RMIT UNIVERSITY School of Science, Engineering & Technology | dewly.ai

SSET CAPSTONE 2026
17 SEPTEMBER 2026

Dewly AI Voice Agent

INBOUND & OUTBOUND CALLS FOR SHORT-STAY OPERATORS

Extending Dewly's AI duty manager from chat to the phone line: a voice agent mapped from the existing LangGraph agentic system, so guest calls and staff dispatch run on one reasoning core across two channels.

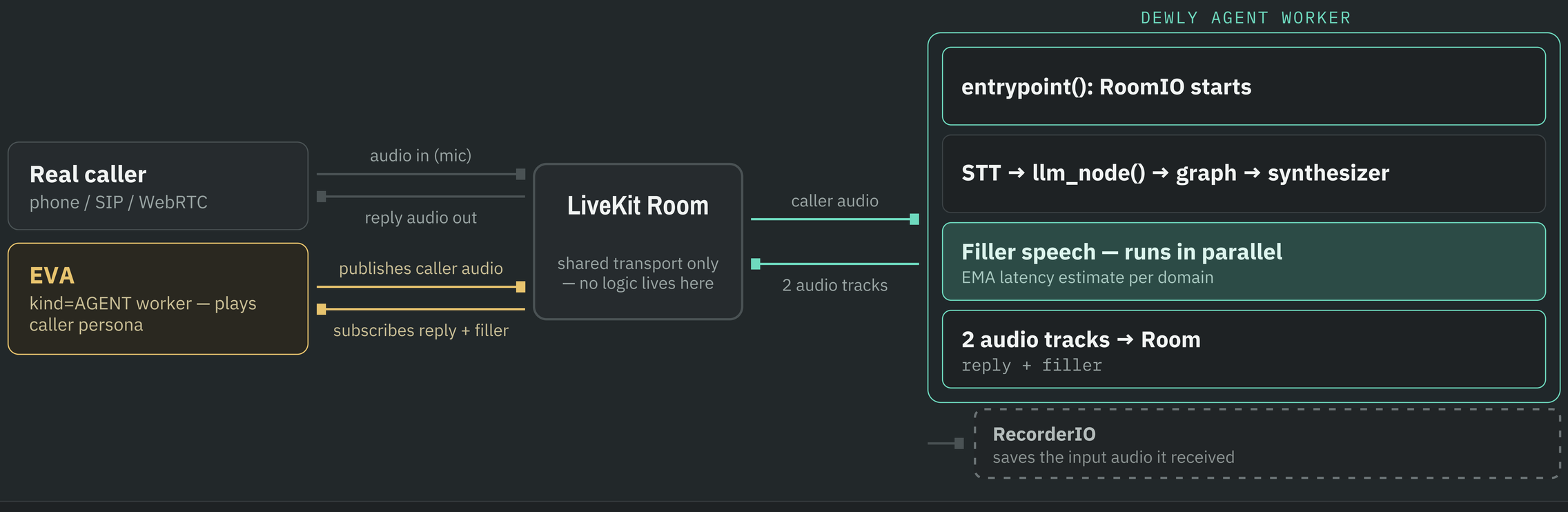
Scan for the live demo rmitvn-showcase.com/voiceagent

P50 ≤ 1.5s
turn latency, 800ms first audio

4 Domains Agents
pricing/availability, cancel/modify, escalation, search

12 quality metrics
experience, diagnostic & latency axes

One reasoning core, two channels: chat and voice. A fix to the duty-manager logic ships to both at once.



Background & Motivation

Short-stay operators run on the phone: guests call to extend a stay, report a broken air-conditioner or ask for a late check-out, while one duty manager covers dozens of rooms. Dewly already answers these requests in chat, but the phone line stays manual calls are missed at night, context is lost when staff are dispatched, and every conversation is re-typed into the log. Voice carries the most urgency and the least automation, and it is also the hardest to judge: a wrong tone or a two-second silence breaks a call in a way a chat message never would.

Objectives

01 Map the chat agent onto a real-time voice pipeline without forking the reasoning core.

02 Handle both directions: inbound guest calls and outbound dispatch to housekeeping and maintenance.

03 Hold a P50 of 1.5 s per turn, with first audio inside 800 ms.

04 Build a repeatable evaluation framework so voice quality is measured per turn, not judged by ear.

05 Stop within 300ms on a real interruption, without cutting off natural backchannels

Tech Stack

LangGraph · LiveKit · Gemini · Deepgram · Cartesia · Twilio telephony · EVA framework

Methodology

01 Channel & requirement mapping
Audited the chat agent's tools and intents against real call recordings to scope what voice must support.

02 Porting the LangGraph core
Wrapped the graph behind a transport-agnostic interface so chat and voice share one set of nodes, tools and prompts.

03 Voice pipeline integration
Connected Twilio telephony to a LiveKit agent worker with Deepgram STT and Cartesia TTS, tuning turn detection and barge-in.

04 EVA evaluation loop
Scored every prototype call on responsiveness, speakability and tone, then fed failing turns back into prompt and pipeline changes.

05 Pilot & deployment
Ran the agent against hotel test lines and human callers for 60 days, logging every room for review and regression.

Conclusion & Future Work

Routing voice through the existing LangGraph core proved the cheaper path: the duty manager gained a phone line without a second agent to maintain, and EVA turned "does it sound right" into a number the team could act on each iteration. Next: sub-second P50 with speculative TTS, full-call rubrics such as task completion and escalation accuracy, and a link into the property management system so a call updates the booking directly.

Evaluation Framework - EVA

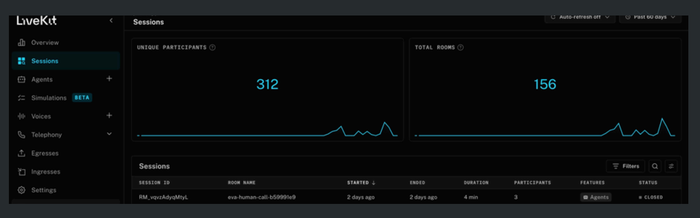
An LLM/audio/deterministic judge mix scores 12 metrics per turn - no composite score; each reports independently


```
"speakability": {
  "name": "speakability",
  "score": 0.857,
  "normalized_score": 0.857,
  "details": {
    "per_turn_ratings": {
      "1": 0,
      "5": 1,
      "6": 1,
      "7": 1,
      "8": 1,
      "10": 1,
      "12": 1
    },
    "per_turn_explanations": {
      "1": "The text contains markdown bullet points (*) and bold formatting (**863 Pierce Ave**), which violate th
      "5": "The text is written in natural language without any markdown formatting, emojis, missing spaces, or unp
      "6": "The text contains only standard punctuation and words. There are no visual formatting elements, emojis,
      "7": "The response is plain text, easily spoken, and free of any markdown, emojis, or formatting violations."
      "8": "The response consists of regular sentences with standard punctuation. It does not contain any formattin
      "10": "The text includes standard punctuation and a number sign (#659) which is naturally spoken as 'number'.
      "12": "The text is purely conversational and contains no visual formatting, markdown, emojis, or unpronounce
    },
    "judge_prompt": "You are an expert evaluator analyzing whether text is voice-friendly and appropriate for text-
    "judge_raw_response": "```json\n[\n  {\n    \"turn_id\": 1,\n    \"explanation\": \"The text contains markdown
    "aggregation": "mean",
    "num_turns": 7,
    "num_evaluated": 7
  },
  "error": null,
  "is_not_applicable": false
```

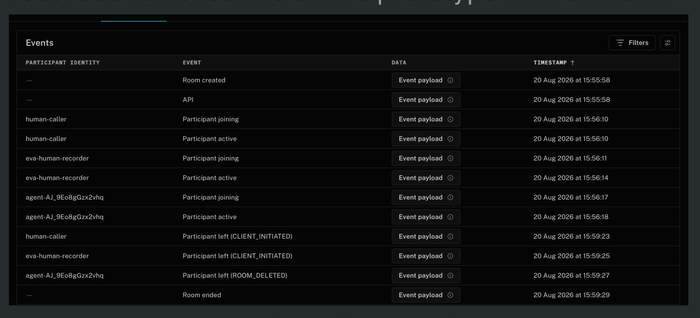

EXPLANATION

Every call produces a structured, per-turn breakdown - a score and a written reason for each metric, not a single opaque number.

Results



LiveKit Sessions dashboard, showing an overview of recorded call sessions on the prototype environment.



LiveKit session event log, showing the participant join/active/leave timeline for a single voice call.

TEAM MORIMORI
Minh Anh Nguyen
Ngoc Dung Vo

ACADEMIC SUPERVISOR
Dr. Thao Nguyen Hieu
RMIT University Vietnam

INDUSTRY SUPERVISOR
Mr. Viet Ha
Co-Founder | CTO @ Dewly.AI

SHOWCASE
SSET Capstone 2026
17 September 2026