# Citcom Responsible AI Label – Assessment Guidelines for TEF sites

The Citcom Responsible AI Label is a programme through which TEF site partners assess AI systems deployed in smart city contexts. It covers systems across the full range of smart city domains: infrastructure, mobility, energy grids, citizen services, urban planning, and related areas.

This document serves as a guideline for TEF sites on how to conduct the assessment and compile the assessment report, a template for which can be found [add link later]. The report itself is composed of two parts:

- a **general part**, where information about the company, the system under assessment, and the context in which the assessment is performed is recorded;
- a **dedicated part for each badge**, which contains a list of questions to be answered. These questions represent the minimal set that must be addressed; a TEF site may decide to include additional ones where the specifics of the system or context warrant it.

Participating in the Label process means submitting an AI system, or part of an AI system (such as model, data etc.), to independent expert assessment by a TEF site partner. The outcome of that assessment is the award of one or more badges, each corresponding to a specific evaluated dimension. The badges are the tangible result of the Label process: a provider awarded a badge can use it to signal to municipalities, procurers, and the public, that their system has been independently scrutinised on that dimension.

The three badge dimensions are:

- **Technical Testing:** how the system has been evaluated for trustworthiness, including performance, robustness, fairness, safety, explainability and several others
- **Governance:** the governance, oversight, and operational practices around the system
- **Impact:** the effects of the system on people, communities, and society

Badges can be awarded independently. A provider may go through the Label process and receive all three, or only one or two, depending on what was assessed and what the assessment found. Receiving a badge on one dimension says nothing about the others. The Label process is not pass/fail: it is an expert judgement, grounded in evidence, about whether a system, or part of it, has been seriously considered on that dimension.

**The role of the TEF site**

TEF sites conduct assessments independently. TEF sites are not auditing compliance with a standard, they are forming a professional opinion, as a domain expert, about whether this system has been developed and operated responsibly in its smart city context. The guidelines that follow are designed to harmonise how that judgement is exercised and documented across TEF sites, not to replace it.

The Label process typically begins with a conversation: with the provider, who may request a specific badge or may be unfamiliar with the framework and need it explained, or internally at the TEF site, where the TEF site may identify a deployment that warrants assessment. There is no fixed entry point. What matters is that the scope of the assessment, and which badges are in play, is clear before substantive work begins.

To finalise the report, we encourage the four eyes principle: the completed report should be reviewed by a colleague who has not taken part in the assessment, to check that conclusions follow from findings and that the overall picture is coherent. Where possible, this reviewer should be from a different TEF site, though this will depend on availability.

In the following, each badge is described along four guiding aspects — what you're assessing, key areas of inquiry, smart city considerations, and report guidance — to provide general tips on how to approach the assessment and exercise judgement when completing the corresponding questions in the report template.

# Badge 1: Technical Testing

**What you're assessing**

Whether the system has been tested rigorously and honestly, and whether the provider genuinely understands how it behaves, including at the edges and under conditions that don't favour it.

**Key areas of inquiry**

1. **Testing methodology:** What was tested, under what conditions, and against what benchmarks or datasets? The central question is not just what metrics were reported, but whether the right things were measured in the right way, against data that reflects the actual deployment environment. A computer vision system for pedestrian detection benchmarked only on daytime, clear-weather images has not been adequately tested for urban deployment. A demand forecasting tool for public transport should have been evaluated against data capturing strikes, major incidents, and seasonal variation, not only normal operating conditions.
2. **Independence of testing:** Who conducted the testing, the development team, a separate internal team, or an external party? Note where the assessment sits on that spectrum and what it implies for the reliability of reported results. There is a meaningful difference between self-reported performance figures and findings produced by parties with no stake in the outcome. Independence does not guarantee quality, but its absence is a relevant factor in judging credibility.
3. **Performance:** Are the chosen metrics appropriate for the task and deployment context? Results reported on unrepresentative data, or without acknowledgment of uncertainty or variability, should be treated with caution. A single aggregate accuracy figure tells you little if the system will operate across heterogeneous environments or under varying load. Ask whether the provider can account for performance variability, not just central tendency.
4. **Fairness:** Does the testing evidence show how the system performs across different groups and contexts, by neighbourhood, demographic, or infrastructure type, or was it evaluated only on aggregate metrics that could mask uneven performance? Aggregate figures can conceal substantial disparities. Ask for disaggregated results and, where they are absent, treat that absence as a finding.
5. **Robustness and failure modes:** Can the provider speak concretely about how their system fails, and what mitigations are in place? A provider who can articulate specific failure modes and the conditions that produce them is more credible than one who can only describe strengths. Ask about behaviour under atypical conditions, edge cases, data quality degradation, adversarial inputs, infrastructure failures. Where the provider cannot answer these questions with specificity, that is itself informative.
6. **Safety and harmful outputs:** Have outputs been evaluated for safety, bias, or harmful behaviour? How rigorous and comprehensive was that evaluation relative to the deployment context? The bar here should be proportionate to the stakes: a system making decisions that affect public access to services, mobility, or resource allocation warrants more thorough safety evaluation than one operating in a lower-stakes context. Ask what failure scenarios were explicitly tested, not just assumed away.
7. **Explainability:** To what extent can the system's outputs be explained, and has explainability been tested or documented for the relevant use cases? Explainability is not an abstract property, it has to be assessed in relation to who needs to understand the outputs and for what purpose. An operator appealing a decision, a regulator auditing the system, and a member of the public seeking to understand why a service was unavailable all have different explainability needs. Ask whether the provider has considered these concretely, not just in general terms.

8. **Gaps and limitations:** Where was evidence unavailable, access restricted, or a confident assessment not possible? Incomplete or evasive documentation is itself a finding, not a reason to defer judgement. Be explicit about what could not be assessed and what that implies for the reliability of this badge overall. A badge awarded on partial evidence should say so.

**Smart city considerations**

Smart city systems operate in variable, real-world conditions with populations that are more diverse than most test sets. This cuts across all eight dimensions above. Ask specifically whether testing reflected the actual deployment environment: different neighbourhoods, seasonal variation, infrastructure heterogeneity, the social and cultural background of citizens, and variation in traffic and usage patterns. Systems that perform well on average but poorly in specific areas or for specific populations can entrench existing inequalities rather than address them. Where disaggregated evidence across these dimensions is absent, that absence should be noted explicitly in the assessment.

**Report guidance**

Ground the assessment in specifics: what was examined, what it showed, and where the gaps are. Where testing was strong in some areas and weak in others, say so precisely. An assessment that finds rigorous performance benchmarking but no disaggregated fairness data, or strong internal testing but no independent validation, should reflect that clearly rather than averaging it into a single impression.

# Badge 2: Governance

**What you're assessing**

Whether the organisation managing the system, or the part being assessed, has the governance structures, oversight mechanisms, and operational practices to run it responsibly over time, not just at the point of deployment.

**Key areas of inquiry**

1. **Risk management:** Is there a genuine, working approach to identifying and managing risks, specific to this system and updated as circumstances change? A risk framework written at procurement and never revisited is a weak signal, particularly for systems that evolve or whose deployment context shifts. An energy grid optimisation system that has not revisited its risk assessment since integrating new renewable sources is a concrete example. Ask whether risk management is treated as a living practice or as a box ticked at the start of the project.
2. **Human oversight:** Is oversight meaningful, with real authority to intervene? Nominal arrangements that look good on paper but lack operational substance are common. For a system managing traffic signal timing across a city, it matters whether a human operator can actually override it in real time, or whether the override process is so cumbersome it is never used. Ask not just whether an override mechanism exists, but whether it has been used, tested, and is practically accessible to those responsible.
3. **Accountability:** Are responsibilities clearly assigned, and do they reach senior decision-makers? In deployments involving both a private provider and a public authority, which covers most smart city deployments, ask concretely where accountability sits when something goes wrong. Ambiguity between provider and municipal authority is common and should be noted explicitly where it exists.
4. **Monitoring and incident response:** Is the system's behaviour monitored after deployment, and is there a credible response process for failures? Ask not only whether a monitoring and incident response process exists, but whether it has actually been exercised. A process that has never been tested offers weaker assurance than one that has been stress-tested against real or simulated incidents.

5. **Governance over time:** Smart city deployments often outlast the governance arrangements designed at the start: contracts change, providers change, systems get updated. Ask how governance adapts over time, not just how it was initially designed. Has the provider seriously considered what happens when the system evolves, ownership shifts, or the deployment context changes? A mobility-as-a-service platform that has changed hands since deployment is a good test case for whether continuity of accountability was built in.
6. **Documentation:** Is the documentation complete, specific, and candid? Incomplete, generic, or evasive documentation is itself a finding. Note where materials were requested but not provided, where answers were vague where specificity was expected, or where documentation appears to describe intended practice rather than actual practice.
7. **Gaps and limitations:** Where was evidence unavailable, access restricted, or a confident assessment not possible? As with the Technical Testing badge, absence of evidence is not a reason to defer judgement but a finding in its own right. Be explicit about what could not be assessed and what that implies for the reliability of this badge overall.

**Smart city considerations**

Smart city deployments often involve multiple parties, long operational lifespans, and systems that evolve well beyond their initial design. Governance arrangements that were adequate at deployment may become inadequate as the system changes or as the institutional context shifts. Ask specifically how accountability is maintained across organisational boundaries and over time, not just how it was structured at the outset.

**Report guidance**

The key distinction is between organisations that have genuinely operationalised governance and those that have produced governance documents. Ground that distinction in what you observed in conversations and in the evidence reviewed, not just in what the documentation says. Where governance looks strong on paper but shows signs of being nominal in practice, say so directly. An assessment that finds clear accountability structures but an untested incident response process, or robust monitoring but no evidence of governance adaptation over time, should reflect those distinctions precisely rather than treating governance as uniformly strong or weak.

# Badge 3: Impact

**What you're assessing**

The actual and potential effects of the system on the people it affects, including those who never interact with it directly, and whether those effects have been seriously considered by the provider.

**Key areas of inquiry**

1. **Affected populations:** A provider who describes impact only in terms of their primary users is missing most of the picture. A smart parking system affects not just drivers but residents of the streets where it redirects traffic, and businesses whose footfall changes as a result. Mapping direct and indirect affected populations is the starting point. Ask whether the provider can speak concretely about effects on people beyond their primary use case, and treat a narrow or user-centric impact analysis as a gap in itself.
2. **Fairness and differential impact:** Are there systematic differences in how the system performs or what it delivers across different groups, by neighbourhood, age, socioeconomic status, disability, or other relevant dimensions? Aggregate impact assessments can conceal substantial disparities. A municipality chatbot should not provide different answers about schooling or social housing to residents depending on their ethnic background. Ask for disaggregated evidence and, where it is absent, note that absence explicitly.

3. **Transparency and recourse:** Are people informed when AI is being used in ways that affect them, and do they have meaningful recourse? In a system that influences access to public services such as housing allocation or social benefit processing, the absence of explanation and appeal mechanisms is a significant gap. Municipalities deploying such systems have a particular responsibility to ensure that recourse is accessible and not merely nominal, and the assessment should reflect whether that responsibility is being taken seriously.
4. **Broader societal effects:** For systems operating at scale, consider whether the deployment changes how public space functions, creates new dependencies, or has effects on employment, public trust, or social systems more broadly. A city that has fully automated its traffic management around a single proprietary system has created a fragility that did not previously exist. Ask whether the provider has thought carefully about these second-order effects, or whether their impact analysis stops at the intended use case.
5. **Participation and community input:** Have the communities affected by the system had any role in its design, validation, or deployment? This is often where the seriousness of a provider's commitment to impact becomes clearest. A provider who can speak in detail about how community input shaped the system, or who acknowledges honestly where it was absent, is more credible than one who offers generic statements about stakeholder engagement. Note how willing and transparent the provider has been about effects on communities beyond their primary users.
6. **Environmental impact:** What is known about the system's energy consumption and environmental footprint, particularly for compute-intensive components? This is an area where evidence is often limited, but the question should still be asked and the response noted. Where environmental impact has not been assessed at all, that is itself a finding.
7. **Gaps and limitations:** What could not be assessed at this point in time, and what additional evidence or time would be needed for a fuller picture? This dimension is the most contextual and the least reducible to a checklist, and it is therefore the one where honest acknowledgment of uncertainty is most important. Be explicit about what was not possible to assess and what that implies.

**Smart city considerations**

This is where the public nature of smart city deployments matters most. These systems affect everyone, including people with no alternatives: elderly residents dependent on public transport, people with disabilities navigating public space, communities without the resources to opt out. Unlike private services, smart city systems cannot be avoided by those who find them harmful or inadequate. Ask specifically whether the provider has considered how communities, including those without resources or technical literacy, can participate in or at least be meaningfully informed about the systems that shape their daily lives.

**Report guidance**

This dimension is the most contextual and the least reducible to a checklist. Be specific about which populations and effects were considered, honest about what could and could not be assessed, and direct about concerns even where they do not amount to clear findings. Generic statements about potential risks are not useful; concrete observations about what was found or notably absent are. Where the provider's impact analysis is narrow, optimistic, or focused exclusively on intended users, say so directly rather than hedging.