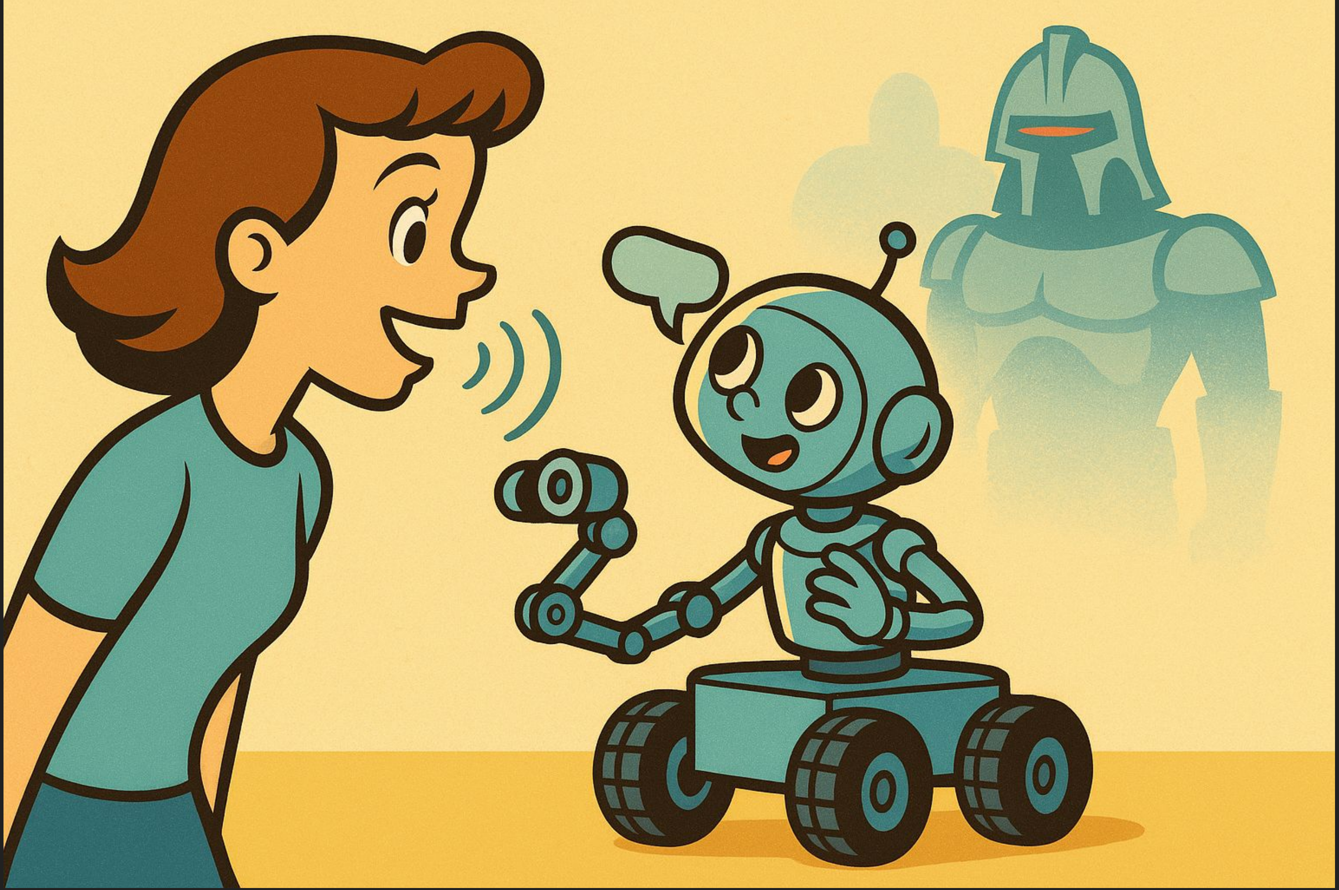


By Your Command

BY YOUR COMMAND

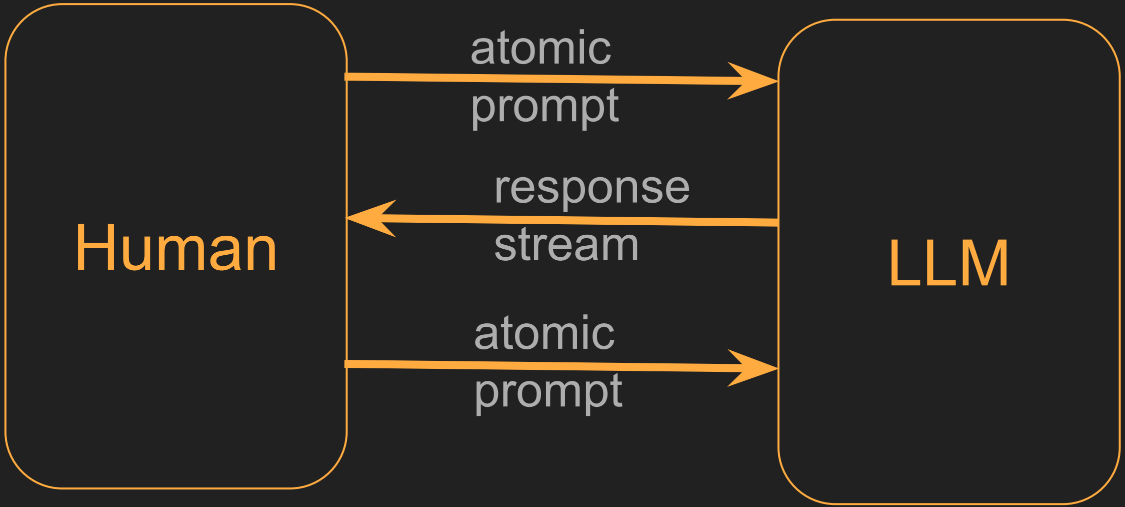
User Story

1. I want to have ***fluid*** voice conversations with my robot, where it responds like a *frontier AI model*.
2. I want my robot to understand objectives, and to carry them out.
3. Reach goal: I want my robot to progressively develop a semantic understanding of its environment as it explores, so that we can coordinate objectives with that common understanding.
4. I want these capabilities at a reasonable cost. $$$

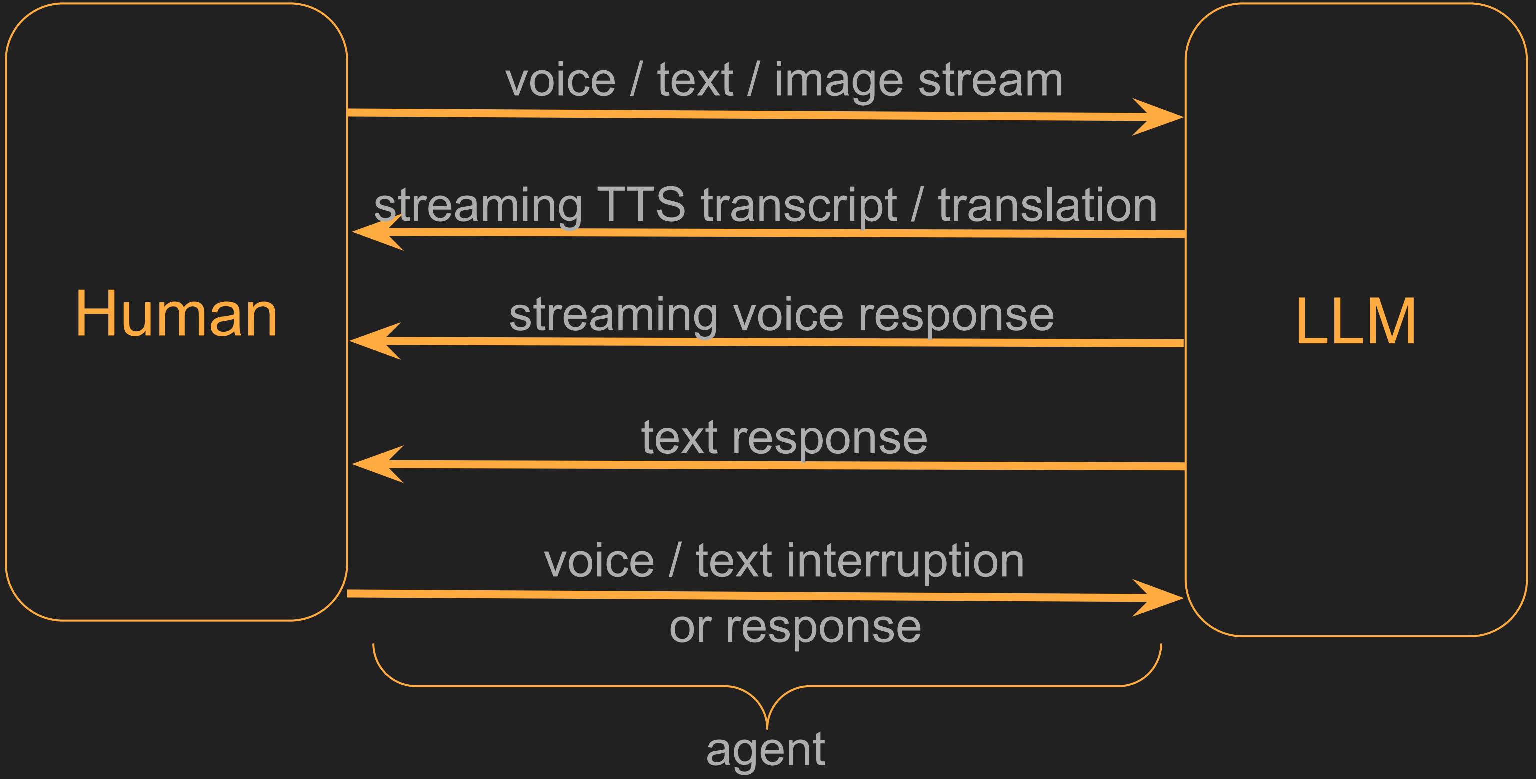
Fluid Voice - People are Picky

- So damn particular - an extra half second pause causes discomfort and uncanny valley feelings
  - People listening continuously predict/update the end-of-turn
  - Results in ~200ms turn around time
- Stupid human tricks - can use motion to buy time - implies "i'm thinking"
- Ultimately this necessitates streaming to lower response latency
- Work is ongoing on semantically predicting end-of-turn for minimum latency
- Invest in hardware echo cancellation
- Robots don't know how to shut up!

Normal Turn-Based LLM Chat



Voice Driven OpenAI RealTime / Gemini Live LLM APIs



Human

LLM

voice / text / image stream

streaming TTS transcript / translation

streaming voice response

text response

voice / text interruption or response

agent

Conversations vs Commands

Conversations

- Fluid / Languid
- Persistent
- Descriptive
- Dynamic
- Casual
- Verbal

Commands

- Abrupt / Occasional
- Short
- Specific
- Validated
- Structured Text
- Confirmable

AI's struggle with bifurcated responses

- Differing modes - command vs conversational
- If / Then style conditionals are not reliably followed
- Can't silence or respond differently on a specific type of output (voice vs text responses)
- Gemini can offer alternate responses (A/B), but they are variations and aren't separately steerable
- Use multiple agents to get around obstacles - parallel voice to agents with different instructions (expensive)

- Same agent, 2 instances, different system prompts
- Shared voice & video input
- Shared context history & reinjection

Conversation Agent

- Everyday chat
- LLM things
- Recapitulate commands
- Scene description
- Voice out

Command Agent

- Arm presets
- Pan camera
- Move Robot
- Identify & locate objects in scene
- Sleep / Wake
- Command out

Me: Conversations, Commands, Behaviors & Plans Not Control:

VLM - Vision Language Model (visual understanding + language)

VLA - Vision Language Action Models (high hz vision + language -> fine motor control, with explicit fine tuning, arm-based manipulation)

- OpenVLA
- πo (PI Zero)
- RDT-1B
- Gemini Robotics (Google) - private
- SmolVLA - mini model - LeRobot

*generally no voice front-end, so far

autonomous, 1x speed

Pan Handle?

Setting the bar much lower now …

Let’s see if we can get the arm to move…

Complications in Human Spaces

- Agents on their own likely don’t know when they aren’t being addressed*
- This means they’ll respond to anything / everything
- Address this with wake words: “Hey Siri”
- Auto-sleep (time-out) is not always appropriate
- Explicit sleep commands are better
- Remote wake is important
- Behavior wake is a good idea (classic visual presence detection, etc)

*Gemini 2.5 has “proactive audio”

Now what?

Good night!

OpenAI's Realtime API is Pricey

- You get to pre-pay for tokens up front
- Gets you access to the OpenAI Playground
- They charge an exponential premium for long (>2min) sessions and for voice token buildup.
- Use good local VAD to keep "off the air" when no one is talking
- Session cycling can help with keeping costs reasonable
- Overly aggressive session cycling causes loss of accuracy **even if context reinjection is happening** - the lost voice tokens carry meaning.
- Directive/operational NL conversations with robots tend to be choppy and infrequent - not a normal conversational flow. Has implications for session management. Still need context.
- What’s more expensive? Optimizing too early.

Google Gemini Generous Free Tier + Video

Live API - Free Tier, concurrent, tokens/day, requests/day

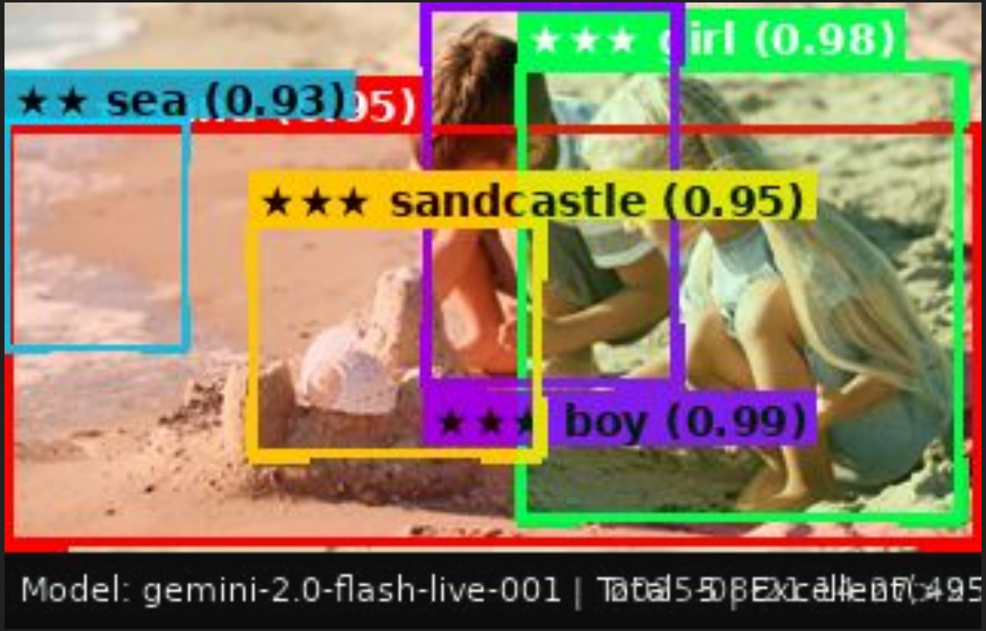
Gemini 2.5 Flash Live, 3 sessions, 1,000,000, *

Gemini 2.5 Flash Preview Native Audio Dialog, 1 session, 25,000, 5

Gemini 2.5 Flash Experimental Native Audio Thinking Dialog, 1 session, 10,000,5

Gemini 2.0 Flash Live 3 sessions, 1,000,000, *

Whatcha Lookin' At?



```
{
  "label": "boy",
  "box_2d": [
    259,
    239,
    993,
    713
  ],
  "confidence": 0.986
},
{
  "label": "shell",
  "box_2d": [
    549,
    323,
    737,
    396
  ],
  "confidence": 0.985
},
{
  "label": "girl",
  "box_2d": [
    275,
    646,
    996,
    885
```

Under the Hood

Prompt Construction

Node Graph

Project Structure

Questions

Sanity Check Time

- Are realtime LLMs necessary?
- Are frontier models necessary?
- Local compute and small model effectiveness go brrrr.
- Orchestration progressively subsumed by frontier model progression.
- Security Robots - do we really want to put them in contention with trespassers?
- Human-in-the-loop

Resources

by_your_command on github - Start with README.md Also check the specs folder for PRDs, research and analysis. This package is early-experimental.

My notes:

Multimodal LLMs for Robotics	ROS2 Learning Links
Insightful AI Talks	Bridging ROS2 with AI Models
Coding has Shifted with AI	

Robot: overview, hardware (chassis, arm, camera), software(chassis, arm - my fork, camera)