

Motivation & Problem

Advancing robots from **lab research** to **real-world deployment** is a central challenge in service robotics.

Gap in existing approaches:

- Competitions test TRL 2–5 in structured, reproducible environments
- Mostly single-dimensional (task performance only)
- No applicable EU safety standard for robots in *public spaces*

Our contribution: An interdisciplinary, multi-phase benchmarking framework evaluated across three real-world use cases over **3 years**.

Three Public-Space Use Cases



Public Library SoftBank Pepper

Book recommendations & reading samples, Duisburg



Park Cleaning Angsa robot

Cigarette butts & bottle caps, Munich Glyptothek



Pedestrian Underpass Cleaning

Adlatus SR1300 — Floor cleaning among pedestrians, Ulm Main Station

Benchmarking Method

Three incremental phases:

Phase 1	Phase 2	Phase 3
Consensus workshop	Lab/controlled tests	Field tests in practice
Eval. concept	Bottrop, Stuttgart,	Duisburg, Munich,
Bonn, Dec 2023	Ulm, 2024/25	Ulm, 2024/25

Benchmarking Categories

Four benchmarking categories — applied across all phases:

Task Fulfillment Success rates, efficiency, area covered, dialog success	Interaction Quality ISO 9241-110, usability, trust, diversity sensitivity
Safety Functional & mechanical, e-stop, obstacle detection	Economic Viability Business model, value proposition, stakeholder network

Benchmarking Panel

≈7 members, reflecting experts from all four categories:

- Roboticians
- HRI Psychologists
- Safety Experts
- Business Economists
- Science Communicators

Inspired by the **EuRoC** freestyle stage: each team defines their own metrics, refined via consensus workshops.

Phase 1 — Evaluation Concept

Use-case teams proposed metrics in four categories. The benchmarking panel reviewed and refined them in a joint workshop (Consensus Workshop (Bonn, Dec 2023)), establishing an evaluation concept for all three use cases.

Library	Park	Underpass

Phase 2 — Lab Evaluation

Library (Bottrop)	Park (Stuttgart)	Underpass (Ulm)

Phase 3 — Field Deployment

Library (Duisburg)	Park (Munich)	Underpass (Ulm)

Task Fulfillment Results

Public Library (Pepper):

- Phase 2 dialog success: **35.7%** (weighted checklist)
- Phase 3 dialog success: **43.2%** — modest improvement
- Missing wishlist features; unintuitive feedback; LLM latency

Park Cleaning (Angsa):

- Phase 2: 24 items collected, 7 missed (out-of-distribution objects)
- Phase 3 Exp. 2: **100%** collection on gravel; grass embeds bottle caps

Pedestrian Underpass (Adlatus):

- Phase 2: **95%** completeness (67.6 m², 5 min, 13.5 m²/min)
- Phase 3: 14.5 m²/min in live deployment
- Coffee stains spread rather than cleaned

Interaction Quality Results

Shared **13-dimension guideline** developed (ISO 9241, trust, accessibility, ...). Guideline **too generic** for direct cross-use-case application — contextual tailoring required. Assessment remained **predominantly expert-based and subjective**; passers did not use questionnaires.

Library: long LLM processing times; non-adaptive gestures; inconsistent gaze direction.

Park: rapid turns alarmed pedestrians; colour status LEDs mis-understood.

Underpass: multi-modal signals (lights, icons, speech) caused perceptual overload in confined space.

Safety Results

No applicable EU standard for robots in public spaces.

Custom test pieces (park robot):

- Squatted/standing five-year-old child (denim/molleton)
- Hiking pole; chrome grill leg (hard to detect)

Phase 2 outcomes:

Test	Park	Underpass
Emergency stop	✓ Pass	✓ Pass
Obstacle detect	⚠ Camera not rated	✓ Pass
Stair/edge drop	—	✓ Pass

Park camera has *no performance level* → bumper is sole safety function. Safety tests concluded after Phase 2 (no hardware changes).

Safety Test Impressions (Phase 2)



Park: safety-stop test

Underpass: drop & detect

Economic Viability Results

Large **variation in business-model knowledge** across teams

Library use case as **best practice**: iterative user-centred toolbox (HRJ → Stakeholder Network → BMC → MVP)

Economic viability must be integrated **from Phase 1** — not retrofitted

Recurring tension: *inclusivity* vs. *cost control*

Platform / B2B ecosystem identified as path to scalability

Discussion & Conclusions

Key lessons learned:

Lab-to-field gap is **real and use-case specific**

Interaction quality remains **inherently subjective**; objective proxies not yet mature

No EU standard for public HRI; **custom test pieces needed**

Economic viability is a **multi-level challenge** (micro / meso / macro)

Interdisciplinarity creates tension but is **essential** at TRL 6–7

Outlook:

Long-term deployments required; short events are insufficient at high TRL

Standardised interaction & safety metrics for public spaces urgently needed

EuRoC-style *freestyle* benchmarking is a promising alternative to traditional competitions

Acknowledgements



Bundesministerium
für Forschung, Technologie
und Raumfahrt

Funded by: German Federal Ministry of Research, Technology and Space (BMFT), grants 16SV8680, 16SV8681, 16SV8683.
Project: Transferzentrum Robotik im Alltag (RimA).