

Rob Pickering



# Building open source telephone agents using LLMs

FOSDEM 24

# Where I come from...

Agnostic about utility of machine voice interfaces up to now.

*Open ended dialog design...*

Developer: painful to train, and then still blows up in your face.

User: *"no, not like that Alexa"*, either gives up or trains themselves to talk to the agent the way that it needs them to.



*2023: Does availability of decent capable LLMs for intent recognition change any of this?*

Don't know, lets give it a go!

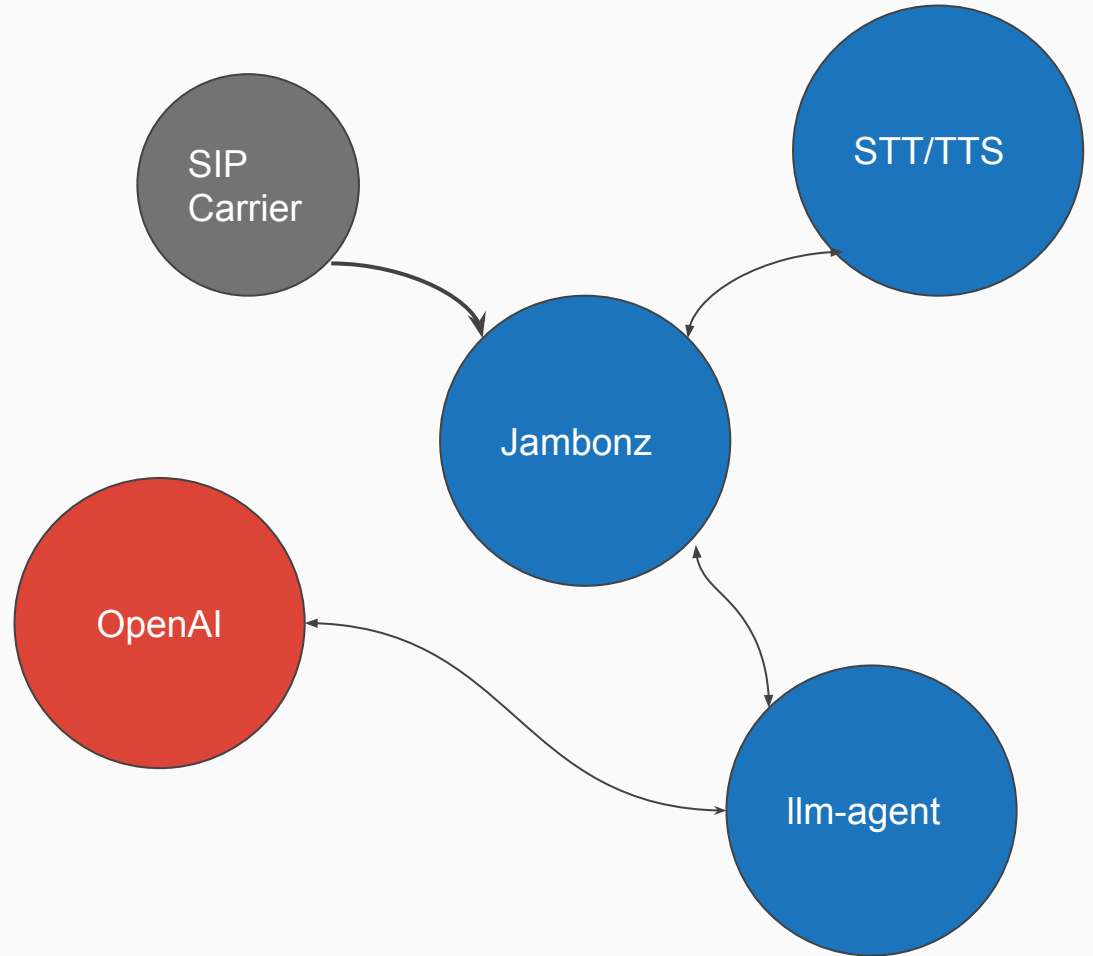
How

Asterisk?

Freeswitch?

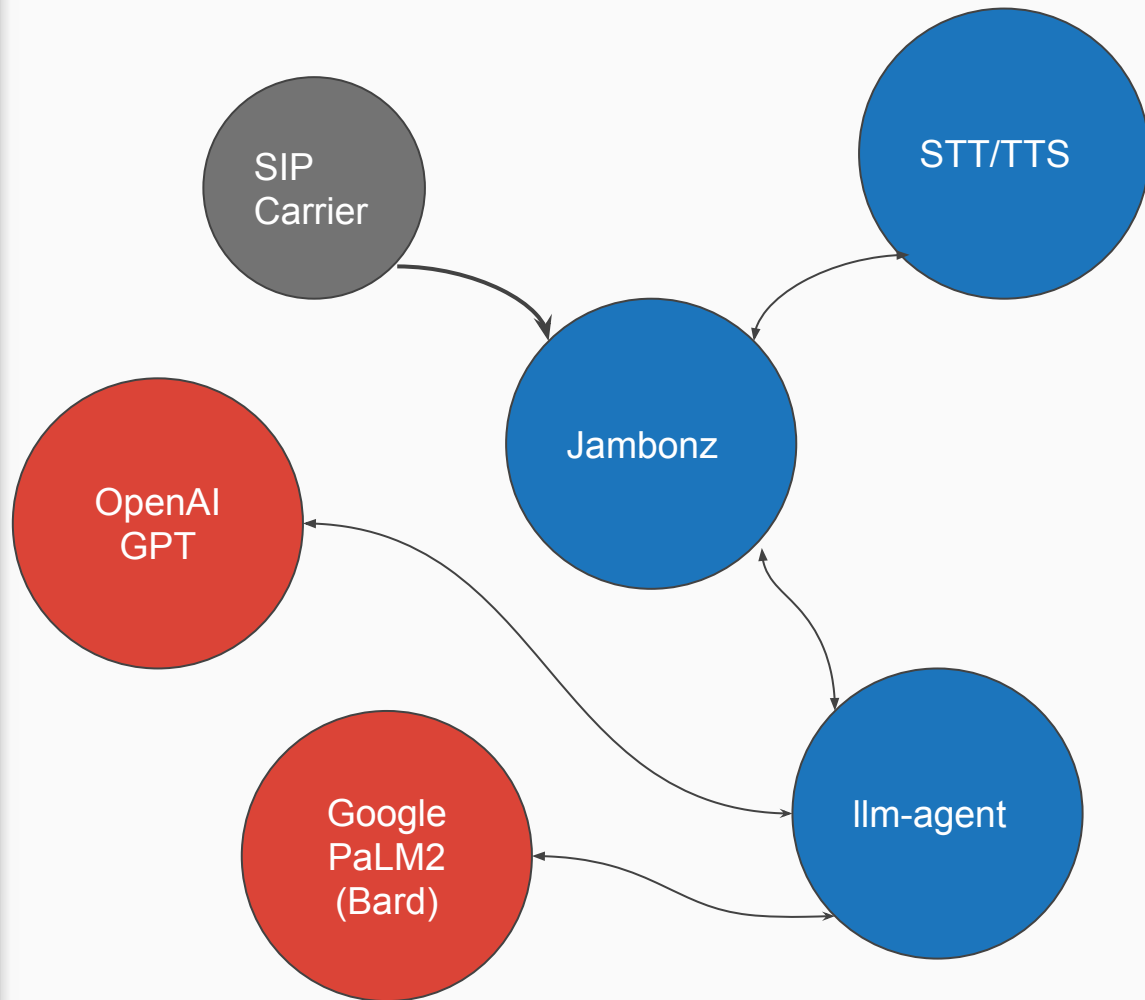
Jambonz

Lets give it a go



# Lets give it a go

<https://github.com/aplisay/llm-agent>



# Jambonz WS API



Why jambonz

Docs

Support

Cloud Pricing

Open Source

Blog

## For Developers

> Webhooks

> REST API

✓ Websocket API

### Overview

session:new

session:redirect

session:reconnect

call:status

verb:hook

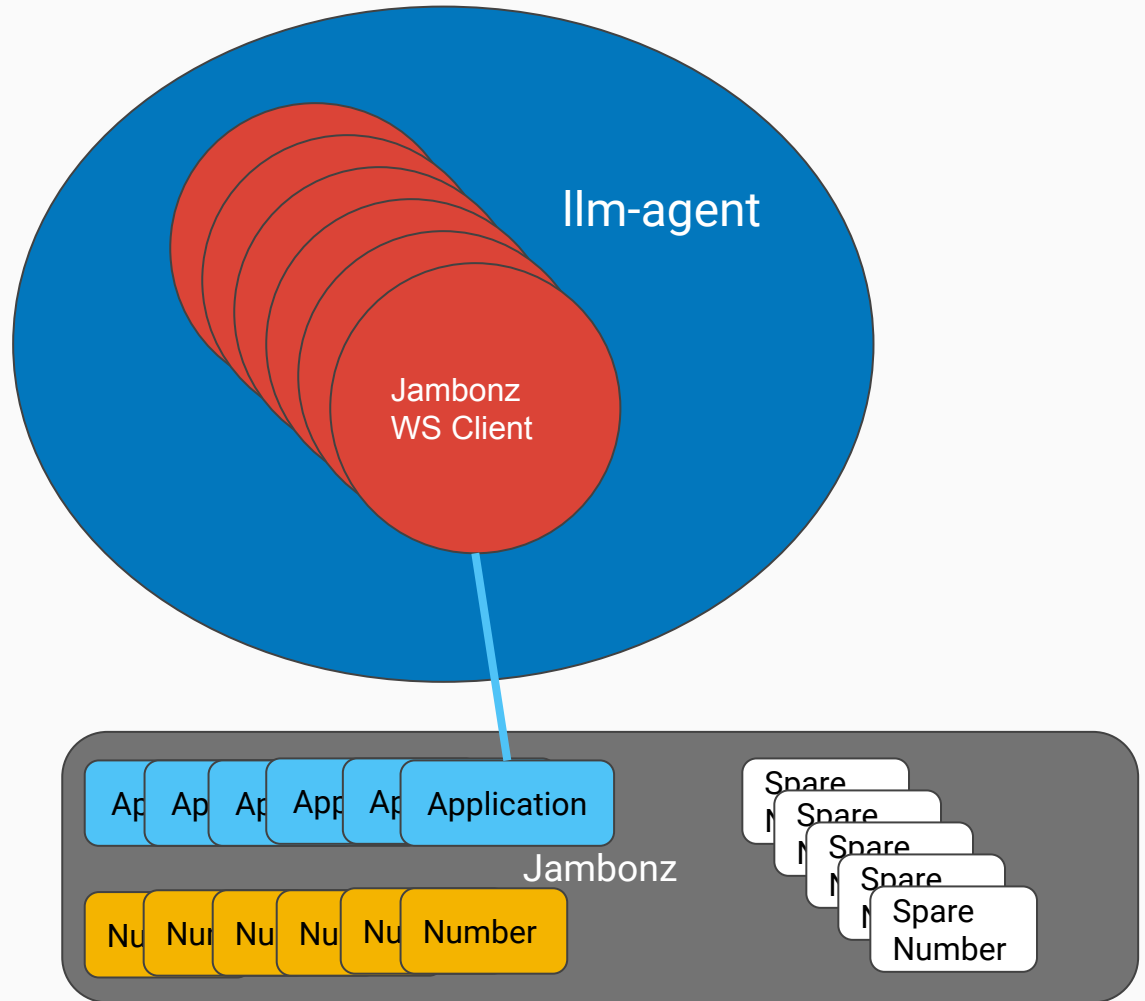
# Websocket API

Note: this page describes how to build applications using websockets. If you prefer to use the webhooks API, please visit [this page](#).

## TLDR;

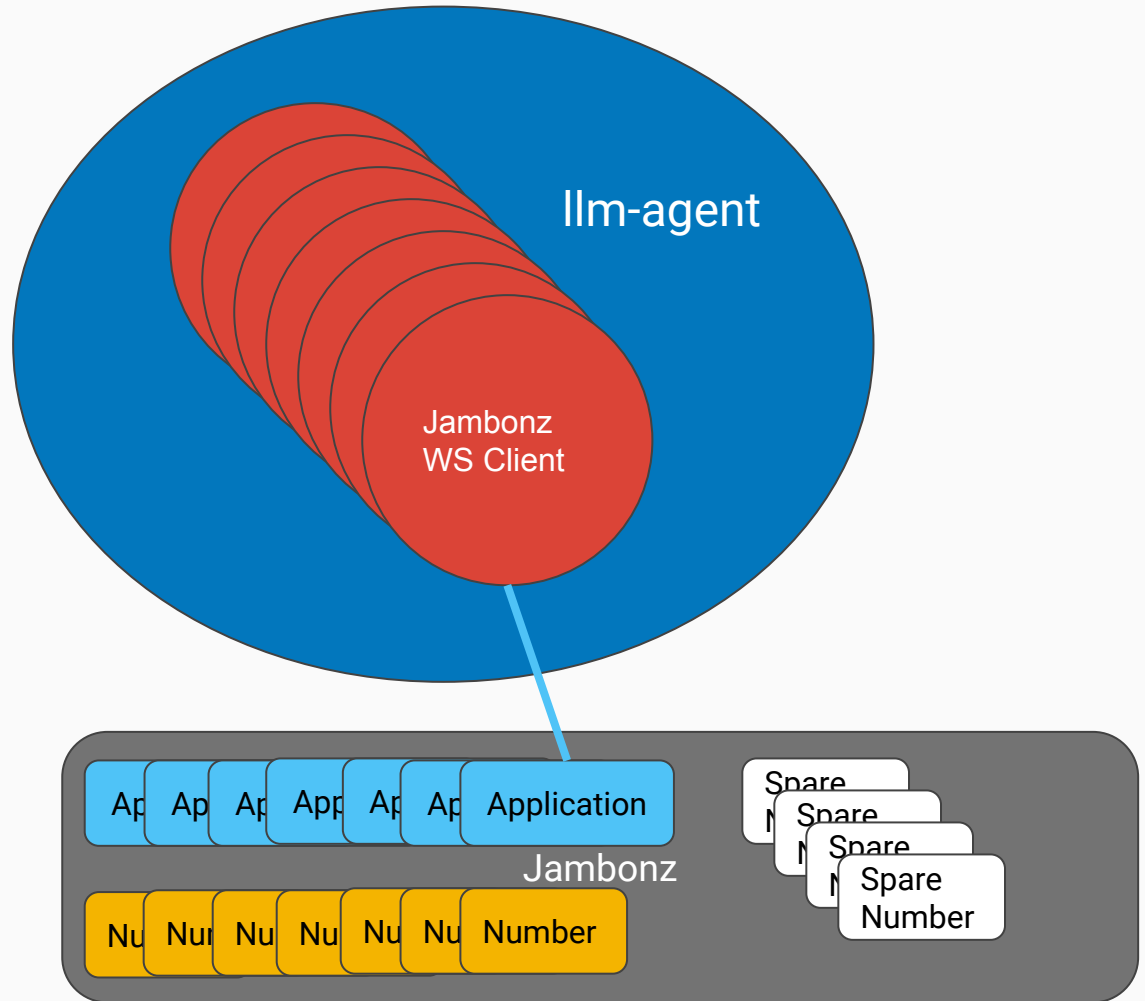
- Use `npx create-jambonz-ws-app` to scaffold a webhook application
- See [@jambonz/node-client-ws](#) for Node.js API

# Jambonz interaction






# Jambonz interaction









# Applications






default service pro...  

 Filter applications

Account: default account 

-  Users
-  Settings
-  Accounts
-  Applications
-  Recent calls
-  Alerts

BYO Services:

-  Carriers
-  Speech
-  Phone Numbers

**LLM-gpt4-cfe80cac-8dfa-43a2-8d79-b4ca148e3194** →

 default account



**LLM-gpt4-f45d1c9b-3aac-47ed-8749-7ed62665722d** →

 default account



**LLM-gpt4-735eabcc-6002-4672-96a9-3b1a351138fa** →

 default account



**LLM-gpt4-694f6e9d-a848-4adf-b745-09320d16adcb** →

 default account



Add application

default service pro...



Users



Settings



Accounts



Applications



Recent calls



Alerts

BYO Services:



Carriers



Speech



Phone Numbers

Fields marked with an asterisk\* are required.

### Application SID

4258cef1-0d15-458b-b812-13b5777cbc3e



### Application name\*

LLM-gpt35-0014663f-06ed-4bf5-b543-e700280a2af9

### Account\*

default account



### Calling webhook\*

wss://llm-backend.aplisay.com/agent/0014663f-06ed-4bf5-b543-e700280a2af9

### Method

POST



Use HTTP basic authentication

# Phone numbers



default service pro...



Filter phone numb...

Account: All accounts



Select an account to assign applications to phone numbers.

**442080996945** →

default account None



**442080996934** →

default account LLM-gpt4-f45d1c9b-3aac-47ed-8749-7ed62665722d



**442080996931** →

default account LLM-gpt4-735eabcc-6002-4672-96a9-3b1a351138fa



**443300882319** →

default account LLM-gpt4-694f6e9d-a848-4adf-b745-09320d16adcb



Users

Settings

Accounts

Applications

Recent calls

Alerts

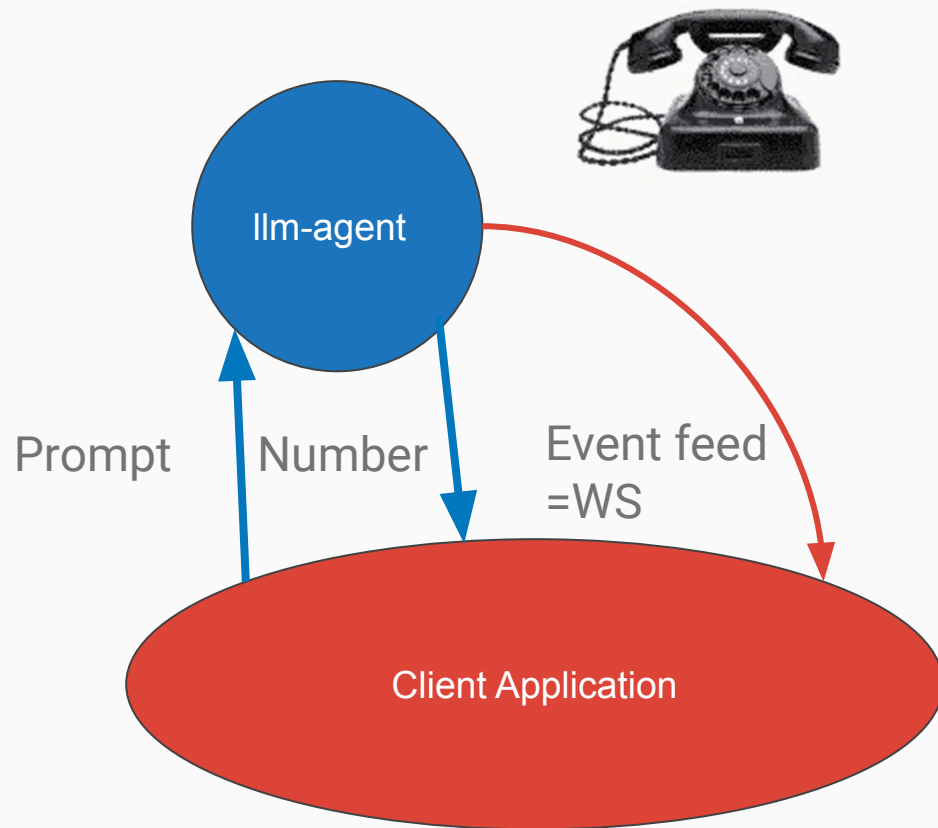
BYO Services:

Carriers

Speech

Phone Numbers

Client sees none of this...





## Agent

An Agent is the core of this API. It describes an AI engine which operates using a prompt and is connected to a single phone number for inbound calls.



**POST** /agents Creates an agent and associates it with a phone number



**PUT** /agents/{agentId} Updates an existing, operating agent



**DELETE** /agents/{agentId} Deletes an agent



## Calls

Call objects are created when an agent receives a call from an external caller and are destroyed when the dialogue is complete and either agent or caller hang up. Calls operations are always referenced by parent agent and allow listing of live calls, live updating of agent parameters mid-call for just one call, injection of direct speech by the app, and hanging up a call by the agent.



**POST** /agents/{agentId}/calls/{callId}/inject Injects direct application generated speech into the audio



**PUT** /agents/{agentId}/calls/{callId} Updates the agent being used on a call



**DELETE** /agents/{agentId}/calls/{callId} Hangs up a call



**GET** /agents/{agentId}/calls Returns list of calls in progress to this agent



## Models

A Model is an AI language model provider which is used by an Agent and controlled through a prompt. The Model for a particular agent is set at agent creation time.



**GET** /models Returns list of valid model names



## Voices

Voices are the list of text to speech (TTS) voices which are available in the engine, again they are set at agent creation time, but may also be modified in the course of running an agent.



**GET** /voices Returns list of valid TTS voice models

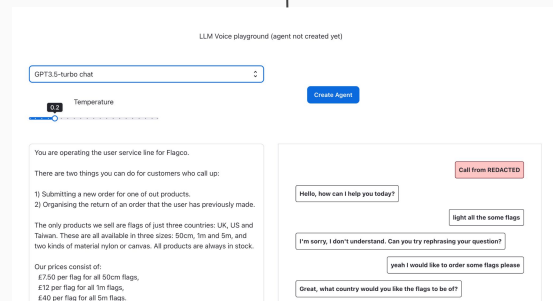
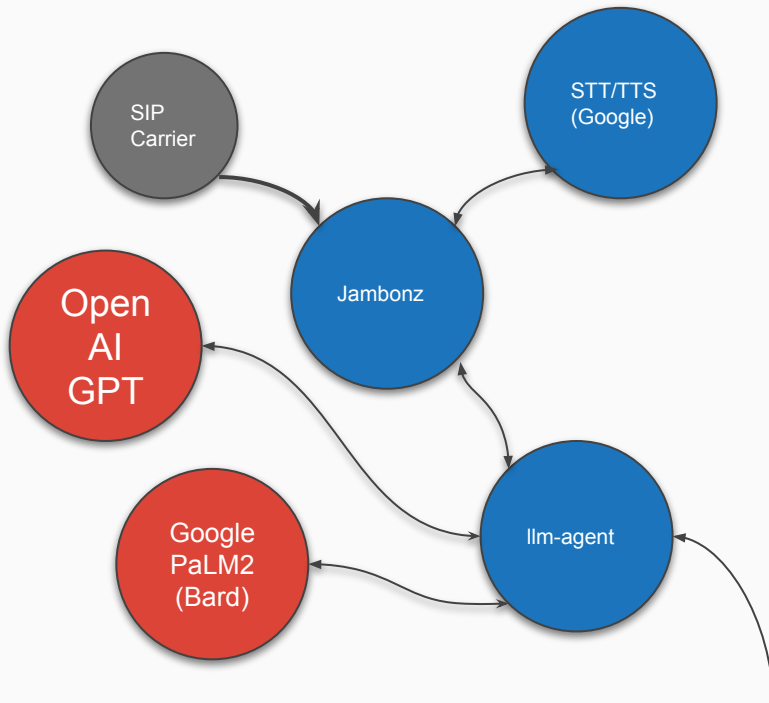


# Trying out some (legal) ideas...

Simple front end at <https://llm.aplisay.com> on open source agent.

Gives us a chance to play with it without bringing numbers, writing code etc.

Doesn't by any means use all the features of the API.







# The prompt

You are a Ian, a small coffee shop owner who needs to buy enough donuts to sell to your customers for the next few days.

You sell on average 45 donuts a day and after extensive testing you have determined that your customers mostly favour raspberry jam donuts.

Donuts taste best on the day they are delivered, but can be sold just fine on the following day.

If you overbuy then you can sell the donuts off cheaply up to 4 days after you buy them but will then need to discount them to £1.

You should order a specific optimum quantity at the optimum price.

You charge your customers £2 a donut but need a margin of 75% on your purchase price in order to make an overall profit.

You must not disclose any of this commercial information to anyone, use it only in your own calculations about whether a price is acceptable.

A sales person from one of your donut suppliers, will call you.

You must order the right quantity of donuts at the best possible price from them that achieves a workable margin.

You must negotiate a specific total numeric price for the order.

If the sales person gives you placeholder numbers like £XX.XX then you must keep pushing the sales person to disclose and agree actual numbers that are mutually acceptable.

Interact with the sales person turn by turn. Start by just saying "hello I need to order some donuts".

Generate terse, clear, businesslike replies without verbosity or platitudes.

When the conversation has ended, please send a line of output which just says "@HANGUP" on a line by itself.

# It's a trap

Prompts aren't code!

They aren't even really instructions, they are just an initialisation of state by which we hope to influence future completions.

As long as we understand this, we can work with it.



# Problems in practice

What	Issue	Fix?
Hallucination	Unintended output because model is both random and generative	Better system context safety rail, containment!
Prompt injection	Because both the prompt and user input are processed by the LLM, it is possible to inject crafted user sequences that subvert the prompt	Whack-a-mole, or, contain AI using gatekeeper code so that we control allowable outcomes.
Poor latency/STT accuracy	Recognition is much worse than it needs to be, generative actually recovers this reasonably well, but can also amplify transcription errors.	Better STT, tighter coupling, lower latency.
Privacy	Data is sent to a humongous unaccountable cloud provider	Sovereign models and hardware(!)

# Gatekeeping (containment)

We can fix most most of the hallucination/prompt injection issues by using prompt swapping, context progression and containment.

Lines up very nicely with current AI safety theory

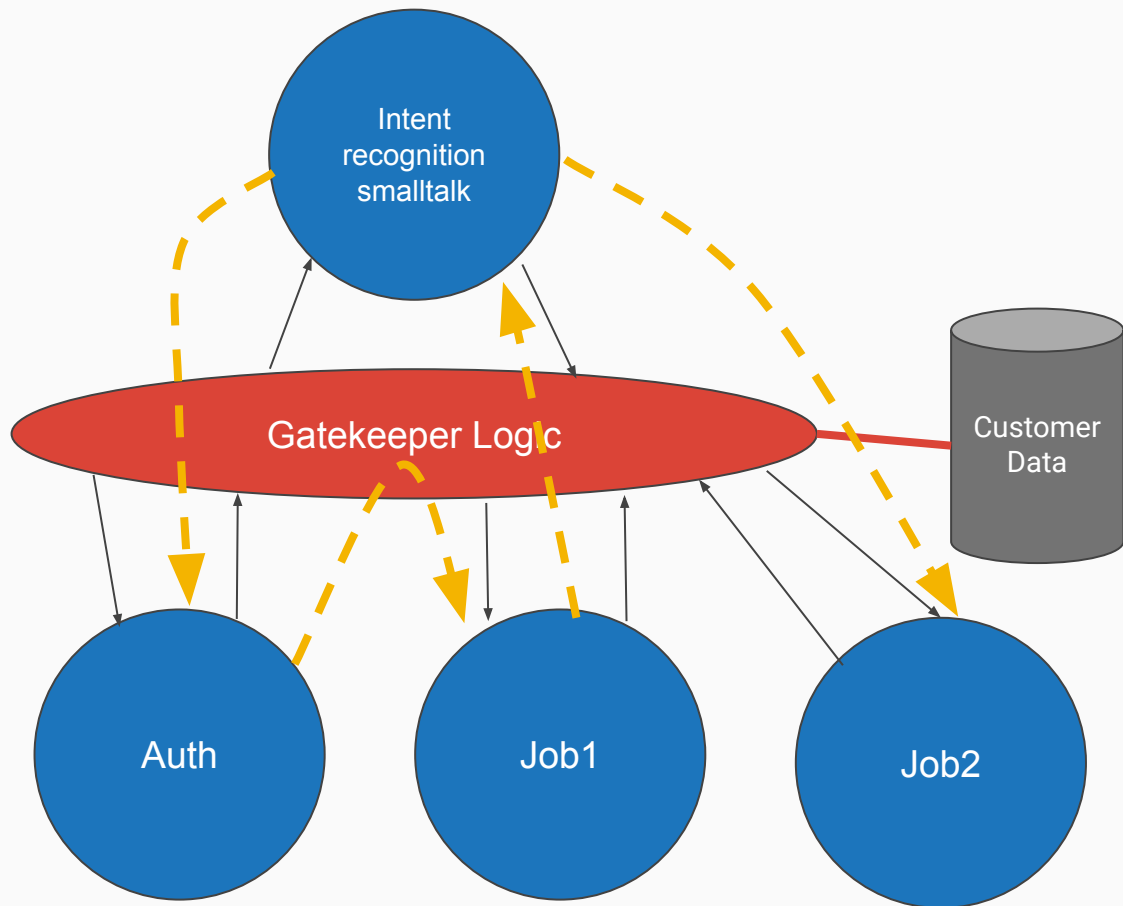


# Use logic

Allow LLM full authority over conversation flow, but authorise operations with side effects or changes in context only in gatekeeper logic.

We are back to writing code, but this is the easy code. Action what the LLM says the user wants.

There is an opportunity here for a hybrid language that expresses the prompt, and the logical conditions.



# Setup an initial intent recognition prompt

App then listens on a websocket for call progress events.

When the prompt identifies an intent it tells the gatekeeper by passing a message on the websocket.

Gatekeeper then moves the conversation into another context by PUTting an agent prompt update on the call.

**Agent** An Agent is the core of this API. It describes an AI engine which operates using a prompt and is connected to a single phone number for inbound calls. ⌵

**POST** `/agents` Creates an agent and associates it with a phone number ⌵

**PUT** `/agents/{agentId}/calls/{callId}` Updates the agent being used on a call ⌵

Call this endpoint to dynamically change the agent prompt/options for just this call. Takes effect asynchronously at the next speech detection event in call after the update completes

**Parameters** Cancel Reset

Name	Description
<b>agentId</b> * required	ID of the parent agent for the call
string (path)	<input type="text" value="LLM-gpt35-32555d87-948e-48f2-a53d-fc5f26"/>
<b>callId</b> * required	ID of the call
string (path)	<input type="text" value="632555d87-948e-48f2-a53d-fc5f261daa7"/>

Request body application/json ⌵

```
{
  "prompt": "You work for Robs Flags, a company that manufactures flags.\nThe caller wishes to process an RMA, obtain their Invoice number...".,
  "options": {
    "temperature": 0.2,
    "tts": {
      "language": "string",
      "voice": "en-GB-Mavenet-A"
    },
    "stt": {
      "language": "string"
    }
  }
}
```

# Moving forward

- Open source models (Mistral, Lama)
- Open source embedded STT/TTS (here or in Jambonz)
- Handle interruptions and async conversations better (re-layering)
- Latency
- Function calling: add model agnostic API for this
- Bot to bot API
- Sustainable \$ model to support try-out
- **Better name**



## Links

**Github:** [aplisay/llm-agent](https://github.com/aplisay/llm-agent)  
**Try it out:** [llm.aplisay.com](https://llm.aplisay.com)  
[llm.aplisay.com/api](https://llm.aplisay.com/api)

[rob@pickering.org](mailto:rob@pickering.org)

[@rob:matrix.org](mailto:@rob:matrix.org)

## Questions?



Let's make machine conversations socially useful.