Conversational AI without the Cloud: A Lightweight, Local Dialogue Pipeline for Non-commercial Social Robots

Lewis Watson Emilia Sobolewska Carl Strathearn
Mayuko Morgan Yanchao Yu

Edinburgh Napier University, Scotland

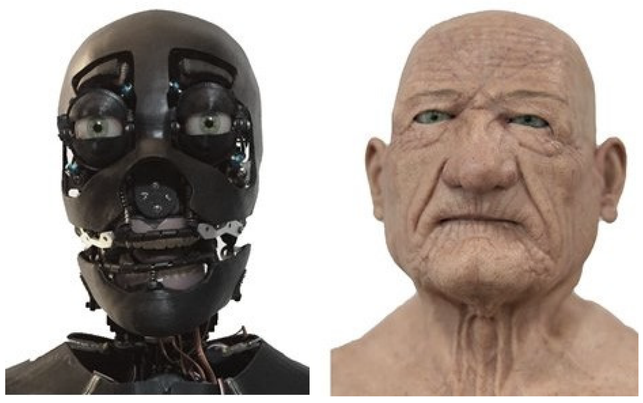
Why Local Conversational Robotics?

Cloud-hosted dialogue stacks are common in social robotics, but they are fragile in venues with poor connectivity, strict privacy requirements, or limited budgets. We built a fully local speech-to-speech pipeline for social robots so that non-commercial humanoid robots can converse without internet access, API keys, or proprietary cloud services.

- **Infrastructure independence:** All inference runs on-device.
- **Privacy by design:** all dialogue content stays local to the deployment machine.
- **Low-latency interaction:** ASR, LLM, and TTS are selected for practical offline use.
- **Modularity:** All components can be swapped out and updated to the latest open-source models without rewriting the dialogue manager.

Euclid the Social Robot

Euclid is a custom lifelike social humanoid robot with an expressive face, microphone, camera, and USB-connected Arduino / Raspberry Pi control stack. This makes it a practical test-bed for updating AI components in real time while keeping the robot body separate from the local inference machine.



Implementation Snapshot

Stage	Model / Tool	Device
VAD	RMS energy gate + Silero	CPU
ASR	Whisper V3 Large Turbo	GPU
LLM	Qwen2.5 7B via LM Studio	GPU
TTS	Kokoro-82M (ONNX)	CPU

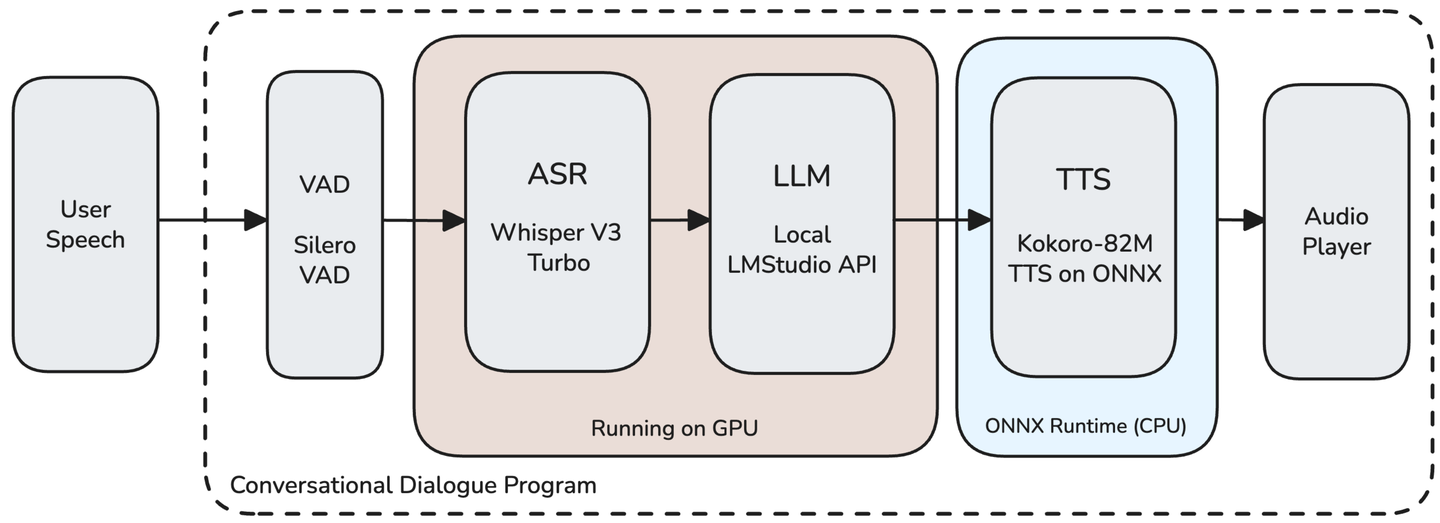
The reference deployment used an NVIDIA RTX 2000 Ada laptop GPU with 8 GB VRAM. Sentence-level TTS queuing reduces time-to-first-audio by letting Euclid speak before the full audio response has been synthesised. Preliminary tests with Qwen3 4B show comparable performance under the same hardware constraints.

Qualitative Field Observations

- **Fragile turn-taking:** When users hesitated or paused mid-thought, Euclid would sometimes treat the pause as the end of a turn and start speaking early.
- **Noisy venues caused false starts:** The always-on microphone could pick up nearby chatter or operator speech, making replies seem directed at the wrong person.
- **Persona shaped engagement:** Many visitors responded positively to Euclid's playful "rusty circuits" character and banter, which made the interaction feel more lifelike.
- **Expectations varied by audience:** Some visitors expected factual, "ChatGPT-style" answers, while academic audiences were more likely to discuss the robot with operators than address Euclid directly.

Offline Speech-to-Speech Architecture

The pipeline follows a fully local sequence: audio capture → voice activity detection → speech recognition → language model response generation → text-to-speech → queued audio playback.



- **Input gating:** RMS energy filtering plus Silero VAD reduces non-speech triggers before transcription.
- **ASR:** Whisper V3 Large Turbo provides the speech-to-text front end on GPU.
- **Language Model Backend:** A local OpenAI-compatible LM Studio API hosts a 4-bit Qwen2.5 7B model.
- **TTS:** Kokoro-82M runs on CPU via ONNX Runtime, freeing the GPU for ASR and LLM inference.

Core Contributions

- A lightweight offline dialogue pipeline for social robots.
- Real-world testing through deployments at four public events using a single laptop.
- Analysis of 5,161 dialogue turns to expose interaction patterns and failure modes.

Field Deployments and Dialogue Data

After filtering empty or one-sided dialogues, the dataset contained 5,161 retained dialogue turns.

Venue	Turns
AI for Good (UN Geneva)	2,723
New Scientist Live	1,894
UK Robotics Expo	428
Edinburgh Fringe	116
Total	**5,161**

Conversations ranged from short greetings to longer exchanges about robotics, AI, and Euclid's *"old robot with rusty circuits"* persona. The system prompt instructed Euclid to be helpful, family-friendly, kind, chatty, and wise.

Interaction Findings

Measure	User	Euclid
Positive sentiment	79%	75%
Assigned BERTopic cluster	72%	66%

Quantitative analysis showed predominantly positive interactions; field observations further showed that persona, audience expectations, and venue context strongly shaped how people engaged with the robot.

Field Deployments and Ethics

- **Four public deployments:** AI for Good (UN Geneva), New Scientist Live, UK Robotics Expo, and Edinburgh Fringe.
- **Mixed audiences:** Adults, children (with guardians), and STEM communities, often in noisy, multi-speaker settings.
- **Privacy safeguards:** Participants were informed on-site that Euclid's responses were generated autonomously; only anonymised text-only dialogue logs were retained, with no audio, video, or demographic data stored.

Remaining System Limitations

- **Weak grounding:** Replies occasionally missed immediate environmental context.
- **Verbosity:** Responses could become over-long and lecture-like.
- **Context leakage:** Prompt states needed reliable resets between events.

Next Steps

- Add interruption-aware turn-taking and optional push-to-talk control.
- Distil more concise responses from curated interaction data.
- Introduce lightweight visual grounding (presence/gaze cues).
- Stream LLM tokens directly to TTS to minimise time-to-first-audio further.
- Compare local and cloud pipelines in controlled user studies (trust, warmth, perceived intelligence).

Takeaway

A fully local conversational robot is practical on accessible hardware: Euclid sustained real public interaction using a single 8 GB laptop GPU. Natural turn-taking and environmental grounding remain the main barriers to fluent dialogue.

doi.org/10.1145/3776734.3794421 HRI Companion '26, March 16-19, 2026, Edinburgh, Scotland, UK
L.Watson@napier.ac.uk