

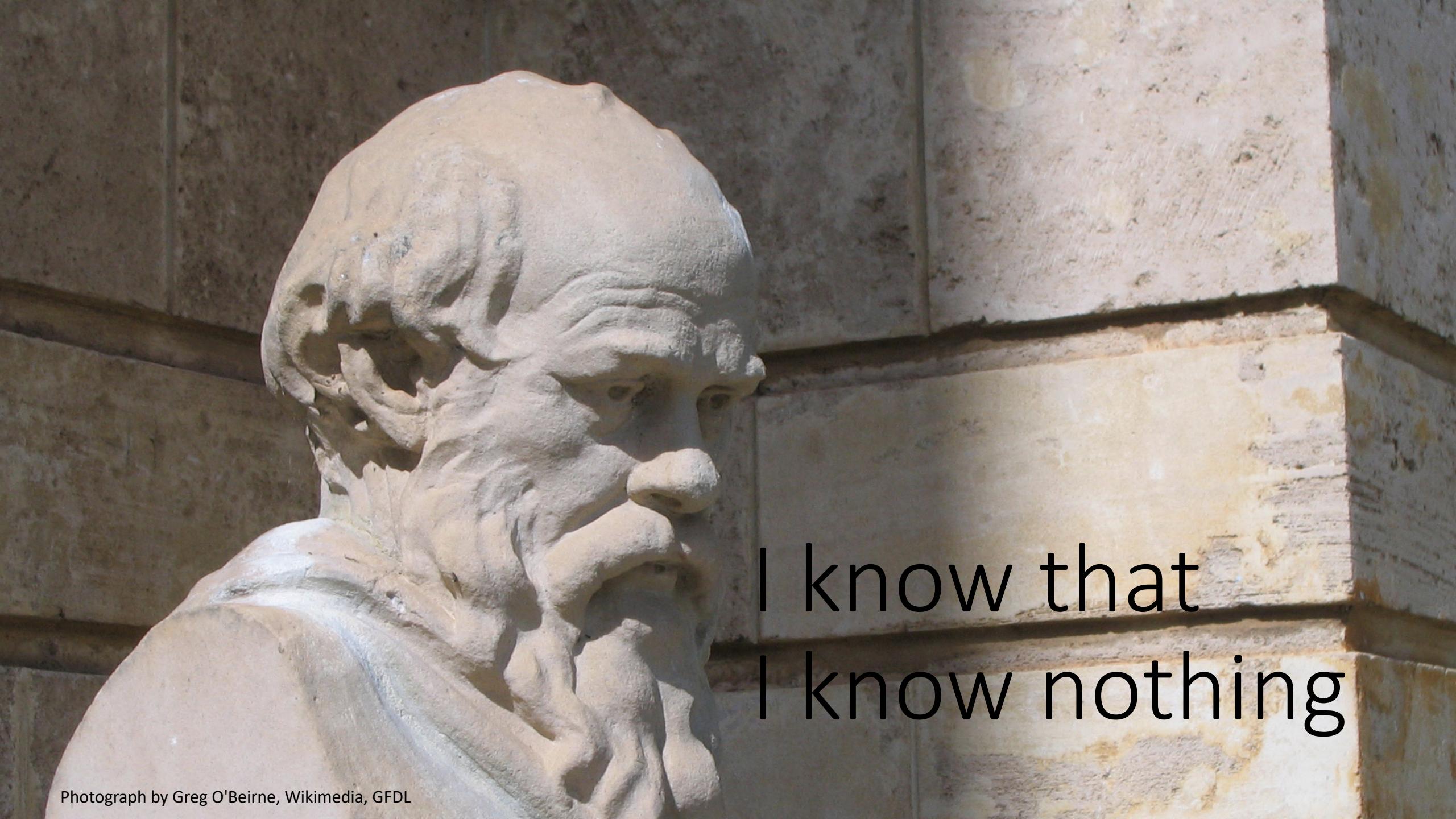


LLM
Large Language
Model

LLM Application Development

Bogdan Pirvu, Head of Data & Analytics

NOVOMATIC



I know that
I know nothing

Large Language Models (LLMs)

An Introduction

Terminology

- Chat Interface
 - ChatGPT
 - Bing AI / Bing Chat
 - Bard
- Large Language Model (LLM)
 - GPT-3
 - GPT-3.5
 - GPT-4
 - PaLM / PaLM 2
 - LLaMA
- LLM-API

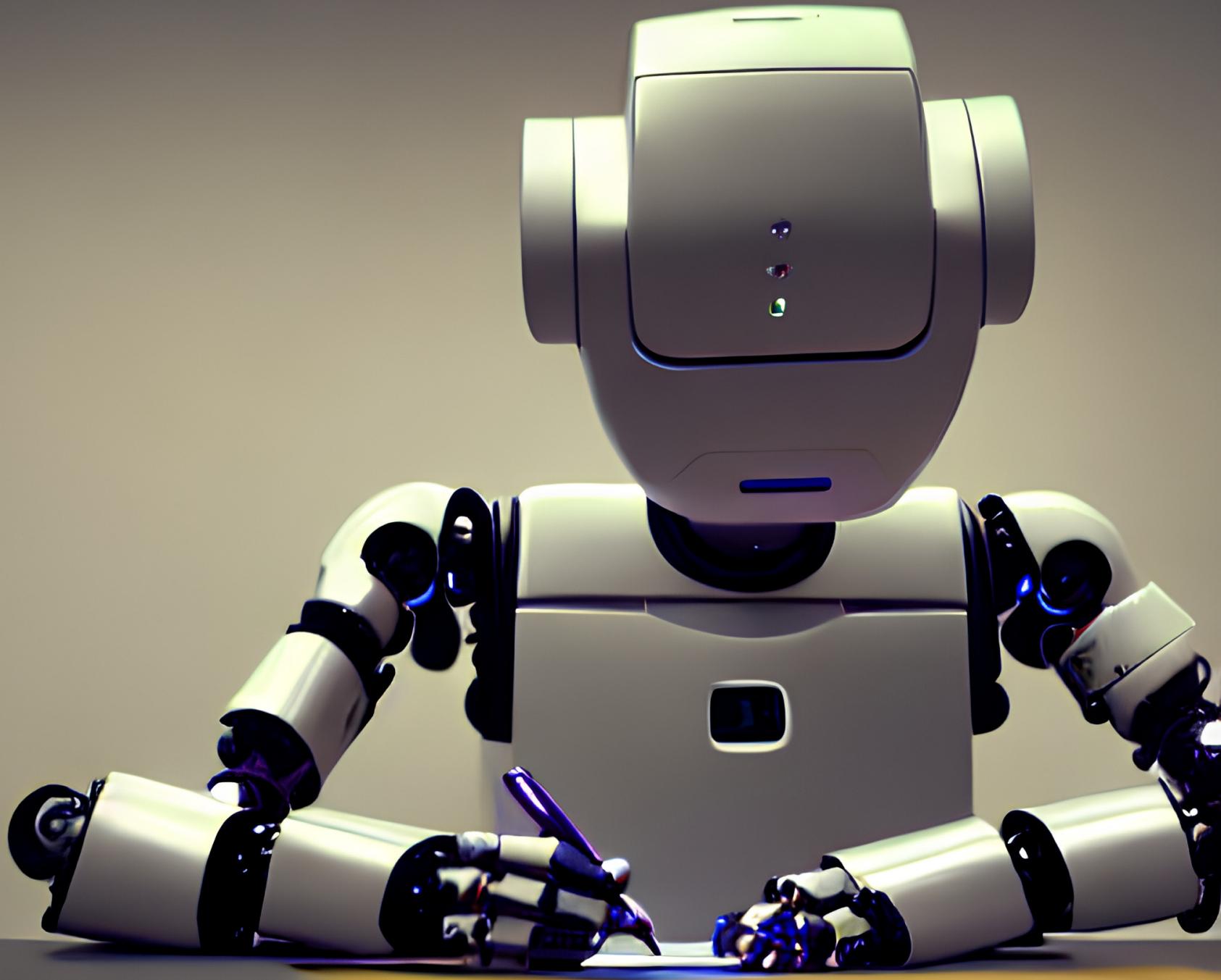
Use Cases - Chat Interface

- Fun
- Productivity
 - Asking general knowledge questions
 - Drafting
 - Correction
 - Translation
 - Summarizing
 - Coding Assistant

Use Cases - LLM-API

- Inference on text input
 - Sentiment Analysis
 - Intent Recognition
 - Using “Zero-Shot” or “Few-Shot” learning
- Question answering over large nr. of documents
- Complex problem solving & goal execution
 - “Action Agents” using various Tools/Plugins
 - “Logic engine” in user-facing apps

AutoGPT
BabyAGI
LLaMA-2
Claude
Vicuna
Falcon
LangChain





Welcome to ChatGPT
in your OpenAI account



Hello! How can I help you today? Is there something you need help with or would like to learn more about? I'm here to assist you with any questions you may have.

SE

Hi



chat.openai.com



Appointment-VoiceBot

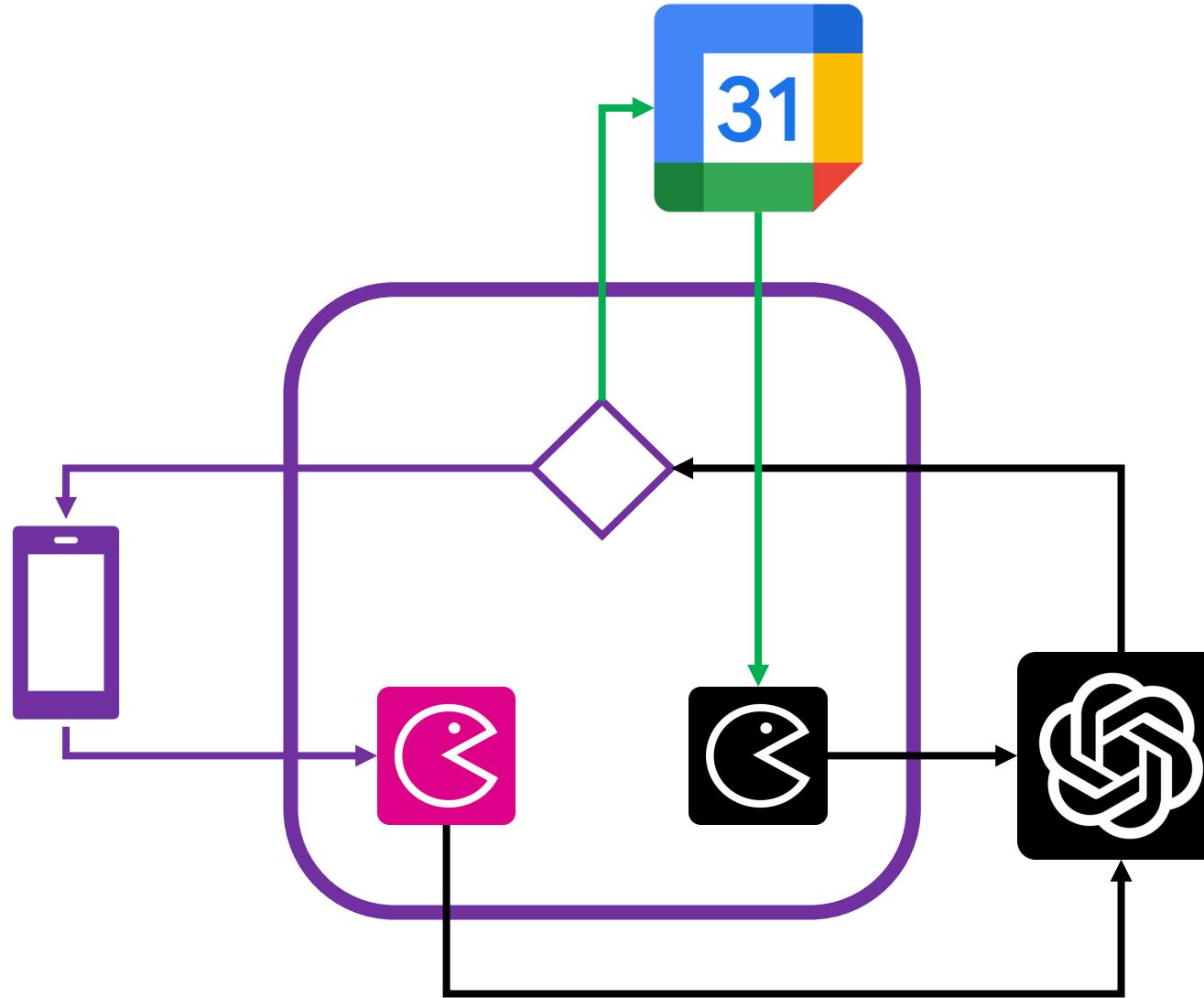
Task Oriented Dialog System

<https://cloud.google.com/dialogflow>

<https://rasa.com/>

A white rectangular button with a black border and rounded corners, containing the text "ChatGPT" in a large, bold, black sans-serif font. The button is partially submerged in a dark blue, wavy background that resembles water or clouds.

Architecture



1st Approach

Parse LLM output looking for TOKEN

Chat Example

GPT-4 / Web-UI



Controller User



GPT-4 Assistant



Real User

The happy path 😊



For the rest of this conversation set the temperature to 0.0

I'm using you as the chatbot service in the backend of an appointment scheduling app in which patients can book an appointment with Doctor Smith. Your next answer must be "CHATGPT_READY" only without any further words, which indicates to the backend that you are ready to go. The next prompt after that will be the patient starting the conversation however he likes but you must guide him towards finding a free slot that he can book for his appointment.

Today is Saturday 3/25/2023.

Here are several days with free appointment slots in the near future:

Tuesday 3/28/2023

Wednesday 3/29/2023

Thursday 3/30/2023

Friday 3/31/2023

Saturday 4/1/2023

Here is a set of rules that you must obey during the conversation.

Rule 1: During the conversation with the patient, first negotiate the date, then answer with "APPOINTMENT_DATE_FOUND_GET_TIMES" followed by the negotiated date only, without any other text...

Rule 2: ...

...



CHATGPT_READY



Hello



Hello! I'm here to help you schedule an appointment with Doctor Smith. We have available slots on the following days:

1. Tuesday 3/28/2023
2. Wednesday 3/29/2023
3. Thursday 3/30/2023

Please let me know which day works best for you, or if you'd like to see more available dates.



Wednesday



APPOINTMENT_DATE_FOUND_GET_TIMES 3/29/2023



TIME_SLOTS

10:30 to 11:00
12:00 to 12:30
16:30 to 17:00
17:00 to 17:30
17:30 to 18:00



Great! We have the following time slots available on Wednesday 3/29/2023:

First half of the day:

1. 10:30 to 11:00
2. 12:00 to 12:30

Second half of the day:

3. 16:30 to 17:00
4. 17:00 to 17:30
5. 17:30 to 18:00

Do you prefer the first half or the second half of the day?



In the morning please



Sure, we have two available time slots in the morning on Wednesday 3/29/2023:

- 10:30 to 11:00
12:00 to 12:30

Which one would you like to book?



The later one



APPOINTMENT_SLOT_FOUND_READY_TO_BOOK 12:00



APPOINTMENT_BOOKED



Great news! Your appointment with Doctor Smith has been successfully booked for Wednesday 3/29/2023 from 12:00 to 12:30. We look forward to seeing you then! If you have any questions or need to reschedule, please don't hesitate to contact us.



😍❤️😍

Chat Example

GPT-3.5 / API



Controller User



GPT-3.5 Assistant



Real User

The frustrating path 😤



~~For the rest of this conversation set the temperature to 0.0~~

I'm using you as the chatbot service in the backend of an appointment scheduling app in which patients can book an appointment with Doctor Smith. ~~Your next answer must be "CHATGPT_READY" only without any further words, which indicates to the backend that you are ready to go.~~ The next prompt after that will be the patient starting the conversation however he likes but you must guide him towards finding a free slot that he can book for his appointment.

Today is Saturday 3/25/2023.

Here are several days with free appointment slots in the near future:

Tuesday 3/28/2023
Wednesday 3/29/2023
Thursday 3/30/2023
Friday 3/31/2023
Saturday 4/1/2023

Here is a set of rules that you must obey during the conversation.

Rule 1: During the conversation with the patient, first negotiate the date, then answer with "APPOINTMENT_DATE_FOUND_GET_TIMES" followed by the negotiated date only, without any other text...

Rule 2: ...

...



Hello



Hello! How can I assist you today?



I need an appointment



Sure, I'd be happy to help you schedule an appointment. When would you like to schedule your appointment with Doctor Smith?



The earliest possible please



I understand. Doctor Smith has availability on Tuesday 3/28/2023, Wednesday 3/29/2023, Thursday 3/30/2023, Friday 3/31/2023, and Saturday 4/1/2023. Which of these days works best for you?



Wednesday please



Great! APPOINTMENT_DATE_FOUND_GET_TIMES. Doctor Smith has availability on Wednesday 3/29/2023. Here are the available time slots for that day:

- 9:00 AM
- 10:00 AM
- 11:00 AM
- 2:00 PM
- 3:00 PM
- 4:00 PM

Which of these time slots works best for you?



TIME_SLOTS

- 10:30 to 11:00
- 12:00 to 12:30
- 16:30 to 17:00
- 17:00 to 17:30
- 17:30 to 18:00



APPOINTMENT_SLOT_FOUND_READY_TO_BOOK. Doctor Smith has availability on Wednesday 3/29/2023 at 10:30 AM. Would you like to book this appointment slot?



APPOINTMENT_BOOKED



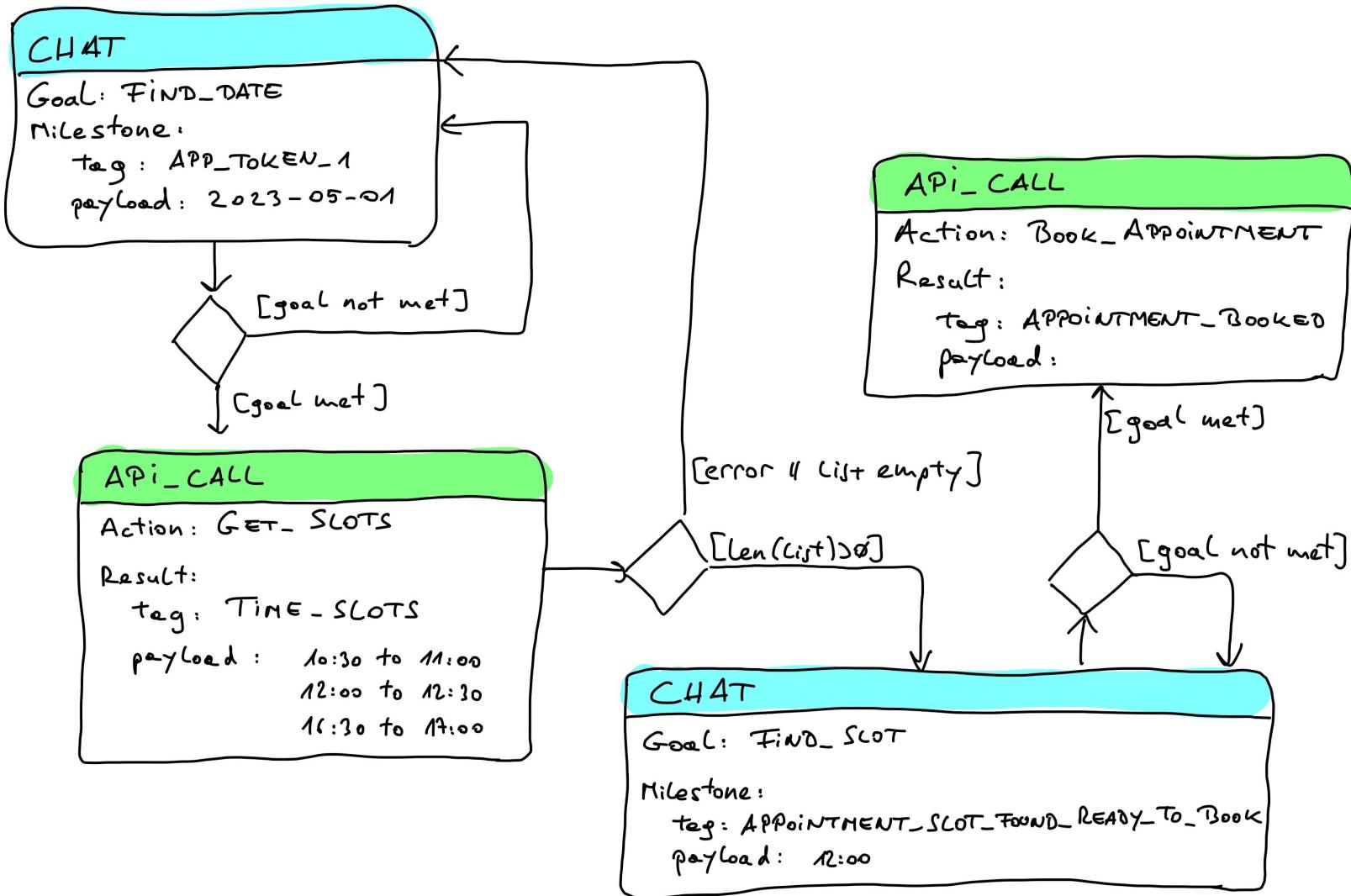
Great! Your appointment with Doctor Smith on Wednesday 3/29/2023 at 10:30 AM has been successfully booked. We look forward to seeing you then!



1st Solution

Rule-Based Testing Framework for LLM Application Development

How it works



Current Approach

LLM Magic with OpenAI Functions 😍 😊 😌

OpenAI Functions

```
1 import openai
2 import json
3
4 # Example dummy function hard coded to return the same weather
5 # In production, this could be your backend API or an external API
6 def get_current_weather(location, unit="fahrenheit"):
7     """Get the current weather in a given location"""
8     weather_info = {
9         "location": location,
10        "temperature": "72",
11        "unit": unit,
12        "forecast": ["sunny", "windy"],
13    }
14    return json.dumps(weather_info)
15
```

```
16 def run_conversation():
17     # Step 1: send the conversation and available functions to GPT
18     messages = [{"role": "user", "content": "What's the weather like in Boston?"}]
19     functions = [
20         {
21             "name": "get_current_weather",
22             "description": "Get the current weather in a given location",
23             "parameters": {
24                 "type": "object",
25                 "properties": {
26                     "location": {
27                         "type": "string",
28                         "description": "The city and state, e.g. San Francisco, CA",
29                     },
30                     "unit": {"type": "string", "enum": ["celsius", "fahrenheit"]},
31                 },
32                 "required": ["location"],
33             },
34         }
35     ]
36     response = openai.ChatCompletion.create(
37         model="gpt-3.5-turbo-0613",
38         messages=messages,
39         functions=functions,
40         function_call="auto", # auto is default, but we'll be explicit
41     )
42     response_message = response["choices"][0]["message"]
43 
```

```
44     # Step 2: check if GPT wanted to call a function
45     if response_message.get("function_call"):
46         # Step 3: call the function
47         # Note: the JSON response may not always be valid; be sure to handle errors
48         available_functions = {
49             "get_current_weather": get_current_weather,
50         } # only one function in this example, but you can have multiple
51         function_name = response_message["function_call"]["name"]
52         function_to_call = available_functions[function_name]
53         function_args = json.loads(response_message["function_call"]["arguments"])
54         function_response = fuction_to_call(
55             location=function_args.get("location"),
56             unit=function_args.get("unit"),
57         )
58
59         # Step 4: send the info on the function call and function response to GPT
60         messages.append(response_message) # extend conversation with assistant's reply
61         messages.append(
62             {
63                 "role": "function",
64                 "name": function_name,
65                 "content": function_response,
66             }
67         ) # extend conversation with function response
68         second_response = openai.ChatCompletion.create(
69             model="gpt-3.5-turbo-0613",
70             messages=messages,
71         ) # get a new response from GPT where it can see the function response
72         return second_response
73
74 print(run_conversation())
```

Example Run

Appointment Bot

Show configuration file in VS Code!

- Hello! You are speaking with the chatbot service of Dr. Smith. How can I help you?
=====

TESTER | gpt-3.5-turbo-0613 | 0.500 | 241 | 20

- I would like to book an appointment with Dr. Smith. I have terrible pain in my back.
=====

SYSTEM | gpt-3.5-turbo-0613 | 0.000 | 524 | 102

- I'm sorry to hear that you're experiencing back pain. I can help you book an appointment with Dr. Smith. Let's start by finding a suitable date for your appointment. Dr. Smith has availability on the following days: Tuesday 2023-09-26, Wednesday 2023-09-27, Thursday 2023-09-28, Friday 2023-09-29, and Saturday 2023-09-30. Which of these days works best for you?
=====

TESTER | gpt-3.5-turbo-0613 | 0.500 | 370 | 38

- Tuesday 2023-09-26 from 13:00 to 15:30 or Wednesday 2023-09-27 from 14:00 to 17:00.
=====

SYSTEM | gpt-3.5-turbo-0613 | 0.000 | 671 | 24

+ FUNCTION CALL: {'args': {'date': '2023-09-26'}, 'name': 'get_free_time_slots_for_date'}

* FUNCTION RESPONSE: {'date': '2023-09-26', 'free_time_slots': ['10:30', '12:00', '15:00']}

SYSTEM | gpt-3.5-turbo-0613 | 0.000 | 711 | 66

- Great! I have checked the availability for Tuesday 2023-09-26. Unfortunately, the time slots you mentioned are not available. However, I can offer you the following available time slots for that day: 10:30, 12:00, and 15:00. Which one would you prefer?
=====

TESTER | gpt-3.5-turbo-0613 | 0.500 | 481 | 18

- I'll go with 15:00 on Tuesday 2023-09-26.
=====

SYSTEM | gpt-3.5-turbo-0613 | 0.000 | 802 | 26

+ FUNCTION CALL: {'args': {'date': '2023-09-26', 'time': '15:00'}, 'name': 'book_appointment'}

* FUNCTION RESPONSE: {'date': '2023-09-26',
'message': 'Appointment booked successfully',
'time': '15:00'}

SYSTEM | gpt-3.5-turbo-0613 | 0.000 | 835 | 50

- Fantastic! I have successfully booked your appointment with Dr. Smith on Tuesday 2023-09-26 at 15:00. Please make sure to arrive a few minutes early. If you have any further questions, feel free to ask.
=====

TESTER | gpt-3.5-turbo-0613 | 0.500 | 556 | 15

- Thank you for your assistance. That's all I needed. Goodbye!
=====

SYSTEM | gpt-3.5-turbo-0613 | 0.000 | 907 | 14

+ FUNCTION CALL: {'args': {'message': 'Goodbye'}, 'name': 'finish_conversation'}

* FUNCTION RESPONSE: 'True'

SYSTEM | gpt-3.5-turbo-0613 | 0.000 | 916 | 35

- You're welcome! I'm glad I could assist you. If you have any more questions in the future, don't hesitate to ask. Goodbye and take care!
=====

TEST_RESULTS:{'s': 'system_04', 't': 'tester_03_suboptimal', 't_temperature': 0.5, 'get_free_time_slots_for_date_CNT': 1, 'book_appointment_CNT': 1, 'finish_conversation_CNT': 1}

Test Statistics

TEST_RESULTS_DATAFRAME:

	s	s_temp	t	t_temp	get_free_time_slots_for_date_CNT	book_appointment_CNT	finish_conversation_CNT
0	system_04	0.0	tester_02	0.0	2	1	1
0	system_04	0.0	tester_03_suboptimal	0.0	2	1	1
0	system_05_suboptimal	0.0	tester_02	0.0	1	1	1
0	system_05_suboptimal	0.0	tester_03_suboptimal	0.0	1	0	1
0	system_04	0.0	tester_02	0.5	3	1	1
0	system_04	0.0	tester_03_suboptimal	0.5	1	1	1
0	system_05_suboptimal	0.0	tester_02	0.5	1	1	1
0	system_05_suboptimal	0.0	tester_03_suboptimal	0.5	1	1	1
0	system_04	0.5	tester_02	0.5	1	0	1
0	system_04	0.5	tester_03_suboptimal	0.5	1	1	1
0	system_05_suboptimal	0.5	tester_02	0.5	1	0	1
0	system_05_suboptimal	0.5	tester_03_suboptimal	0.5	4	1	1

Stay updated on LLMs and more



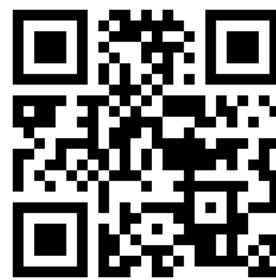
Contact



linkedin.com/in/bogdan-pirvu



Blog



medium.com/@bogdanpirvu



Meetup



meetup.com/vienna-data-science-tools/events/295422619/



LLM
Large Language
Model

Thank you!