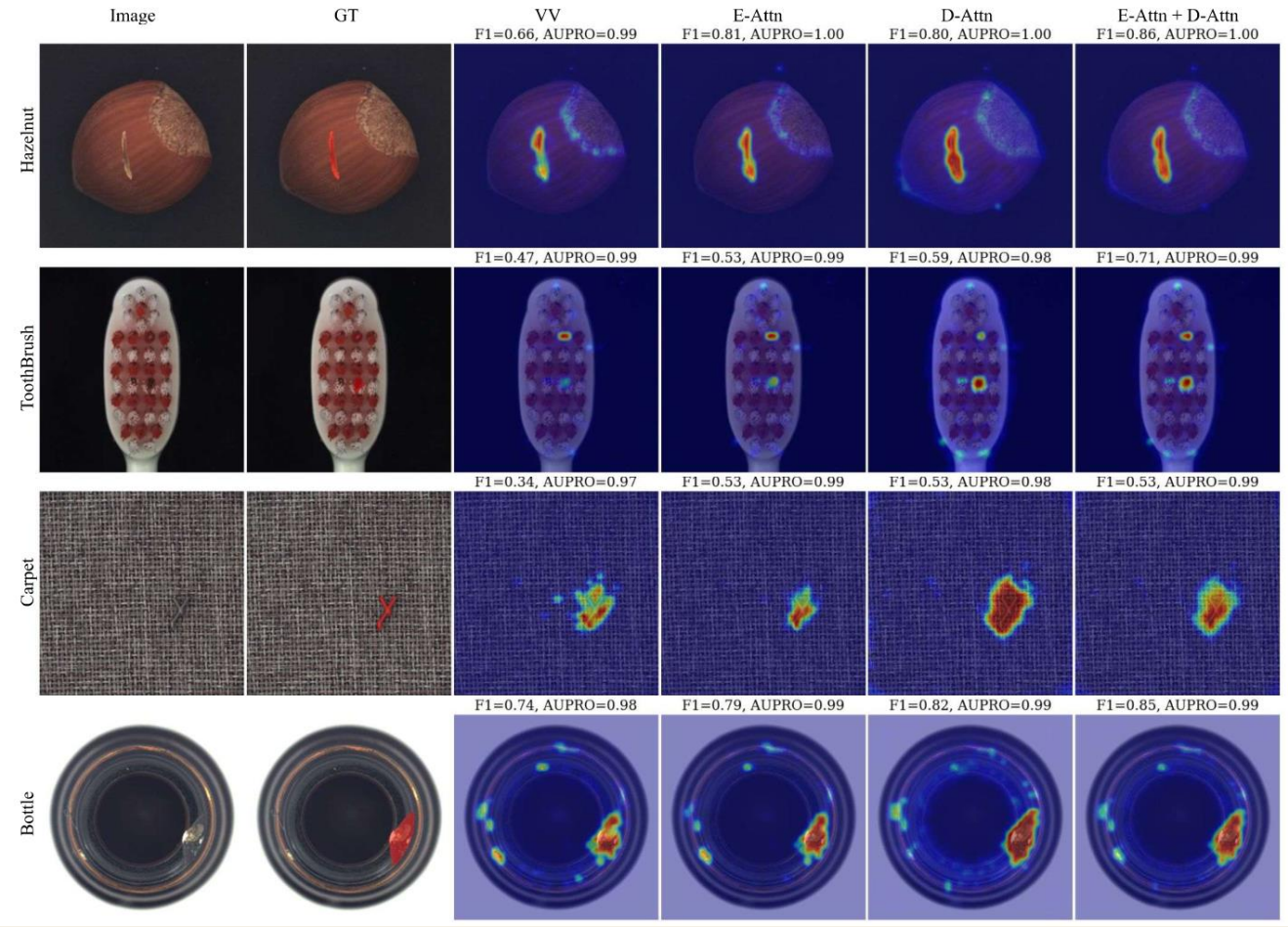
$$W = A(Q) + A(K) + A(V), \quad H = WV$$

3 · Local-to-global fusion

Defect-relevant patches feed the image embedding, so small defects count.

Salehi et al., CRANE, BMVC 2026. arXiv:2504.11055v3, §§3–4. CRANE+ also uses DINOv2.

CRANE: complementary attention branches



Salehi et al., CRANE v3, Fig. 4. Selected examples illustrate behavior. Aggregate benchmark results follow.

E-Attn: semantic structure

Self-correlations in CLIP help select anomaly-relevant regions.

D-Attn: spatial continuity

DINOv2 patch similarities add spatial detail, but can also activate normal regions.

The combined branches balance semantic selectivity and spatial detail.

CRANE: gains and remaining limitations

Setting and variant	Δ AUROC	Δ AP / AUPRO	Δ F1-max
Industrial image · CRANE	+2.4	+4.5	+0.7
Medical image · CRANE	+1.0	+3.1	+1.5
Industrial pixel · CRANE+	+0.5	+9.0	+3.8
Medical pixel · CRANE+	−0.6	+5.0	−1.6

Gain in percentage points over the strongest compared baseline for each metric.
AP measures image detection. AUPRO measures pixel localization.

Industrial localization improves strongly. Medical localization still has trade-offs.

Salehi et al., BMVC 2026, Table 1 (v3). Seven industrial datasets. Medical: 3 image-level and 4 pixel-level datasets.

03

Staying robust and secure

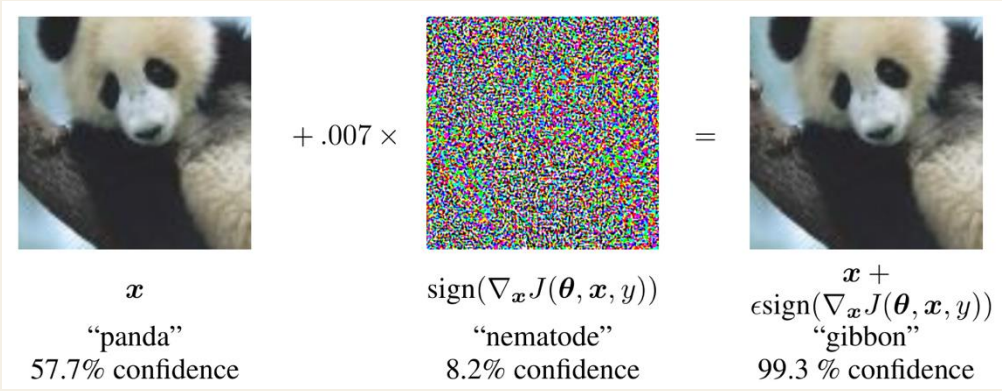
A detector that works on clean data can fail under manipulation.

FEATURED · CBS, TMLR 2024 · TRODO, NEURIPS 2024

Adversarial examples and hidden backdoors

Adversarial examples

Change a test input so the model changes its decision.



$$x_{\text{adv}} = x + \delta, \qquad \|\delta\|_\infty \leq \varepsilon$$

$$f_\theta(x_{\text{adv}}) \neq y$$

Backdoors

Poisoning during training can teach a hidden trigger–target association.

CLEAN INPUT

STOP → Same model → **Stop sign**

INPUT WITH TRIGGER

STOP → Same model → **Speed limit 30**

High clean accuracy can conceal the backdoor.

Adversarial image: Goodfellow et al., ICLR 2015, Fig. 1. Backdoor diagram: illustrative, after BadNets (2017).

Data poisoning and backdoor checks

A backdoor poison changes some training examples to link a hidden trigger with a target prediction.

$$\mathcal{D}_{\text{train}} = \mathcal{D}_{\text{clean}} \cup \mathcal{D}_{\text{poison}}$$

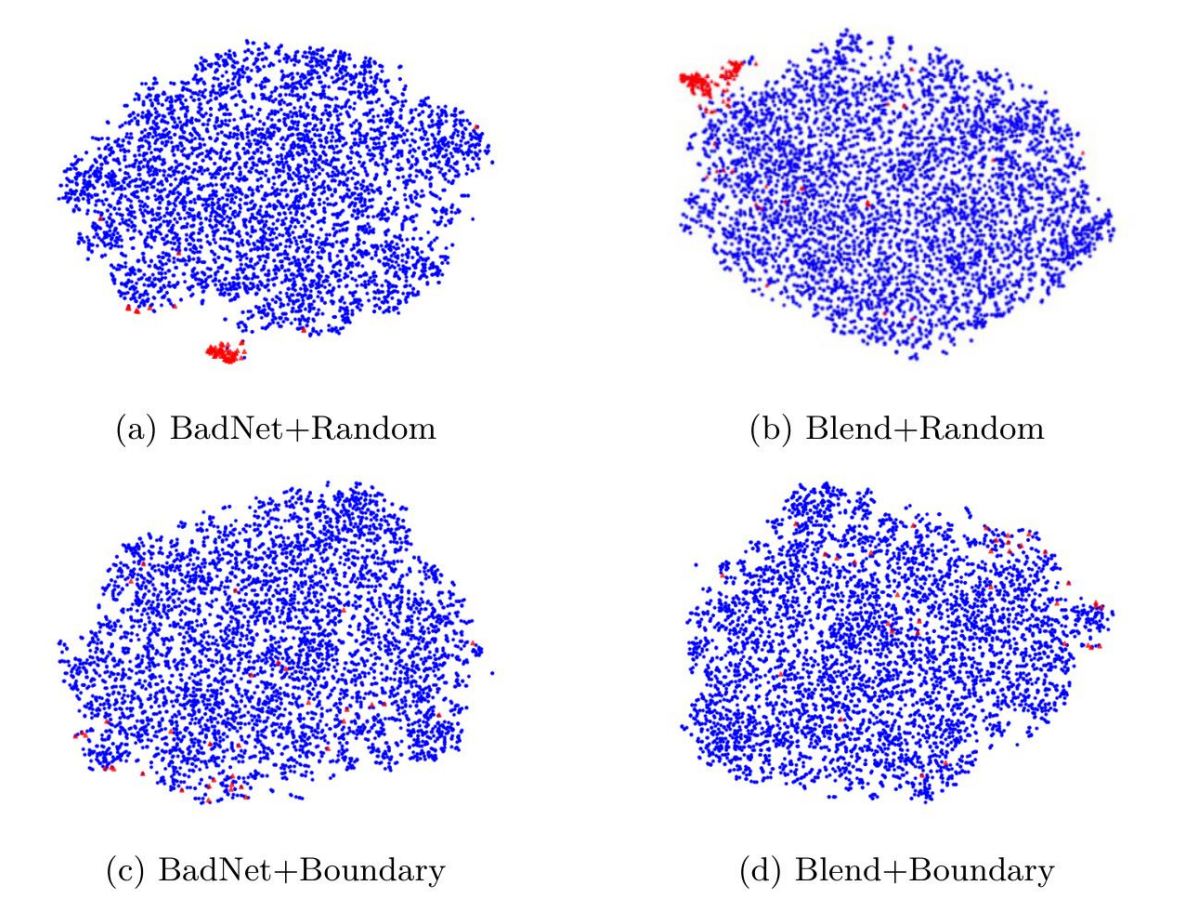
Most examples still teach the normal task. The poisoned subset teaches an additional behavior that appears when the trigger is present.

Access	What can we check?
Training data	Audit suspicious examples, labels and feature clusters.
Trained model	Probe unusual responses and compare a scanning score with clean-model references.

A check can flag risk; its detection and false-alarm rates still require validation.

CBS: confidence-driven boundary sampling

Blue: clean target-class samples Red: poisoned samples



Which samples receive the trigger?

Use a surrogate model to identify examples near the true-class / target-class boundary.

$$|p_\theta(y \mid x) - p_\theta(y_t \mid x)| \leq \varepsilon$$

A small probability gap serves as a boundary proxy. The trigger can stay unchanged.

Boundary sampling makes poisoned features harder to separate.

He et al., Stealthy Backdoor Attack via Confidence-driven Sampling, TMLR 2024, Definition 4.1 and Fig. 1.

Model auditing without the training set

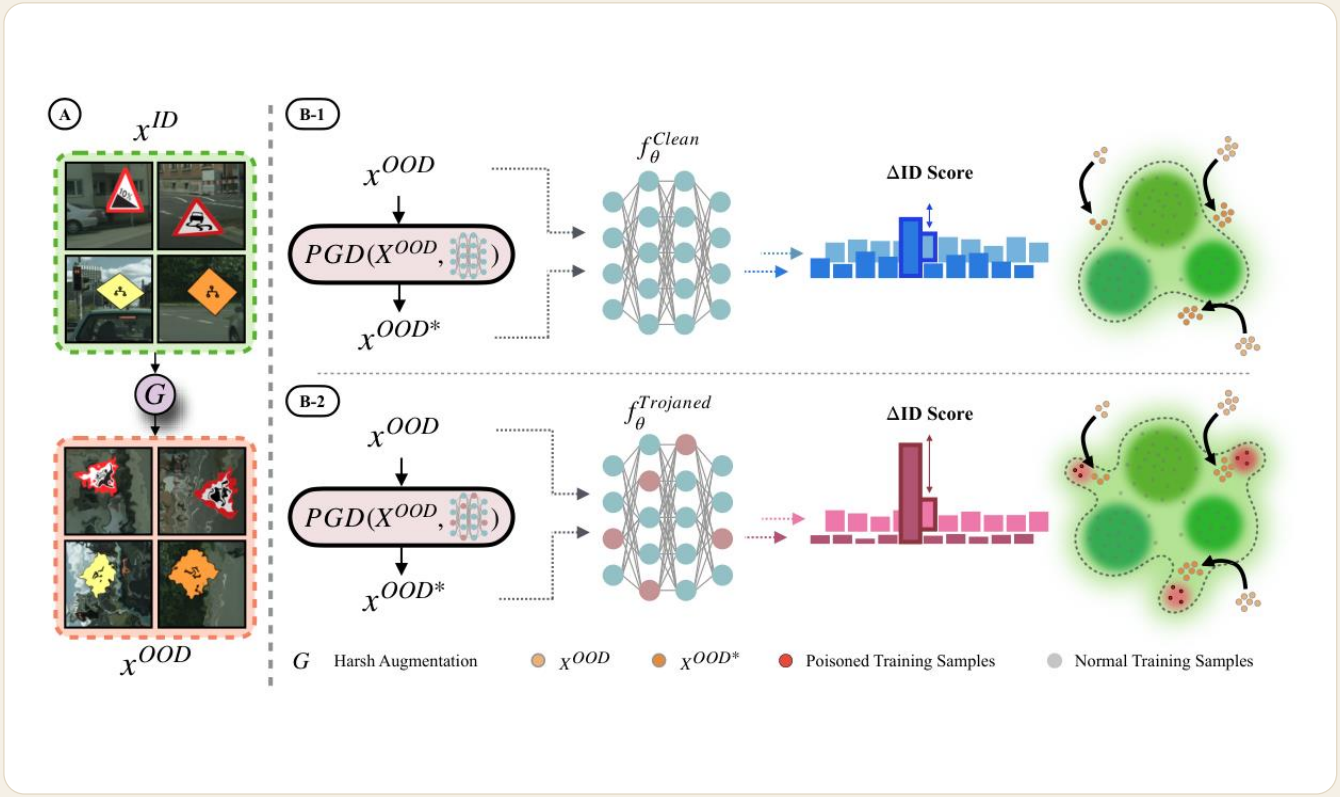
We may receive a pretrained model without the data or history that produced it.

Available to the auditor	Unavailable to the auditor
Model weights and execution access	Original training examples
An external reference set for probes	A list of poisons or the hidden trigger

TRODO-Zero tests how the model responds to external probes and bounded input changes.

"No training data" still permits external reference data and calibration. TRODO-Zero requires model gradients.

TRODO: scanning models for hidden trojans



Probe with OOD inputs

Use augmented clean images or an external reference set.

$$c(x) = \max_k p_\theta(y = k \mid x)$$

Increase confidence locally

PGD searches a bounded neighborhood for a higher ID score.

$$S(f) = \frac{1}{N} \sum_{i=1}^{N} [c(x_i^*) - c(x_i)]$$

Flag if $S(f) > \tau$

Mirzaei et al., Scanning Trojaned Models Using Out-of-Distribution Samples, NeurIPS 2024, Fig. 1

TRODO: scanning across attack settings

Model-scanning accuracy (%), averaged across five datasets in the paper's benchmark.

Scanner	All-to-one standard	All-to-one adv. trained	All-to-all standard	All-to-all adv. trained
FreeEagle	76.3	72.3	58.9	55.4
MM-BD	79.3	62.1	48.7	37.2
UMD	75.4	58.0	76.1	57.7
TRODO-Zero	79.4	77.0	79.2	78.8
TRODO	90.7	87.3	87.5	86.1

The scanning signature remains useful when the evaluated models use adversarial training.

Mirzaei et al., NeurIPS 2024, Table 1. Selected scanners. TRODO-Zero needs no training images, but uses external data.

RESEARCH SYNTHESIS

How the three themes connect

01

Recognizing the unknown

Flag inputs outside what the model knows.

Survey · ALOCC

02

Generalizing to new domains

Keep that judgment reliable in new domains.

UNODE · CRANE

03

Staying robust and secure

Hold up under attacks and hidden faults.

CBS · TRODO

SHARED INGREDIENTS

- Representation geometry
- Contrastive relations
- Evidence at the right scale

TRODO closes the loop: out-of-distribution behavior, the subject of Theme 1, becomes evidence about the model itself.

ONGOING RESEARCH · BASIC IDEA

LLMs, hallucination and activation steering

LLM

A language model predicts the next token from the context. Repeating this process produces an answer.

Hallucination

A fluent answer contains an incorrect or unsupported claim, such as an invented reference.

Steering

Change an internal activation during generation to influence the model's output.

$$h' = h + \alpha v$$

h: hidden state
v: steering direction
α: intervention strength

Activation steering changes behavior. Any factuality benefit must be measured, including damage to correct answers.

ONGOING PROJECT · RESEARCH HYPOTHESIS

Dynamic activation correction for LLMs

Can a small, gated intervention reduce hallucination while preserving correct answers?

Reference model and correction

Fit a reference distribution to hidden states from verified answers.

$$D(h) = (h - \mu)^\top \Sigma^{-1} (h - \mu)$$

$$\delta^* = \arg\min_{\delta} \left[D(h + \delta) + \lambda \|\delta\|_2^2 \right]$$

$$h' = \begin{cases} h + \delta^*, & D(h) > \tau, \\ h, & \text{otherwise.} \end{cases}$$

Evaluation plan

Compare a frozen model, fixed steering, and gated correction.

Measure factuality, changes to correct answers, intervention rate and latency.

Key test: does hidden-state distance predict factual errors on unseen topics?

Research in progress. Hypothesis, candidate objective and evaluation plan.

ONGOING PROJECTS · OPEN RESEARCH QUESTIONS

Privacy, backdoors and safety in large models

Current directions and the evidence each project needs

Direction	Research question	Planned evaluation
Privacy and memorization	When does a model reveal sensitive training knowledge?	Leakage and extraction rates, with retained task performance
Large-model backdoors	Can hidden triggers survive pretraining and transfer?	Attack success after transfer, clean accuracy and detectability
Safety signals and continual learning	Do watermark and safety signals survive model updates?	Detection errors, false positives and retention after updates

Each direction needs a defined failure case, a reproducible test and a useful baseline.

Research in progress. Evaluation criteria describe planned experiments.

What AI alignment asks

How can an AI system pursue intended human goals while respecting agreed values and constraints?

Example: an AI tutor

"Help me understand this proof."

A convincing-looking answer can receive approval even when a step is false.

The intended behavior

Support the student's understanding.

Check the reasoning and admit uncertainty when evidence is missing.

Alignment links the human goal to feedback, training and evaluation.

Gabriel, Minds and Machines 2020. Ouyang et al., NeurIPS 2022. Tutor scenario is illustrative.

Why philosophy matters for AI

Philosophical question	Consequence for an LLM
Knowledge **What justifies a claim?**	Require evidence and distinguish a supported answer from a plausible guess.
Values **Whose preferences count?**	Include affected people, document disagreements and evaluate across languages and contexts.
Responsibility **Who can question a decision?**	Make limitations visible and preserve meaningful human review.

Philosophy helps define what to optimize and what evidence should count as success.

Research framing informed by Gabriel, Artificial Intelligence, Values and Alignment, Minds and Machines 2020.

TAKE-HOME MESSAGES

What reliable AI needs

01	**Unknown inputs**	A model needs a tested way to signal unfamiliarity.
02	**New conditions**	Generalization depends on what changes and what should remain stable.
03	**Manipulation**	Security evaluation needs an explicit threat model.
04	**Real decisions**	Thresholds and human review turn scores into operational choices.

Which failure would make an AI system unusable in your own work?

Thank you

Rahmat! سپاسگزارم

Questions and collaboration

Mohammad Sabokrou

m.sabokrou@newuu.uz

sabokrou.github.io

New Uzbekistan University
Okinawa Institute of Science and Technology, Japan

Papers and research