Instructions for Annotators

November 2025

1 Introduction

You are invited to participate in evaluating test-validator assertions for cyber-physical systems (CPS). Test validators are automatically learned classifiers that characterize input scenarios as safe or low risk, or as violations of system preconditions or operational design domain (ODD) limits. Your participation will help us assess how interpretable and practically valuable these validators are for identifying meaningful test outcomes without needing to run costly or flaky simulations.

Objective. The goal of this study is to assess how closely automatically generated test-validator assertions align with the descriptions in the reference documentation for several CPS:

- **A network router:** assertions that describe conditions on traffic classes and traffic load under which quality of experience are satisfied or violated.
- **Several variants of autonomous driving systems (ADS):** assertions that describe driving scenarios, environmental and traffic conditions, and vehicle behaviours that fall within or outside the intended ODD or violate preconditions.
- **An aircraft autopilot system:** assertions that describe relationships among flight attitude, engine thrust, and related signals that correspond to satisfaction or violation of climb requirements.

Your task is to judge how well each assertion aligns with a small set of related sentences extracted from the system's reference documentation. All information required for your judgments is contained in the assertion and the retrieved sentences. Please do not consult external sources (web, manuals, etc.); use only the text we provide.

2 Background

Each test-validator assertion describes a situation in natural language, such as:

- A violation of preconditions or operational design domain (ODD) limits
- A safe or low-risk operating scenario

For each annotation item you will be given:

- **System name.** For example, a specific router, an ADS configuration, or the autopilot system.
- **Assertion.** A natural language statement that describes a particular situation involving signals, variables, or conditions. For example, *While travelling at a high speed (more than 80 km/h) through a town on a foggy night, the ego-vehicle collides with nearby vehicles in dense traffic.*
- **Retrieved sentences.** Between 2 and 5 sentences that describe requirements, constraints, ODD conditions, or performance characteristics for that system. These sentences are taken verbatim from the documentation.

Your task is to compare the assertion with the retrieved sentences and assign a label using the four-point scale explained next.

3 Rating Scale

For each assertion, choose exactly one of the following labels:

- Aligned
- Overlapping
- Inconsistent
- Unrelated

The definitions below are the standard you should follow throughout the task.

3.1 Aligned

Use Aligned when:

- The assertion matches the situation described in at least one retrieved sentence, or is a more specific instance of it (i.e., the assertion stays entirely within the scope of the retrieved sentences and does not extend beyond what they entail), and
- The assertion is not inconsistent with any of the retrieved sentences.

Example:

- **System:** Insulin infusion pump
- **Assertion.** When the blood glucose level remains within the target range and the patient is not engaged in vigorous physical activity, the pump delivers only a low basal rate of insulin without triggering any alarms.
- **Retrieved sentences from the documentation.**
  - During normal operation, when the measured blood glucose level is within the configured target range, the device maintains a basal insulin delivery rate.
  - Activity-related adjustments are applied only when the user activates an exercise mode or when rapid drops in glucose are detected.
  - Alarms are issued when glucose levels are outside the safe operating range or when delivery errors are detected.
- **Reason.** The assertion is aligned with at least one of the retrieved sentences and is more specific, since it describes a particular situation with glucose in range and no vigorous activity. The assertion stays within the scope of the documentation and is not inconsistent with any of the retrieved sentences. This assertion should be rated *Aligned*.

3.2 Overlapping

Use Overlapping when:

- The assertion is strictly broader in some respect, and only partially overlaps with, the situations described in at least one retrieved sentence (i.e., the assertion generalizes the retrieved sentences, or it covers cases not supported by the retrieved sentences alone), and

- The assertion is not inconsistent with any of the retrieved sentences.

If you are unsure between Aligned and Overlapping, ask yourself: *Does the assertion clearly respect all key constraints from one of the sentences?* If some important conditions are missing or the assertion extends beyond them, choose Overlapping.

Example:

- **System:** Smart building HVAC controller
- **Assertion.** The system keeps the indoor temperature comfortable whenever the outside weather is mild.
- **Retrieved sentences from the documentation.**
  - The controller is designed to maintain indoor temperature within the comfort band when the outdoor temperature is between 10°C and 25°C, provided that all windows are closed.
  - Energy-saving mode may widen the comfort band by up to 2°C but still keeps the indoor temperature within the specified limits during occupied hours.
- **Reason.** The assertion captures the idea that the system maintains comfort under mild outdoor conditions, which overlaps with the documentation. However, it omits important constraints such as the exact temperature bounds, whether windows are closed, and whether the building is occupied. The assertion generalizes beyond what is explicitly stated and leaves out required conditions. This assertion should be rated *Overlapping*.

3.3 Inconsistent

Use Inconsistent when the assertion is inconsistent with at least one of the retrieved sentences. An inconsistency means the assertion and a retrieved sentence cannot both be true for the same situation. In other words, the assertion describes a situation as safe or acceptable, while the documentation clearly states that the same situation is outside the ODD, unsafe, or not allowed. Conversely, the assertion describes a situation as a violation, while the documentation clearly indicates it is normal or within the ODD. For example, the assertion claims a signal should be high where the documentation states it must be low (or the opposite) under the same conditions.

Example:

- **System:** Industrial robotic arm
- **Assertion.** When a human operator enters the robot workspace while the arm is moving at full speed, the system continues normal operation and does not stop the arm.
- **Retrieved sentences from the documentation.**
  - The robot must immediately stop all motion when a presence sensor detects a human in the safeguarded workspace.
  - Full-speed motion is permitted only when all safety interlocks indicate that the workspace is clear of human operators.
- **Reason.** The assertion claims that the robot continues normal operation at full speed while a human is in the workspace. The documentation requires the robot to stop all motion under the same condition and only allows full speed when the workspace is clear. These statements cannot both be true for the same situation. This assertion should be rated *Inconsistent*.

3.4 Unrelated

Use Unrelated when the assertion and the retrieved sentences are not aligned, overlapping or inconsistent. In this case, the documentation sentences concern situations or aspects of behaviour that are essentially different from what the assertion describes. There is no meaningful alignment to judge. For example, the documentation discusses braking performance while the assertion describes lane-keeping behaviour with no clear connection.

Extra classification for Unrelated. If you label an assertion as Unrelated, you must answer an additional question:

Based on the assertion alone, does the situation described correspond to

- a safe or low-risk situation,
- a violation of the ODD or preconditions, or
- neither of these

Use your best judgment based only on the wording of the assertion.

Example:

- **System:** Hospital patient monitoring system

- **Assertion.** When all measured vital signs remain within their configured normal ranges and no alarms are active, the system shows a green status indicator and does not notify clinical staff, indicating that the patient is in a stable, low-risk condition.
- **Retrieved sentences from the documentation.**
  - The system encrypts stored patient data using a hospital-approved key management infrastructure.
  - Archived monitoring records are transferred daily to the central server for long-term retention and audit purposes.
  - Network connectivity between bedside monitors and the central station is supervised using periodic heartbeat messages.
- **Reason.** The assertion explicitly describes a safe, low-risk clinical situation in terms of vital signs, alarms, and patient stability. The retrieved sentences, however, focus on data encryption, archival procedures, and network connectivity supervision, and do not address alarm logic, vital sign ranges, or patient risk level. The assertion and the sentences concern different aspects of the monitoring system, so this assertion should be rated *Unrelated*. Based on the assertion alone, the situation is a *safe or low-risk* situation.

4 How to annotate each assertion

For each assertion and its retrieved sentences, follow this procedure:

1. Read the system name. This gives context (router, ADS, autopilot), but you do not need outside knowledge about the system.
2. Read the assertion carefully. Identify what conditions or signals it mentions and whether it is describing a safe or low-risk situation, or a violation of preconditions or ODD limits.
3. Read all retrieved sentences. Understand what they state about required behaviour, constraints, ODD conditions, or performance.
4. Compare assertion and sentences. Ask yourself whether they refer to the same type of scenario? Is the assertion more specific, more general, conflicting, or unrelated?
5. Select one label from the four-point scale: Aligned, Overlapping, Inconsistent and Unrelated.

6. If you select Unrelated, answer the extra question about whether the assertion describes a safe/low-risk situation, a violation, or neither.

5 General Guidelines

Below are some general guidelines:

- Use the given texts only. Do not rely on outside knowledge of vehicles, networks, or aircraft beyond what is written.
- Differences in phrasing or synonyms are acceptable if the underlying meaning is the same.

If you have any concerns/questions, please do not hesitate to reach out to us at (balia034@uottawa.ca). The quality of your work is our top priority.

Thank you for your careful work on this task. Your judgments are essential for evaluating how well these automatically generated test validator assertions align with the behaviours described in the system documentation.