

Typically, conversational agents discard respiratory sounds as background noise. HealthCUES listens to them instead — and decides, in real time, whether the agent should escalate, without ever breaking the flow of the conversation.

1 The discarded signal

Respiratory sounds are **clinical biomarkers**: in a healthcare conversation, how a patient breathes and coughs changes what a clinician asks next. Audio-based respiratory assessment is among the most valuable remote diagnostic tools available, yet **no conversational agent exploits it** — the signal arrives in every telehealth call and is thrown away before it reaches the dialogue policy.

PRIOR SYSTEMS
Respiratory sound → **acoustic noise** → discarded

HEALTHCUES
Respiratory sound → **clinical signal** → agent acts

2 Why it is hard

LATENCY
Must detect inside the inter-turn pause — under 500 ms.

SIGNAL
8 kHz μ-law telephony, SNR down to 5 dB.

CLINICAL UTILITY
Binary "cough / no cough" is not actionable.

DIALOGUE FIT
Alerting on every cough is intrusive and robotic.

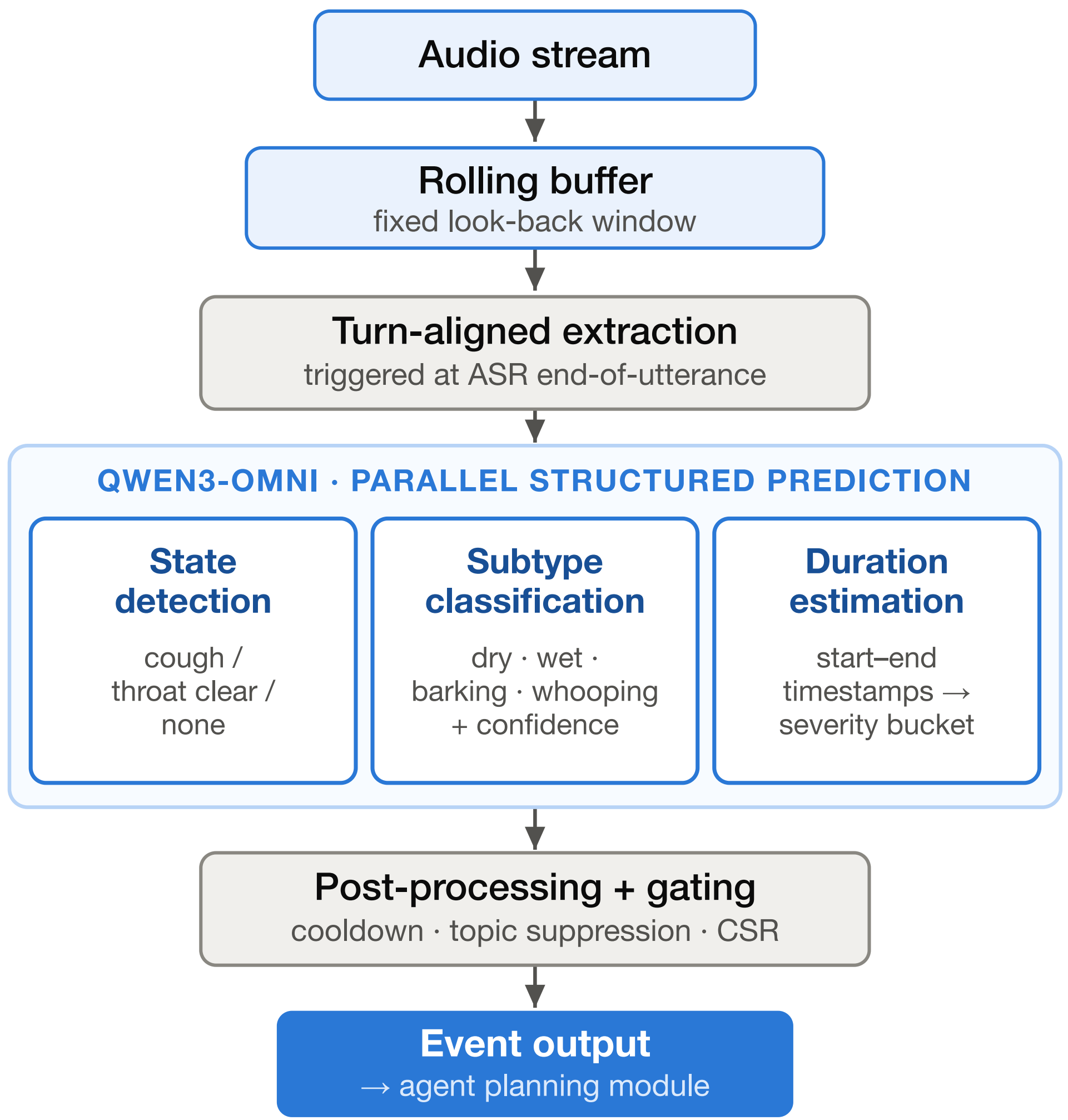
SYSTEM	STRM	THROAT	SUBTYPE	DUR.	DIAL.
BEATs	✓	×	×	×	×
PANNs	×	×	×	×	×
CoughVID FT	✓	×	×	×	×
Coswara CNN	×	×	×	×	×
HealthCUES	✓	✓	✓	✓	✓

Five functions are required for conversational integration. No off-the-shelf detector provides more than one; HealthCUES provides all five, from a single general-purpose MLLM.

3 Contributions

- Streaming detection** with turn-aligned audio processing, built for live conversational agents.
- Fine-grained analytics**: four-way subtype, confidence scores, and start–end boundaries — absent from all prior cough systems.
- Dialogue-aware gating**: cooldown, topic suppression, and cough-to-speech ratio, to prevent alert fatigue.
- Systematic evaluation** with baselines, external validation, and licensed-clinician review.

4 Architecture



Audio is buffered and extracted at turn boundaries, then processed by three parallel inference tasks. Decomposing the problem allows per-component prompt optimization and graceful degradation if any single task fails.

5 Turn-aligned streaming

A rolling buffer keeps recent audio so the system can look **backwards** when a turn completes: on an ASR end-of-utterance signal it analyzes a **fixed look-back window**. Operating at the turn level matches the granularity at which the agent plans — and because the look-back spans the preceding utterance, respiratory events mid-turn, or just before speech onset, are still captured.

No oracle boundaries. Evaluation uses the same rolling-buffer mechanism that runs in production, not pre-segmented clips.

6 Constrained structured output

All three tasks emit a fixed schema, so the agent never has to parse free text:

State	coughing
Subtype + confidence	[wet][0.88]
Duration	[1.2–4.3]
Severity bucket	medium (2–6 s)

Throat clearing is separated explicitly — it carries different clinical significance and warrants a different conversational response.

7 Dialogue-aware gating

Accurate detection alone would be **counterproductive**: alerting on every cough disrupts flow; repeating the same follow-up feels robotic. Three mechanisms decide whether an event reaches the agent.

- Cough-to-speech ratio.** Over a fixed sliding window, $CSR = \sum \text{cough duration} \div \sum \text{speech duration}$ — separating one isolated cough from a sustained pattern.
- Cooldown.** A quiet period stops re-triggering while the patient is still coughing through the same episode.
- Topic suppression.** If the agent's recent turns already contain cough language, new detections are suppressed.

Clinically urgent subtypes (wet, barking, whooping) and prolonged episodes clear **lower thresholds** than brief dry coughs.

8 What the agent does with it

When a detection clears gating, the pipeline passes a natural-language action to the agent's planner. **The patient only ever hears the agent's turn.**

Dry cough · x2 · 1.8 s
count threshold met

SYSTEM → PLANNER
Coughed twice — dry. Ask if not already addressed.

AGENT
"Has that cough been bothering you?"

Wet cough · 3.0 s
2 events · wet lowers the bar

SYSTEM → PLANNER
Patient has coughed twice. Sounds like a wet cough. Ask about it if not addressed.

AGENT
"I noticed you've been coughing. Does it feel like you're bringing anything up?"

Barking cough · 4.2 s
long duration → immediate

SYSTEM → PLANNER
Prolonged barking cough. Raise it now.

AGENT
"That cough sounds quite harsh. How long have you had it?"

Wet cough · 2.4 s
topic suppression active

SYSTEM → PLANNER
Wet cough — already raised this turn.

No agent turn — suppressed, so the follow-up is not repeated.

Detection surfaces respiratory status **without requiring patient self-report** — the gain clinicians rated most valuable.

9 Results

93%
F1, cough detection
(847 in-house segments)

340 ms
avg end-to-end latency
(p95 = 520 ms)

0.91
macro-F1 on AMI corpus
zero domain adaptation

0.75
weighted-F1, wet vs dry
the clinically central call

847 held-out conversational segments (12.3 h): 8 kHz μ-law telephony and 16 kHz mic capture, SNR 5–30 dB. Three U.S.-licensed clinicians labeled presence, subtype, and timestamps (κ = 0.82 presence, 0.67 subtype).

10 Detection vs. off-the-shelf

SYSTEM	COUGH F1	F1	MS
BEATs		59.1%	180
PANNs		76.4%	210
CoughVID FT		79.8%	95
HealthCUES		93.0%	340

■ HealthCUES ■ Off-the-shelf baseline

Baselines are not compute- or backbone-matched — they differ in scale, pretraining data, and task formulation, so this is capability-illustrative, not a like-for-like ranking. CoughVID FT is fastest at 95 ms but binary-only; our 340 ms buys multi-task inference.

11 External validation · AMI

On the **AMI Meeting Corpus**, where coughs and throat clears co-occur with multi-party speech, with **no domain adaptation** (n = 719):

THREE-WAY STATE	P	R	F1
Coughing	0.99	0.76	0.86
Throat clearing	0.97	0.91	0.94
None	0.86	0.99	0.92

Accuracy 0.92**Macro-F1 0.91**

12 Subtype breakdown

Dry		0.89
Wet		0.55
Barking		0.23
Whooping		0.24

■ Dependable operating point today ■ Data-limited

Per-class F1, conditional on cough presence. Barking and whooping are acoustically subtler *and* far rarer in deployment, so limited data makes both reliable recognition and *evaluation* hard. We treat **wet/dry as today's operating point**.

13 Clinician study

Three licensed U.S. registered nurses each ran **6 structured telehealth scenarios**, rating response appropriateness, naturalness, and dialogue efficiency.

- Agent **correctly altered its dialogue strategy** on detected respiratory state; check-in phrasing rated natural.
- Unprompted subtype identification **reduced conversational burden**, improving information-gathering efficiency.
- Weaknesses: consistency when coughs overlap *agent* speech; gating during prolonged episodes — now informing gating refinements.

Scope. Informational respiratory analytics, **not medical diagnosis** — it triggers follow-up questions, never clinical recommendations. Audio is processed transiently, never stored.