AI-BOOST

Delivering the next level of European **AI Open competitions**

GENERATIVE AI FOR AUTOMATIC TEST CASE GENERATION FROM CRASH DATABASES & STANDARDS

CHALLENGE DESCRIPTION

GENERATIVE AI FOR AUTOMATIC TEST CASE GENERATION FROM CRASH DATABASES & STANDARDS

1. Context and Objectives

Organisation description

Siemens Industry Software NV (SISW) is an engineering and technology partner serving companies in the automotive, aerospace, and advanced manufacturing sectors. Through its Simcenter portfolio and Engineering teams, SISW supports customers in test-based engineering, 3D simulation, and model-based systems engineering, drawing on extensive industrial experience. SISW also develops solutions in areas such as system dynamics, industrial AI, control engineering, and design optimization. As part of its strategic R&D activities, SISW is expanding its portfolio of AI-driven engineering solutions to support next generation simulation, validation, and digital engineering workflows. In this context, SISW has direct experience with the challenges of applying AI in engineering environments, including fragmented information sources, limited access to high quality data, and the need for trustworthy and domain-specific AI methods.

General context

Recent years have seen increasing regulatory activity and deployment efforts for automated driving across Europe. As automated driving moves closer to broader adoption, the need for transparent, scalable, and evidence-based validation methods is becoming increasingly urgent.

European traffic environments exhibit distinctive characteristics, including dense urban areas, diverse climatic conditions, extensive cycling infrastructure, complex road layouts, and significant variation across national road networks and traffic conditions. These characteristics shape the safety-critical situations that should be represented in validation. As a result, methods and scenario resources developed primarily from non-European data may fail to capture important European conflict patterns and validation priorities.

Simulation-based validation, risk assessment, and automated driving regulation depend on the availability of realistic and safety-relevant scenarios. Across Europe, however, the evidence needed to derive such scenarios remains fragmented across accident databases, crash records, investigation reports, infrastructure data, digital maps, images, and videos. Although highly valuable for safety engineering, this information is rarely available in a structured form that can directly support scenario extraction, scenario catalog development, or validation coverage assessment.

Recent advances in generative AI, multimodal foundation models, retrieval-augmented generation (RAG), knowledge graphs, agentic AI workflows, and explainable AI create new opportunities to address this challenge. These technologies have the potential to transform fragmented safety evidence into structured scenario knowledge, support the discovery of recurring traffic conflicts, enrich sparse accident descriptions with contextual information, and assist engineers in identifying which safety-relevant situations are already reflected in validation scenario resources and which remain insufficiently covered.

This challenge proposes a benchmark for generative AI and multimodal AI methods that transform fragmented European safety evidence into structured, explainable, and validation-relevant scenario knowledge for automated driving. Participants are asked to extract Functional Scenarios from accident data and enrich them into Logical Scenarios using multimodal evidence. In the final validation step, AI methods are evaluated on how accurately they can match reference validation scenario descriptions, for example, a Euro NCAP-aligned subset, to the accident-derived scenario catalog generated by the participants. These match results are then used to support validation coverage analysis and gap identification.

The central benchmark artifact is a Structured Scenario Representation that combines scenario abstractions, enriched attributes, supporting evidence, confidence information, and validation relevance metadata. By bridging real-world accident evidence and scenario-based validation workflows, the benchmark addresses a critical gap not covered by existing automated driving benchmarks, which primarily focus on perception, prediction, planning, or driving policy performance.

The benchmark is intended to support researchers, practitioners, and other stakeholders working at the intersection of AI, safety engineering, and automated driving validation. Submissions will be evaluated not only on extraction accuracy and scenario completeness, but also on explainability, evidence traceability, and usefulness for scenario matching and validation coverage assessment. In this way, the challenge aims to stimulate methods that are technically strong while also being relevant for real-world safety validation workflows in Europe.

SCOPE

The main objective of the challenge is to develop AI-driven methods capable of transforming heterogeneous accident and safety evidence into structured scenario representations that can support simulation-based validation of automated driving systems.

Participants are invited to design solutions capable of:

- extracting relevant information from accident datasets, reports, and related evidence;
- identifying and harmonizing recurring **Functional Scenarios** across heterogeneous sources;
- enriching these scenarios into Logical Scenarios using multimodal evidence such as text, images, maps, or infrastructure context;
- representing the results in a common Structured Scenario Representation; and
- matching reference validation scenario descriptions to the participant-generated accident-derived scenario catalog in order to support validation coverage analysis and identify gaps.

The challenge is intentionally technology-agnostic. Participants may employ generative AI, Large Language Models (LLMs), Vision-Language Models (VLMs), Retrieval-Augmented Generation (RAG), knowledge graphs, agentic AI workflows, multimodal foundation models, or hybrid approaches. Participants may address individual benchmark components or develop integrated end-to-end solutions spanning the full workflow. While modular participation is permitted, integrated solutions spanning all three components will be considered particularly favorably in the overall assessment. Solutions that make use of open, shareable, and reproducible tools, or models will be viewed positively, particularly where this improves transparency, portability, and reuse.

The long-term vision is to enable AI systems capable of converting fragmented safety evidence into structured, traceable, and validation relevant scenario knowledge, thereby reducing manual engineering effort and improving scenario coverage in automated driving validation workflows.

The challenge is organized into three interconnected benchmark components:

- **Component 1 – Functional Scenario Extraction and Harmonization:** identify and cluster recurring scenarios from heterogeneous accident datasets and map them to a common taxonomy.
- **Component 2 – Logical Scenario Enrichment from Multimodal Evidence:** enrich Functional Scenarios with structured attributes and parameter ranges derived from multimodal evidence.
- **Component 3 – Scenario Matching and Validation Coverage Assessment:** match reference validation scenario descriptions - for example, a Euro NCAP-aligned subset - to the participant-generated accident-derived scenario catalog in order to assess validation coverage and identify potential gaps.

These components support both modular participation and integrated end-to-end benchmarking. Solutions that build on the outputs of earlier components to improve the completeness, consistency, and validation relevance of downstream results will be given greater weight in the overall evaluation.

2. DATASETS PROVIDED

The benchmark may draw on a combination of structured, semi-structured, and multimodal data sources, including accident databases, crash reports, textual descriptions, images, videos, maps, infrastructure information, and validation-oriented scenario resources.

Representative data sources may include:

1. **Component 1 – Functional Scenario Extraction and Harmonization:** European accident databases, national crash statistics, and related open road safety data sources. Participants may also draw on comparable open datasets from other countries or regions. Representative examples include:
    - UK Road Safety Open Data: https://www.gov.uk/government/statistical-data-sets/road-safety-open-data
    - France Accidentologie: https://www.data.gouv.fr/fr/datasets/accidentologie-base-victimes/
    - Belgium road traffic accidents: https://statbel.fgov.be/en/open-data/geolocation-road-traffic-accidents-2017-2024 and https://statbel.fgov.be/en/open-data
    - Other open data released by European administrations (for instance, data by the city of Barcelona https://opendata-ajuntament.barcelona.cat/data/es/dataset?q=accidents)

2. **Component 2 – Logical Scenario Enrichment from Multimodal Evidence:** multimodal datasets and related scenario representation resources providing environmental and contextual information such as images, videos, textual crash descriptions, maps, road

geometry, infrastructure metadata, weather, other environmental context, and formal scenario models. Representative examples include:

- KITScenes-LongTail: https://huggingface.co/datasets/KIT-MRT/KITScenes-LongTail
- KITScenes multimodal data: https://huggingface.co/datasets/KIT-MRT/KITScenes-Multimodal/tree/main/data
- CycleCrash: https://github.com/DeSinister/CycleCrash
- DeepAccident: https://deepaccident.github.io/data.html
- CYCLANDS: https://github.com/U-Shift/cyclands
- PREPER: https://github.com/AsymptoticAI/PREPER
- US NHTSA Crash Investigation Sampling System: https://www.nhtsa.gov/crash-data-systems/crash-investigation-sampling-system
- ASAM OpenSCENARIO (v1 or v2) and related examples:
  https://www.asam.net/standards/detail/openscenario/v200/
  https://github.com/ika-rwth-aachen/alks-scenarios
  https://github.com/vectorgrp/OSC-NCAP-scenarios

3. **Component 3 – Scenario Matching and Validation Coverage Assessment:** reference scenario resources and validation-oriented scenario descriptions. Participants may draw on Euro NCAP scenario descriptions and protocol documents, as well as public validation scenario catalogs. Datasets such as GIDAS, IGLAD, and RAIDS may also be referenced where access is available; however, these are not required for participation. Representative examples include:

- GIDAS: https://www.gidas.org/about-en.html
- IGLAD: https://www.iglad.net
- RAIDS: https://www.raidsuk.org
- Euro NCAP: https://www.euroncap.com/crash-avoidance/

3. KEY METRICS

Evaluation metrics may be specified independently for each benchmark component, reflecting the distinct objectives of functional scenario extraction, logical scenario enrichment, and validation coverage assessment. Accordingly, evaluation should consider both task-specific performance and broader qualities such as explainability, consistency, scalability, and usefulness for downstream validation workflows.

Primary metric dimensions may include:

- quality of functional scenario extraction and taxonomy alignment,
- completeness and correctness of logical scenario enrichment,
- robustness of evidence grounding and traceability,
- accuracy of scenario-to-catalog matching,

- usefulness of identified validation coverage gaps.

Secondary metric dimensions may include:

- explainability of the extraction and enrichment process,
- consistency across heterogeneous data sources,
- diversity and representativeness of generated scenarios,
- scalability to larger accident datasets,
- suitability of outputs for downstream validation preparation,
- visualization of scenarios.

The benchmark focuses on evidence-grounded scenario knowledge generation and validation coverage reasoning. It is not intended to directly assess legal compliance, homologation, or certification readiness.

Compliance and Ethical Requirements

It is recommended that participants consider the EU AI Act, particularly regarding transparency, human oversight, robustness, and traceability requirements for AI systems used in safety-critical automotive contexts. Any use of third-party AI services or large language models within the pipeline must be explicitly disclosed. Participants must ensure that no confidential, sensitive, or personally identifiable information (PII) is transmitted to external services during processing.

- EU AI Act requirements (transparency, traceability, human oversight, robustness)
- Data protection regulations (GDPR compliance)
- Strict prohibition of transmitting PII or sensitive accident data to external services unless explicitly anonymized
- Mandatory disclosure of external AI models, APIs, or cloud services used

Solutions must NOT:

- Generate physically impossible motion or trajectories
- Violate basic traffic physics or kinematic constraints
- Produce non-reproducible simulation outputs
- Hallucinate missing data without uncertainty representation
- Leak or infer personally identifiable information
- Produce inconsistent multi-agent interactions

4. Objectives

The challenge aims to benchmark how generative AI and multimodal AI methods can transform fragmented accident and safety evidence into structured, explainable, and validation-relevant scenario knowledge for automated driving.

Specific objectives are to:

- automatically extract recurring Functional Scenarios from heterogeneous accident datasets;
- harmonize scenario descriptions across datasets and countries;
- enrich sparse accident-derived scenarios into Logical Scenarios using multimodal evidence;
- package outputs into a common Structured Scenario Representation;
- match accident-derived scenarios to a validation scenarios catalog;
- identify meaningful coverage gaps and prioritization insights for simulation-based validation.

More broadly, the challenge seeks to establish a benchmark that links real-world accident evidence with scenario-based validation practice, thereby supporting more scalable, transparent, and evidence-driven safety engineering workflows.

5. Expected Outcomes and TRL Level

The challenge aims to deliver proof-of-concept (PoC) solutions that demonstrate the feasibility of using generative AI and multimodal AI to transform accident and safety evidence into structured scenario knowledge for validation-oriented use.

Expected outcomes include:

1. AI methods capable of extracting functional scenarios from heterogeneous accident datasets.
2. AI methods capable of enriching those scenarios into logical scenarios using linked multimodal databases, including handling of uncertainty where evidence is incomplete, ambiguous, or conflicting.
3. Structured Scenario Representations that provide traceable and comparable benchmark outputs, including uncertainty-related information where relevant.
4. Methods for matching accident-derived scenarios against a validation scenario catalog and identifying coverage gaps.
5. Where applicable, methods for generating and visualizing simulation-oriented scenario descriptions, such as structured outputs compatible with OpenSCENARIO or equivalent scenario formalisms. This may also include the use of optimization or related techniques to derive parameterized concrete scenarios, as well as the use of simulation tools, either open-source (e.g., CARLA) or commercial (e.g., Simcenter Prescan), to support visualization, demonstration, and simulation model creation through relevant APIs.

Technology Readiness Level (TRL)

The expected outcome is to advance solutions from TRL 3 (proof of concept) to TRL 5 (validation in relevant environments), demonstrating both technical feasibility and integration potential in industrial workflows.

6. EXPECTED IMPACTS AND KPIS

Expected impacts: The proposed challenge has the potential to generate impact across research, engineering practice, public-sector decision-making, and the broader European AI and mobility ecosystem.

From a research perspective, the challenge advances generative AI for structured knowledge extraction, multimodal reasoning, evidence-grounded scenario abstraction, uncertainty-aware scenario enrichment, validation coverage reasoning, and explainable AI. It introduces a benchmark problem that remains underexplored in current AI ecosystems, yet is highly relevant to the development of trustworthy automated driving systems. In particular, it encourages methods that not only infer missing scenario information, but also represent the associated uncertainty in a transparent and usable way.

From an industrial perspective, the challenge supports scenario database creation, scenario prioritization, simulation, engineering traceability, and AI-assisted safety and validation workflows. Uncertainty-aware outputs can further improve industrial usefulness by helping engineers distinguish between well-supported scenario elements and more speculative inferences, thereby supporting more robust prioritization, review, and validation decisions.

From a public-sector and ecosystem perspective, the challenge can support regulators, policymakers, consumer assessment bodies, insurers, and standardization communities by enabling more transparent, traceable, and evidence-grounded links between real-world accident data and scenario-based validation practice. Explicit treatment of uncertainty can also help these stakeholders interpret the confidence and limitations of derived scenario knowledge, which is important for safety assessment, risk-informed decision-making, and policy discussions.

The challenge also contributes to the European AI ecosystem by creating benchmark assets, datasets, and evaluation methodologies derived from European safety evidence and validation practices. In doing so, it helps reduce dependence on non-European scenario resources and supports European technological sovereignty.

KPIs: Indicative key performance indicators include:

- **KPI 1. Functional scenario extraction quality:** quality of scenario classification, clustering, or taxonomy alignment.
- **KPI 2. Logical scenario enrichment quality:** completeness and correctness of inferred scenario attributes and parameter ranges.
- **KPI 3. Evidence grounding and traceability:** proportion of generated outputs supported by explicit source evidence references and explanations.
- **KPI 4. Validation coverage assessment accuracy:** correctness and usefulness of scenario-to-catalog matching and coverage estimation.
- **KPI 5. Scalability and reproducibility:** ability to process large or additional datasets in a reproducible and well-documented manner.

- **KPI 6. Uncertainty representation and calibration:** ability to quantify, communicate, and appropriately calibrate uncertainty in inferred scenario attributes, scenario enrichments, and matching outcomes.

7. Infrastructure and reproducibility

No specific hardware or HPC infrastructure is mandated. However, solutions should be designed to scale to large crash datasets without prohibitive computational cost. Participants must document all external dependencies, and the evaluation pipeline must be reproducible in a self-contained environment to ensure fair benchmarking.

Submissions should:

- scale to large or new country/region datasets;
- be reproducible in containerized environments;
- document all external dependencies (e.g., LLMs, APIs, and cloud services).

8. Support offered by the challenge owner

Participants will benefit from direct engagement with domain experts and access to a multidisciplinary innovation ecosystem throughout the challenge.

- Technical mentorship from RobustifAI experts in simulation, artificial intelligence, autonomous systems, scenario-based validation, accident-data interpretation, and automotive safety.
- Clarification of benchmark tasks, expected outputs, and evaluation criteria, together with feedback on the relevance of generated outputs for validation workflows.
- Access to a multidisciplinary EU consortium network, including industrial and research partners such as Siemens, AIT, University of Liverpool, TU Wien, Collins Aerospace, LOXO, PPM, and Thales.
- Feedback from industry experts to help assess the feasibility, scalability, and practical relevance of proposed solutions.
- Opportunities for follow-up discussion regarding co-development activities, pilot projects, research collaborations, or possible integration into future product and service portfolios.

Funded by the European Union

ANNEX 1. DOCUMENTATION AND INSTRUCTIONS DATASETS

Explanation/Documentation Required For Users To Gain Access

The challenge does not impose a single mandatory dataset access procedure for all participants. Instead, participants may rely on publicly available or otherwise lawfully accessible datasets relevant to the benchmark, provided that all applicable access conditions, license terms, and usage restrictions are respected.

Where datasets are openly available, participants may access them directly through the original providers. Where datasets require prior authorization, registration, or approval, participants are responsible for obtaining such access independently and for complying with all associated legal, ethical, and confidentiality obligations.

Background resources

The following non-mandatory resources may be useful as background inspiration for scenario extraction and scenario enrichment approaches:

1. From text to scenario:
https://www.linkedin.com/posts/tongduyson_automotive-safety-adas-activity-7370406583710687232-TT4E

2. Scenario enhancement with video data:
https://www.linkedin.com/posts/tongduyson_autonomousdriving-adas-generative-activity-7227069053188673536-LY_w