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 feedback is crucial to understanding the usefulness and alignment of these test validators with the systems' documented requirements and operating conditions.

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 systems:

- **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 relevant sentences extracted from the system's reference documentation. All information required for your judgments is contained in the assertion and the relevant sentences. Please do not consult external sources (web, manuals, etc.); use only the text we provide. We will use your review results to identify which assertions are fully aligned, partially aligned, or misaligned by the

reference documentation. You are asked to provide a rating for each assertion and provide feedback on identified issues.

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.*
- **Relevant 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 relevant 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:

- Fully aligned
- Partially aligned
- Misaligned
- Irrelevant

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

3.1 Fully aligned

Use Fully aligned when:

- The assertion is conceptually consistent with at least one relevant sentence, and
- The assertion is not logically inconsistent with any of the relevant sentences.

Note that the situation described by the assertion might be narrower or more specific than the situations described in the relevant sentences but it stays inside the boundaries defined by the relevant sentences.

Example:

- **System:** AP–DHB (autopilot climb requirement)
- **Assertion.** When the autopilot is engaged while the aircraft's nose remains pitched downward and the thrust applied to the engine is low, the aircraft does not reach the required altitude.
- **Relevant sentences from the documentation.**
  - During climb, the pilot must establish and maintain a nose-up attitude to gain altitude.
  - Excess thrust is required to initiate and sustain a climb.
  - Failure to maintain a positive pitch attitude or adequate thrust during climb can result in loss of altitude.
- **Reason.** The assertion describes a situation that clearly violates the climb requirement as stated in the sentences above. It is fully aligned with them and more specific, since it spells out a particular combination of nose-down attitude and low thrust. This assertion should be rated *Fully aligned*.

3.2 Partially aligned

Use Partially aligned when:

- The assertion is conceptually consistent with at least one relevant sentence, and
- The assertion is not logically inconsistent with any of the relevant sentences.

Note that the situation described by the assertion might be broader than or only partially overlaps with the situations described in the relevant sentences. In other words, the assertion and the documentation are compatible in spirit, yet the assertion generalizes beyond what is explicitly stated or omits important conditions present in the documentation.

If you are unsure between Fully aligned and Partially aligned, 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 Partially aligned.

Example:

- **System:** DAVE2 (ADS)
- **Assertion.** The vehicle keeps its lane when driving at low speed on a sunny day.
- **Relevant sentences from the documentation.**
  - The system is intended to operate safely during daytime in clear weather, with dry pavement, low wind speed, and air temperature within the specified range.
  - Lane-keeping performance is validated for clear sky conditions where the road markings are clearly visible and the pavement is free of snow, ice, or standing water.
- **Reason.** The assertion is compatible with the idea that the system operates safely in benign environmental conditions and maintains its lane, so it is conceptually consistent with the sentences. However, it omits several important constraints, such as road surface condition, wind speed, and temperature. The assertion generalizes beyond the explicitly stated conditions and does not mention some required factors. This assertion should be rated *Partially aligned*.

3.3 Misaligned

Use Misaligned when the assertion is logically inconsistent with at least one of the relevant sentences. A logical inconsistency means the assertion and a relevant 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:** Network router
- **Assertion.** Under heavy load, streaming traffic in the lowest priority class is guaranteed so that the quality of experience for that traffic is always maintained.
- **Relevant sentences from the documentation.**
  - Under congestion, packets in the lowest priority class may experience significant delay and loss.
  - When the link is saturated, there is no guarantee for the lowest priority class, which may be starved to protect higher priority classes.
- **Reason.** The assertion claims that the lowest priority class is guaranteed a minimum bandwidth and that its quality of experience is always maintained under heavy load. The documentation explicitly states that there is no minimum bandwidth guarantee for this class and that it may be starved. These statements cannot both be true for the same situation. This assertion should be rated *Misaligned*.

3.4 Irrelevant

Use Irrelevant when the assertion and the relevant sentences are neither conceptually consistent nor clearly 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 Irrelevant. If you label an assertion as Irrelevant, 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:** AP-TWN (R2) (ADS)
- **Assertion.** When there are no nearby vehicles, the ego-vehicle does not collide with nearby vehicles.
- **Relevant sentences from the documentation.**
  - The system is designed to keep the vehicle centred within its lane on roads with clearly visible lane markings.
  - Lane-keeping performance is evaluated in terms of lateral deviation from the lane centerline and the frequency of lane departures.
  - The documentation specifies environmental conditions such as visibility and road marking quality that are required for valid lane-keeping operation.
- **Reason.** The assertion describes a safe situation in which collisions with nearby vehicles cannot occur because no nearby vehicles are present. The relevant sentences, however, focus on lane-keeping behaviour, lateral deviation, and environmental conditions for lane-keeping, and do not mention collision risk or the presence of other vehicles. The assertion and the sentences concern different aspects of the ADS, so this assertion should be rated *Irrelevant*. Based on the assertion alone, the situation appears to be a *safe or low-risk situation*.

3.5 Feedback

For any assertion that is not rated Fully aligned (i.e., Partially aligned, Misaligned, or Irrelevant), please write a short explanation (one or two sentences) summarizing why you chose that label. For Partially aligned, indicate what is missing, more general, or incomplete. For Misaligned, identify the main contradiction. For irrelevant, state why there is no conceptual match.

The goal is to capture the key reason for your decision, not to write a long analysis.

4 How to annotate each item

For each assertion and its relevant 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 relevant 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: Fully aligned, partially aligned, misaligned and irrelevant.

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

7. Provide a brief explanation when required (see Section 3.5).

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.

- Once you get a sense of what you consider Fully vs Partially aligned, or Partially vs Irrelevant, apply the same logic across all items.

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 reflect the behaviours intended by the system documentation.