V1 · SCENARIO BANK

Wearable Assistant Context Benchmark

A model-selection benchmark for AI wearable assistants used actively for advice or coaching (smart glasses, ear worn devices), with audio/video/text input and audio/text output. Tests whether the assistant updates to **current context** instead of staying anchored to **prior context** as the user's situation changes between turns.

50 scenarios

4 published runs

PRODUCT PROBLEM

An AI assistant used in-the-moment for advice or coaching fails if the user must constantly restate what they're looking at, holding, or referring to. To be frictionless on an AI wearable, the assistant must silently track the user's situational context as they move, look around, or swap objects.

Examples:

- Task shift: A user asks about a recipe on a tablet, looks down at a pan on the stove, and asks, "Is this done?" The

WHAT THIS BENCHMARK MEASURES

This benchmark measures context tracking. Each scenario has three turns with a deliberate context shift between Turn 1 and Turn 2. The shift is visible only in the video channel; the user does not announce it out loud.

Primary score

Balanced Turn 2 accuracy across the `current` and `prior` classes under the `baseline` prompt condition.

assistant must evaluate the food in the pan, not the text on the tablet.

- Object swap: A user looks at a stripped bolt, picks up a pair of pliers, and asks, "Will this work?" The assistant must evaluate the pliers in hand, not the bolt.

Auxiliary signal

`clarify` and `abstain` per-class accuracy, plus a repair rate after Turn 2 misses.

THREE INPUTS

Every scenario carries three inputs, each with a different audience:

- **Audio.** User speech (`turn_1_user` , `turn_2_user`). Represented as text transcripts in v1, not raw audio; acoustic grounding, speaker attribution, and ambient audio cues are out of scope. Natural phrasing, with references that depend on the scene; the user never names the object outright or announces the shift. Visible to candidate and judge.
- **Video.** Scene descriptions (`turn_1_image` , `turn_2_image`) injected on the user side as `[Camera: …]` blocks (the literal field marker from the scenario JSON). Shape,

SCENARIO BANK

50 scenarios across eight kinds of context shift. The shape of the shift is what the categories describe.

12 object in hand	**8** object state
6 sequential task	**6** location
5 object in view	**5** absent referent
4 screen content	
4 pre-conversation recall	

Target context distribution: **33 current**, **12 prior**, **3 clarify**, **2 abstain**. Difficulty: 15 easy, 20 medium, 15 hard. Coverage spans 16 activity domains.

material, motion, position; no object names. The video input is short text descriptions, used as a proxy for real video frames. Visible to candidate and judge.

- **Ground truth.** Answer keys naming the actual objects in Turn 1 and Turn 2. Visible to the judge only.

HOW TO READ A SCORE

- Score deltas between models on the same release matter more than absolute values.
- Strong on `current`, weak on `prior` means the model defaults to the latest frame and ignores user references to earlier state. The headline number is balanced for that reason.
- `condition_a` and `condition_b` are prompt-sensitivity diagnostics, not replacement scores. A model that only clears the bar with the pre-answer scaffold is a different signal than one that clears it under `baseline`.
- The repair rate stands in for user-correction cost after a miss; it is not part of the primary number.

WHAT THIS BENCHMARK DOES NOT MEASURE

- **Advice quality.** The judge does not check whether the response is correct, safe, or domain-appropriate. A confidently wrong answer can pass if it picks the right context.
- **Multi-turn flow past Turn 2.** The conversation is three turns. Long conversations and branching dialogue are out of scope.
- **Real video.** The video input is text descriptions of the scene, not actual frames. Performance here is not a guarantee of performance on real video.
- **Proactive coaching.** Only direct-question responses are scored.

- **Domain knowledge depth.** Coverage is broad across 16 domains but shallow in any one.

REPO POINTERS

- README.md: product framing and quickstart
- benchmark_spec.md: the benchmark contract
- benchmark_notes.md: score interpretation and limitations
- schema.md: scenario field reference
- scenario_authoring_rules.md: authoring rules and validation checklist
- dataset_card.md: bank statistics

RESULTS

v1 results: 6 runs, 50 + 20 scenarios

Camera ablation: `baseline` 60.6% (CI 54.1–67.1) → `ablation-no-camera` 14.4% (CI 9.1–19.7). A **46.2 pp drop** when the camera description is removed. The model relies heavily on visual input and can't recover the answers from the user's words alone. `baseline-alt` (Gemini-Flash) 77.7%; `baseline-qwen-cross-family` (Qwen3-VL-Plus + Gemini judge) 54.2% with `current` 100% / `prior` 8.3%, showing the model grounds in the latest frame and struggles to refer back. Deictic-repair ablation: deictic gesture-style repairs recover 100% of misses where they apply; named repairs recover 30% of misses on harder scenarios. Adversarial pack 67.3% with cross-LLM agreement κ=0.443. Full table at n-dryer.github.io/wearable-assistant-context-bench.

This benchmark supports a practical model-selection decision for a live AI wearable assistant the user is actively engaging with for advice or coaching. It is not a general multimodal benchmark. A model that fails it cannot serve as an in-the-moment multimodal assistant; a model that passes still needs separate evaluation for advice quality, real video, latency, and everything else outside the context-tracking question.