



# Oltre il Chatbot

Costruisci il Tuo Assistente Vocale AI

# Who am I?



Giuseppe Spina



Let's just say...



# Let's just say...

- Start your own business





# Let's just say...

- Start your own business
- Sales increases



...but





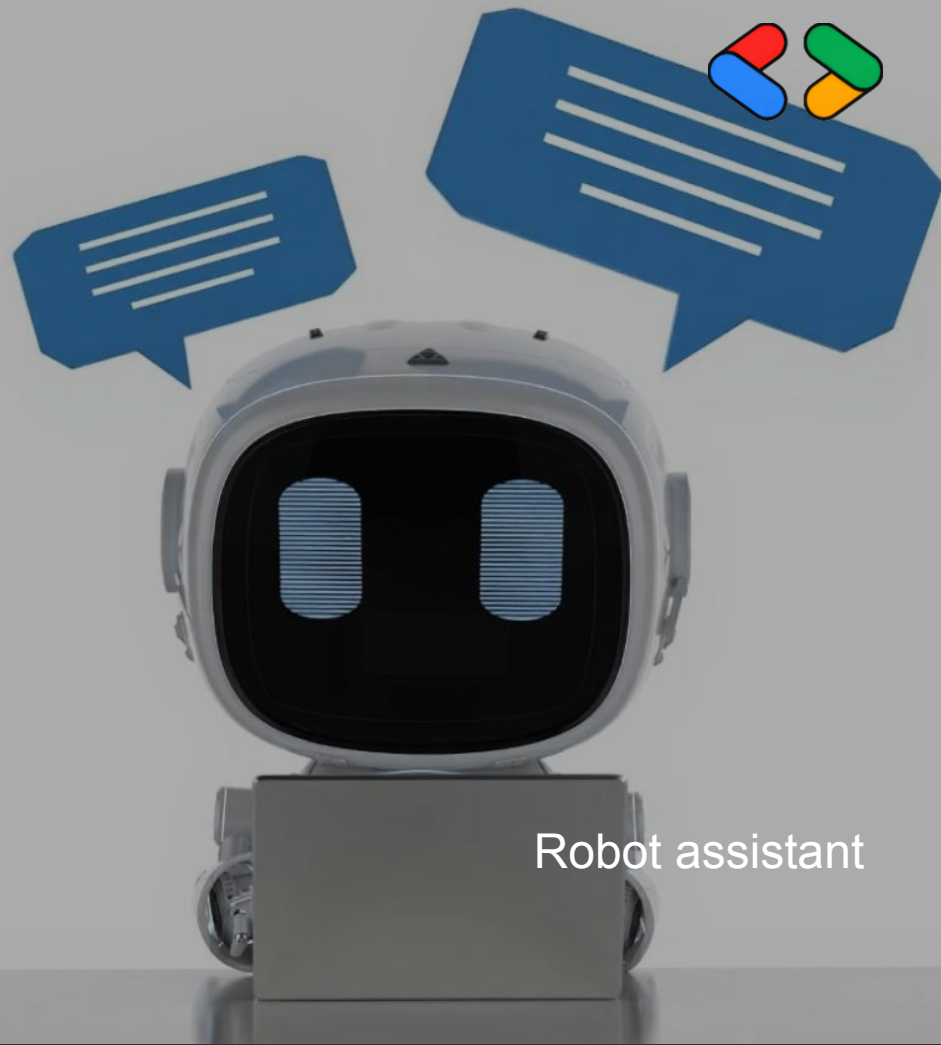
...but

- Always on the PC to solve customer problems



How to solve?

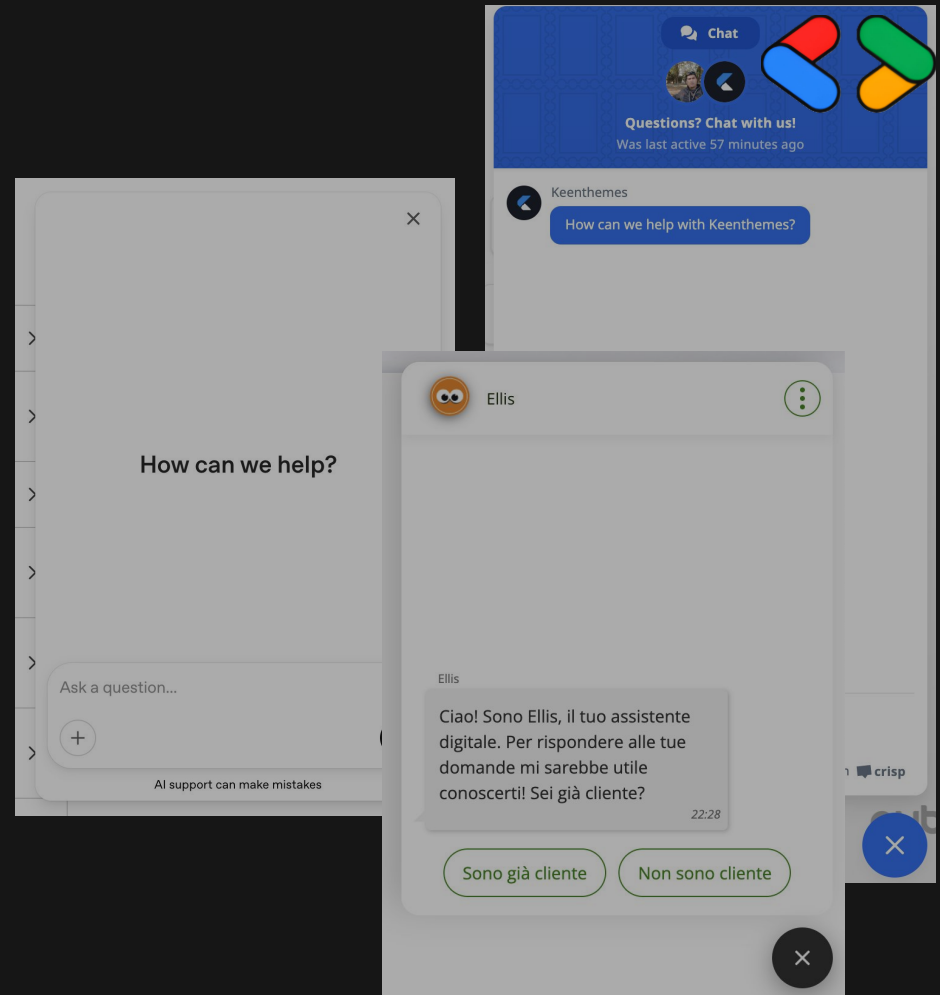
Human assistant



Robot assistant



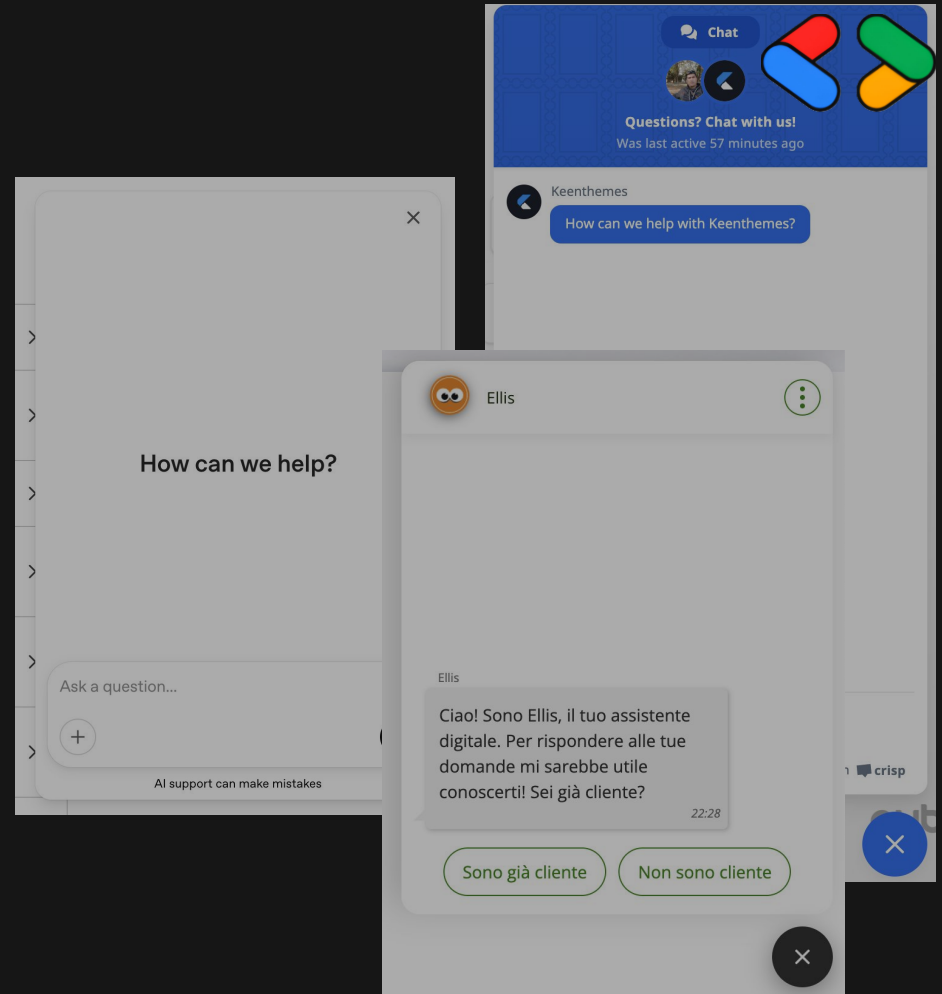
# Common answer: Chatbot



# Common answer: Chatbot

Why:

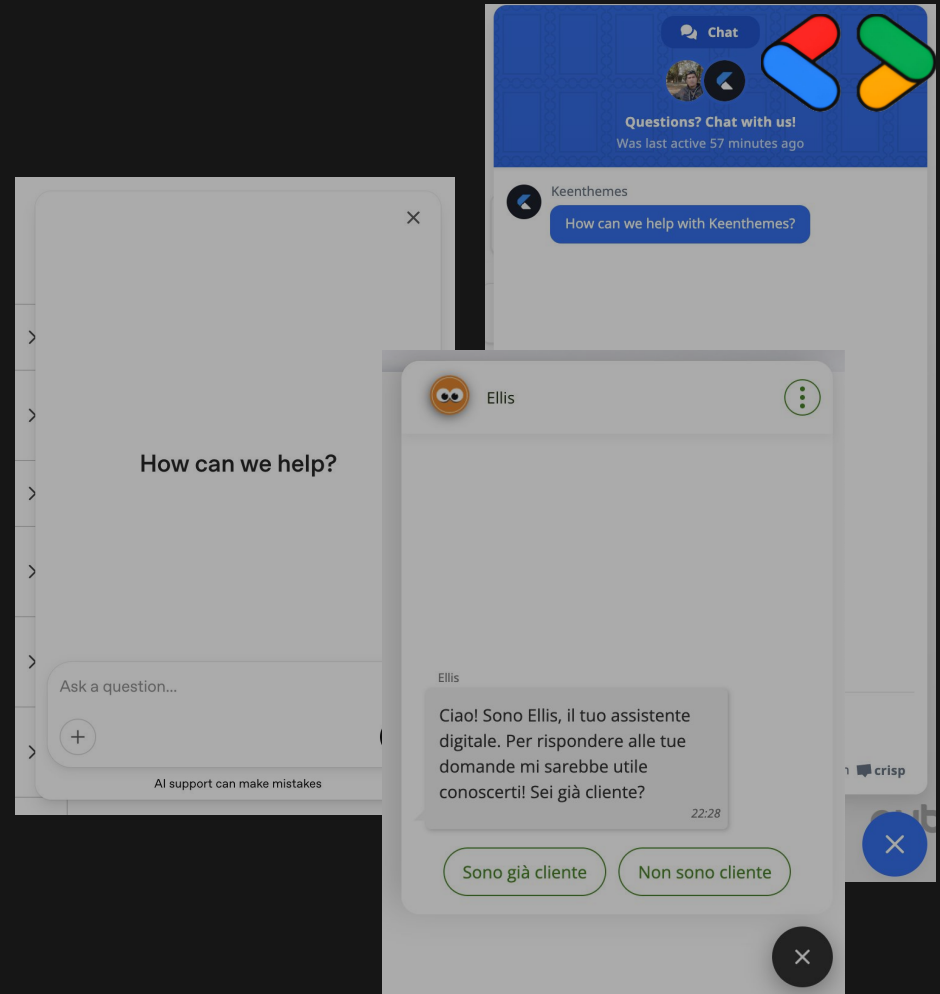
- Cost reduction (30%)



# Common answer: Chatbot

Why:

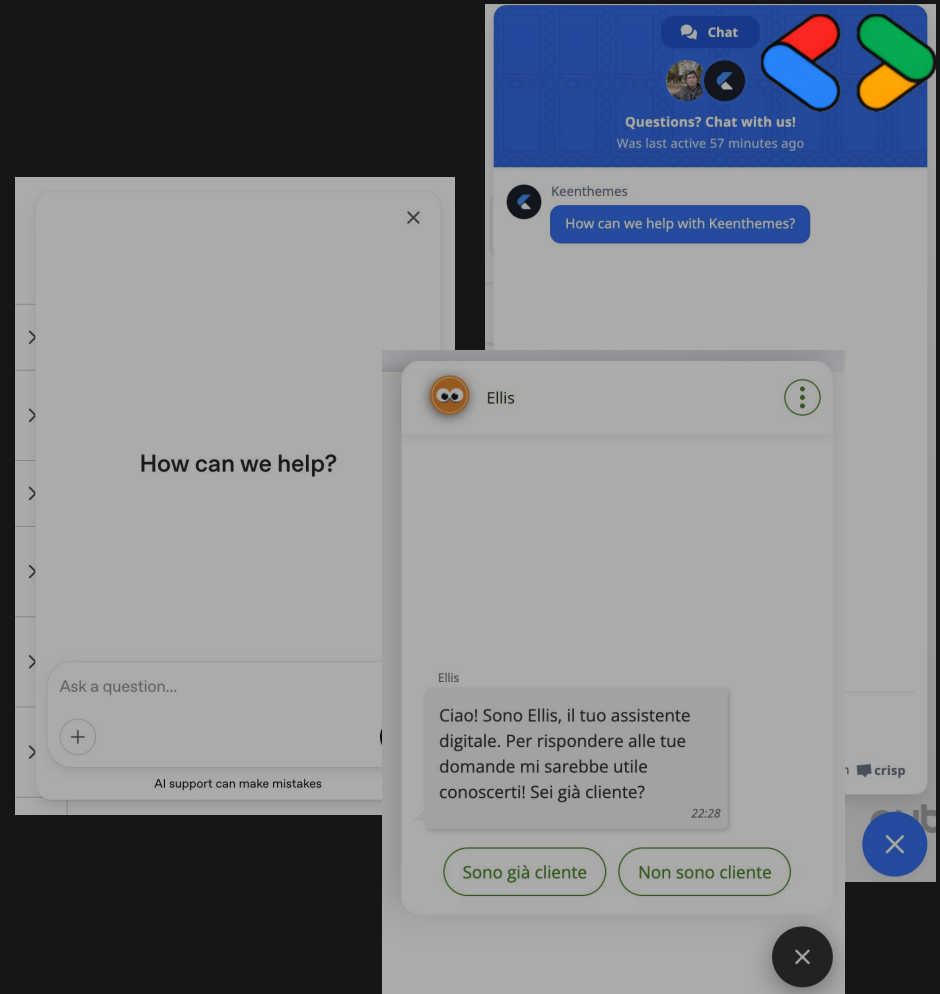
- Cost reduction (30%)
- Active 24/7



# Common answer: Chatbot

Why:

- Cost reduction (30%)
- Active 24/7
- No waiting for customers



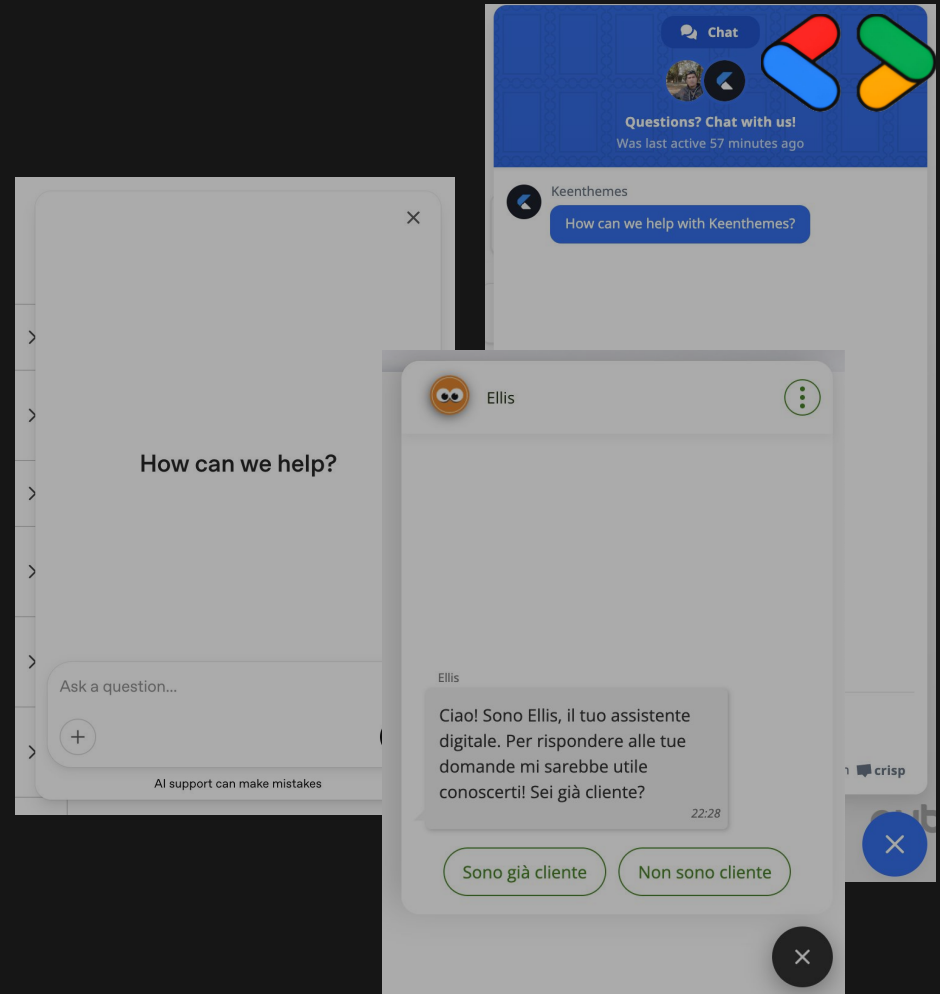


# Common answer: Chatbot

Why:

- Cost reduction (30%)
- Active 24/7
- No waiting for customers

Users use them... if they work



“Why don't you understand what I want?”



# “Why don't you understand what I want?”



Problem:

- What are you saying?



# “Why don't you understand what I want?”



Problem:

- What are you saying?
- No Context





# “Why don't you understand what I want?”



Problem:

- What are you saying?
- No Context

Over 38% of users hate them for this reason





It's time for evolution

# Voice assistant with AI





# Voice assistant with AI

- Go beyond the limits of chatbots







# Voice assistant with AI

- Go beyond the limits of chatbots
- Understands the context well





# Voice assistant with AI

- Go beyond the limits of chatbots
- Understands the context well
- More human-like interaction





What makes this magic possible?



# Anatomy of our assistant







# Anatomy of our assistant

- Brain: LLM model





# Anatomy of our assistant

- Brain: LLM model
- Voice: Websocket



# Brain - LLM model





# Brain - LLM model

- AI trained on huge amounts of data



# Brain - LLM model

- AI trained on huge amounts of data
- Understands/generates human language



# Brain - LLM model

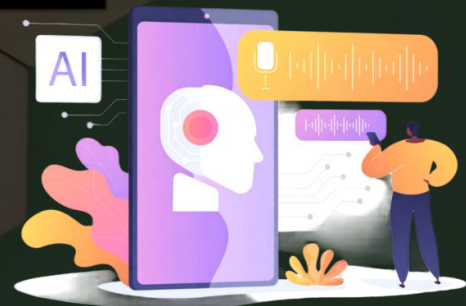
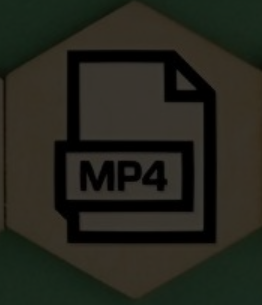
- AI trained on huge amounts of data
- Understands/generates human language

This enables the LLM to use the Context





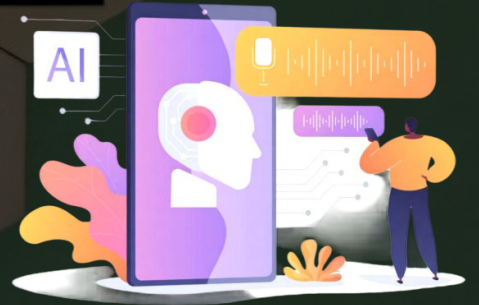
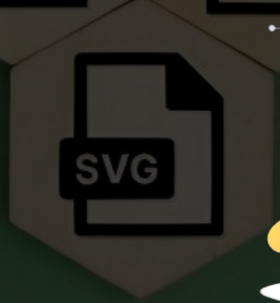
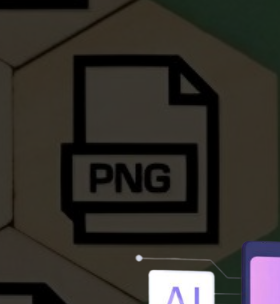
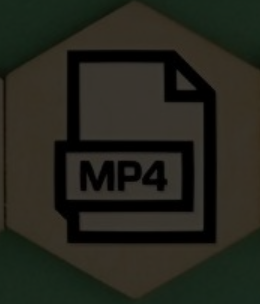
# Brain - Multimodal LLM Model





# Brain - Multimodal LLM Model

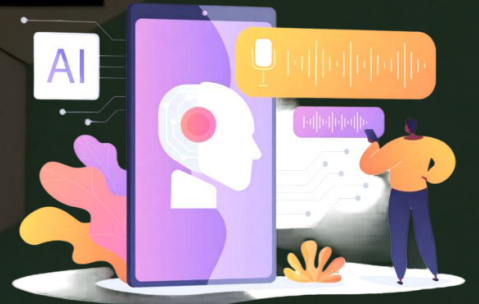
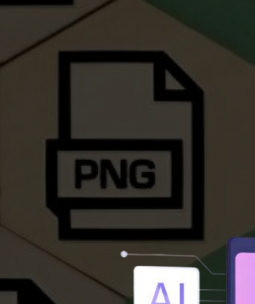
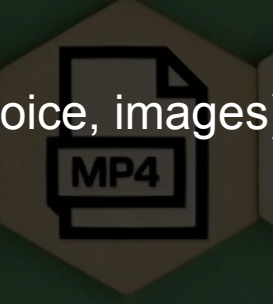
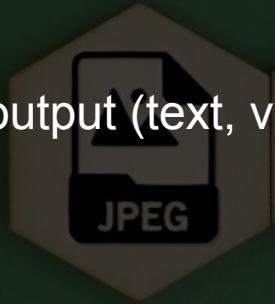
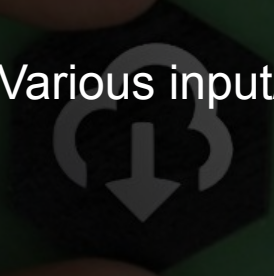
- Not just text





# Brain - Multimodal LLM Model

- Not just text
- Various input/output (text, voice, images)

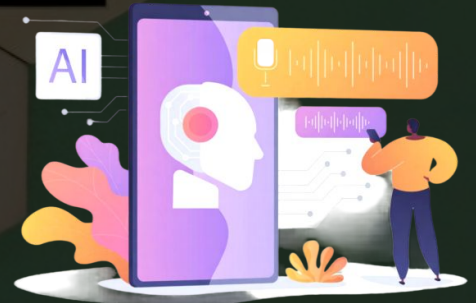




# Brain - Multimodal LLM Model

- Not just text
- Various input/output (text, voice, images)

A more human-like interaction



Brain - Gemini Live Service



✦  
Gemini 2.0  
Flash





# Brain - Gemini Live Service

Why:

- Easy integration

✦  
Gemini 2.0  
Flash





# Brain - Gemini Live Service

Why:

- Easy integration
- Real-time response

✦  
Gemini 2.0  
Flash



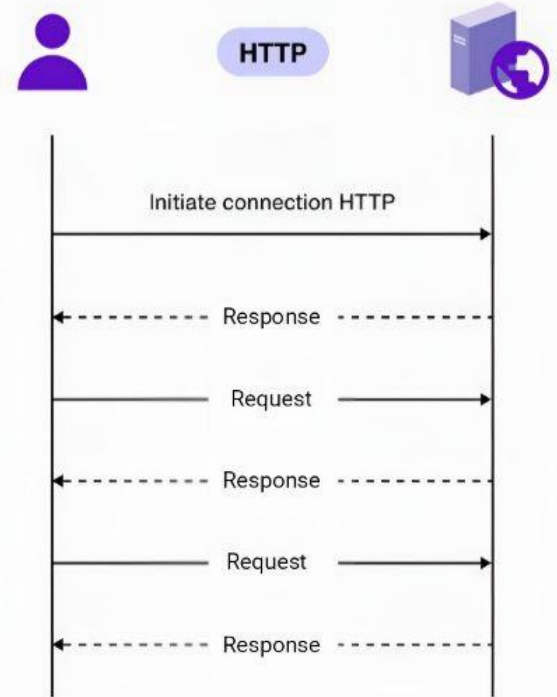
# Brain - Gemini Live Service

Why:

- Easy integration
- Real-time response
- Smoother interaction

✦  
Gemini 2.0  
Flash

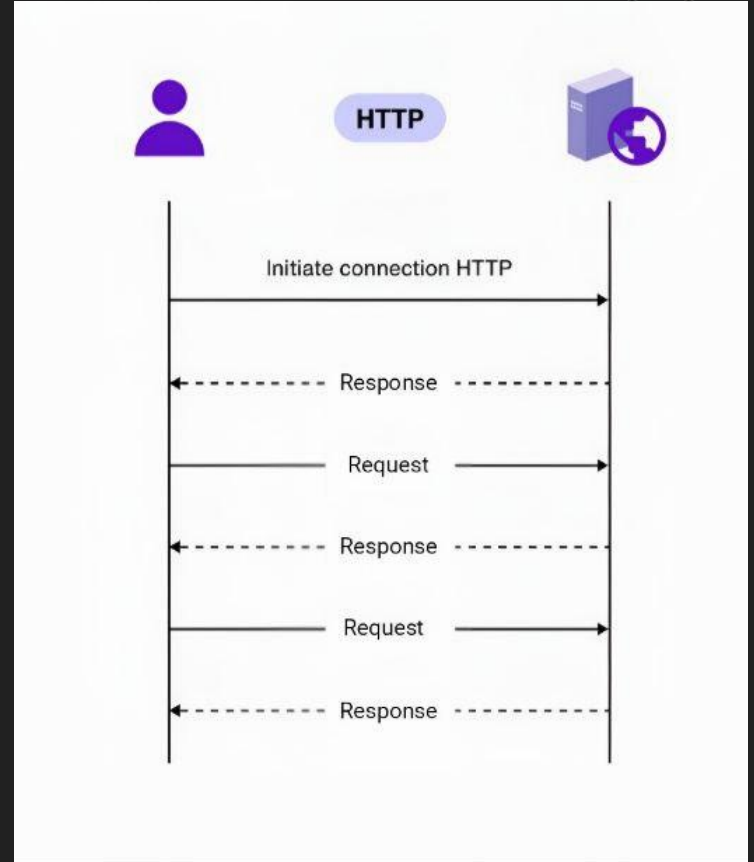
# Voice - Websocket





# Voice - Websocket

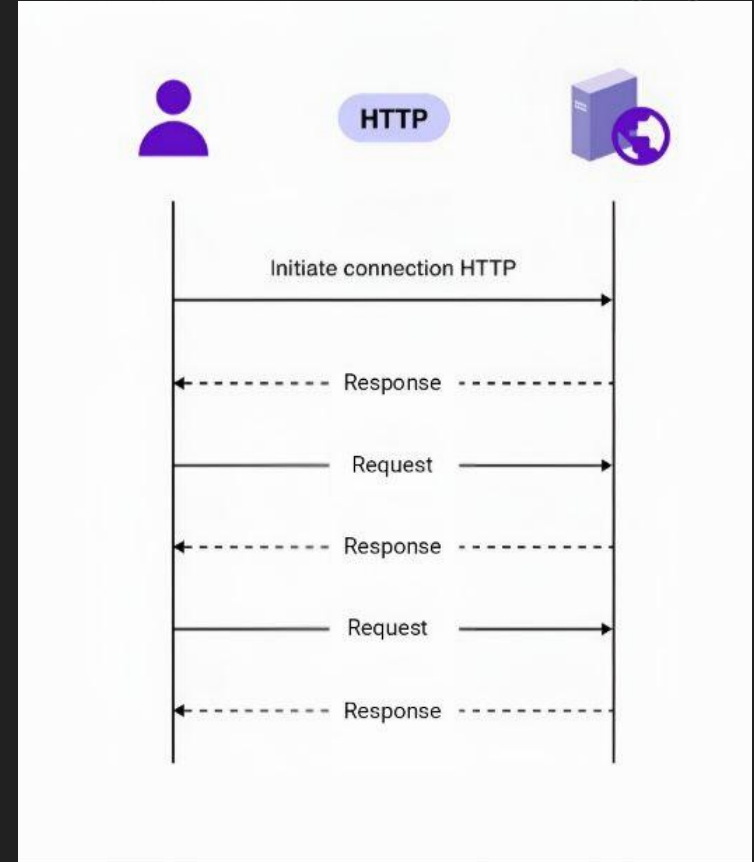
- HTTP Request





# Voice - Websocket

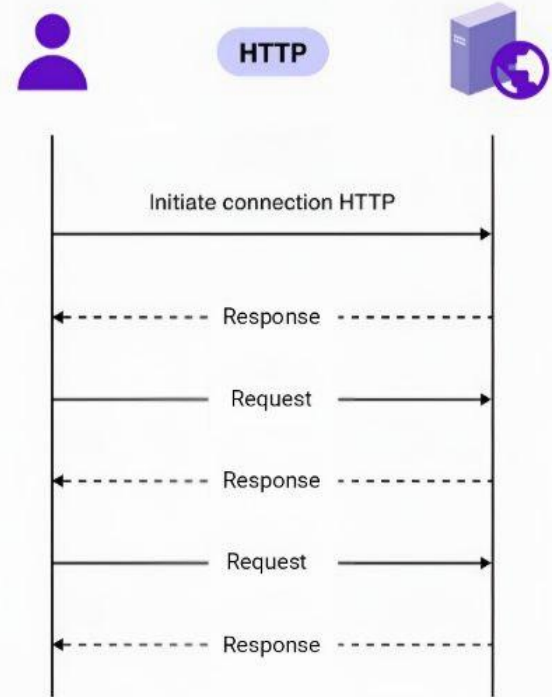
- HTTP  
Request → Response





# Voice - Websocket

- HTTP  
Request → Response ... Request

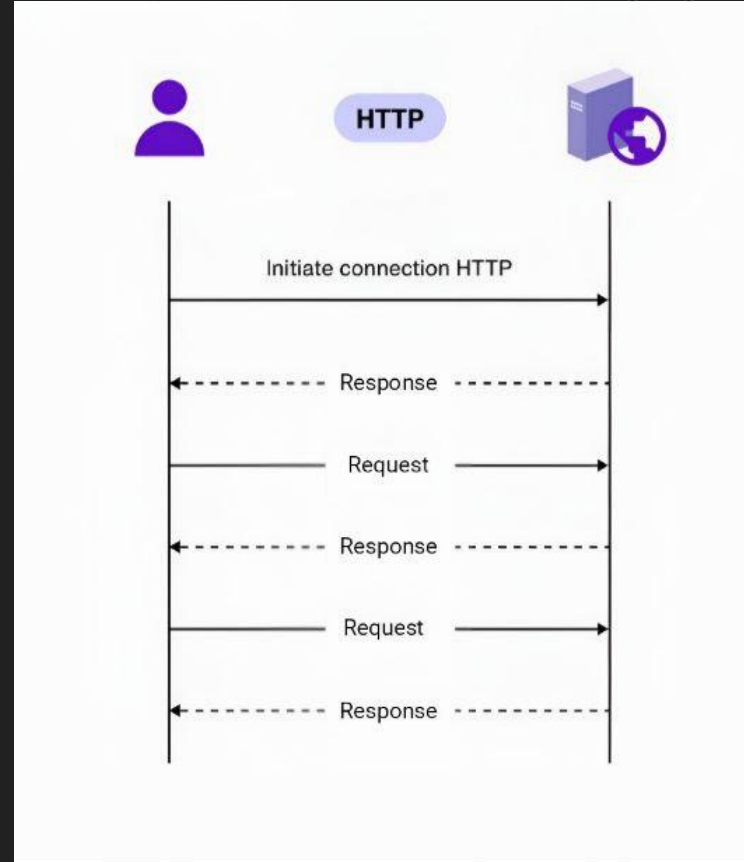






# Voice - Websocket

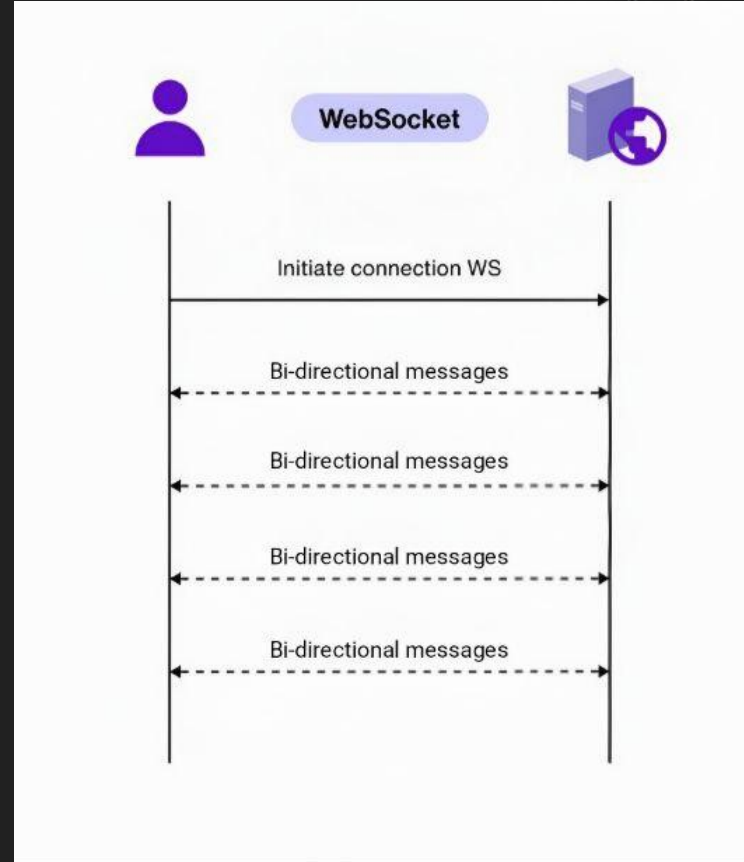
- HTTP  
Request → Response ... Request (etc)  
(**Slow**)





# Voice - Websocket

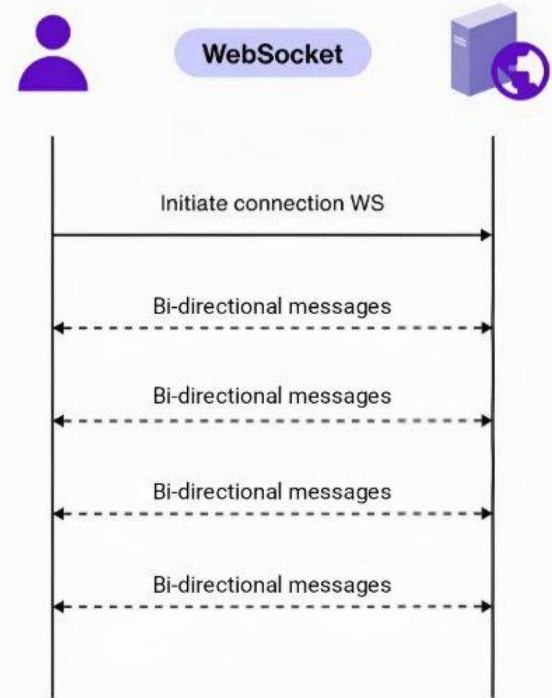
- HTTP  
Request → Response ... Request (etc)  
(**Slow**)
- Websocket





# Voice - Websocket

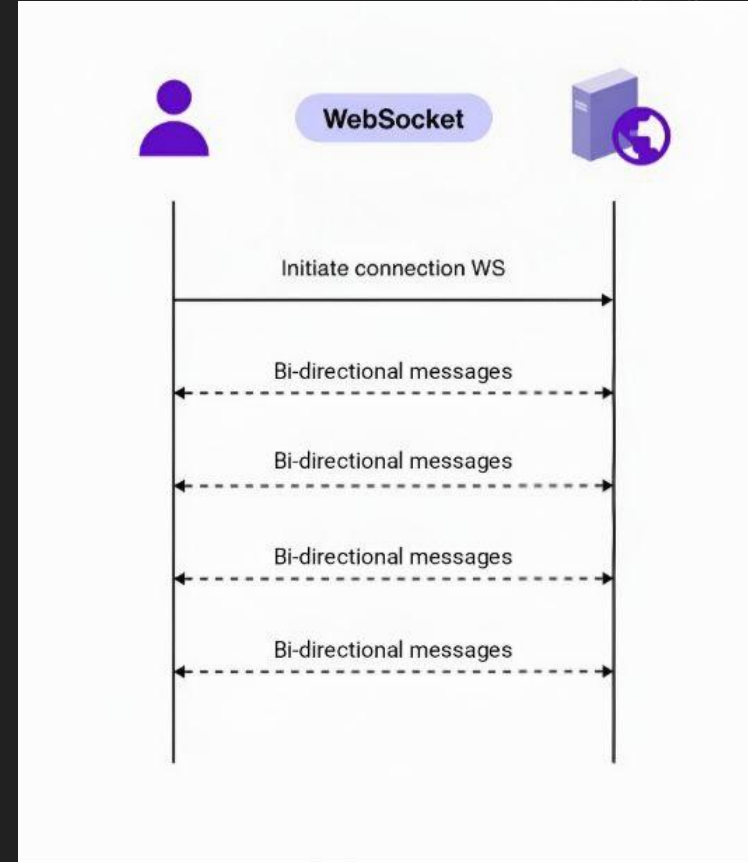
- HTTP  
Request → Response ... Request (etc)  
(**Slow**)
- Websocket  
Always open channel





# Voice - Websocket

- HTTP  
Request → Response ... Request (etc)  
(**Slow**)
- Websocket  
Always open channel  
(**Instantaneous**)





Where to start...

ARE YOU  
READY?



...with useless information





...with useless information

- Google API KEY



... but first, useless information

- Google API KEY
- Models list



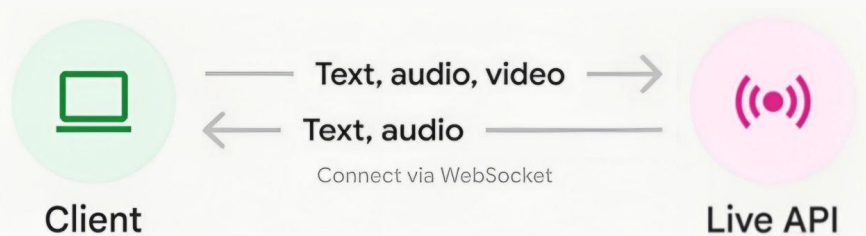


... but first, useless information

- [Google API KEY](#)
- [Models list](#)
- [Google Live Doc](#)

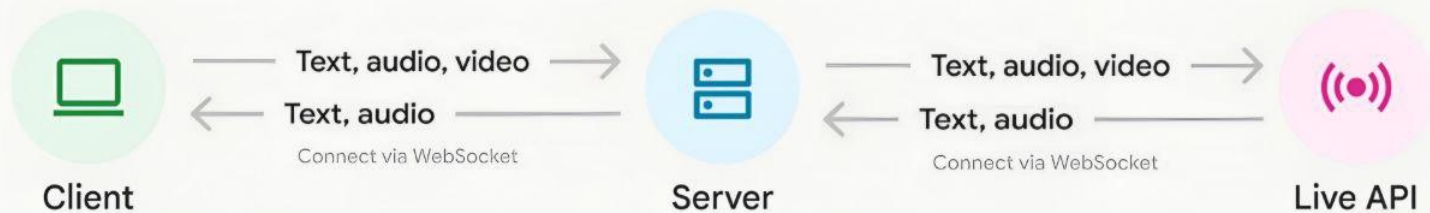


# Client - Gemini solution





# Client - Server - Gemini solution



# Ephemeral token solution





# Thank you for your attention

Leave me some feedback if you like

Github Repo

